Use Cases in Big Data Software and Analytics

Vol. 3, Fall 2017

Bloomington, Indiana

Tuesday 19th December, 2017, 20:55

Editor:
Gregor von Laszewski
Department of Intelligent Systems
Engeneering
Indiana University
laszewski@gmail.com

Contents

Chapter 1

Preface

1.1 Disclaimer

The papers provided are contributed by students of the i523 class thought at Indiana University in Fall of 2017. The students were educated in plagiarizm and we hope that all papers meet the high standrads provided by the policies set at Indiana University in regrads to plagiarizm. In case you notice any issues, please contact Gregor von Laszewski (laszewski@gmail.com) so we cn address the issue with the student.

1.2 Citation

The proceedings is at this time available as a draft. To cite this proceedings you can use the following citation entry:

```
@Book{las17-i523,
  editor =    {Gregor von Laszewski},
  title =        {Use Cases in Big Data Software and Analytics},
  publisher =    {Indiana University},
  year =         {2017},
  volume =    {1},
  series =    {i523},
  address =   {Bloomington, IN},
  edition =   {1},
  month =     dec,
  url={https://github.com/laszewski/laszewski.github.io/raw/master/papers/vonLaszewski-i
}
```

Contributors to the volume can cite their contribution as follows. They just need to *FILLIN* the missing information

```
@InBook{las17-,
```

```
  author =    {FILLIN},
  editor =    {Gregor von Laszewski},
  title =        {Use Cases in Big Data Software and Analytics},
  chapter =      {FILLIN},
  publisher =    {Indiana University},
  year =         {2017},
  volume =    {1},
  series =    {i523},
  address =   {Bloomington, IN},
  edition =   {1},
  month =     dec,
  url={https://github.com/laszewski/laszewski.github.io/raw/master/papers/vonLaszewski-i
  pages =     {FILLIN},
}
```

1.3 List of Papers

HID	Author	Title
101, 230	Huiyi Chen, Yuanming Huang	Big Data in Job Recommendation Systems
102	Dianprakasa, Arif	TBD
104, 216	Jones, Gabriel, Millard, Mathew	Big Bias? An Analysis of Google Search Suggestions
105	Lipe-Melton, Josh	Predictive Model For English Premier League Games
107	Ni,Juan	Big Data Analytics in Support Filtering Wrong Informations On Social Networking Sites
109, 106	Shiqi Shen, Qiaoyi Liu	Diversification of Big Data
201	Arnav, Arnav	IoT Application Using MQTT and Raspberry Pi Robot Car
202, 205	Himani Bhatt, Mrunal Chaudhary	Big Data Analysis in E-Commerce
209, 213, 214	Han, Wenxuan, Liu, Yuchen, Lu, Junjie	Analysis of Digit Recognizer classification algorithms in big data
211	Khamkar, Ajinkya	Continuous motion tracking using Deep Neural Networks and Recurrent Neural Networks
212, 225, 210	Kumar, Saurabh; Schwartzer, Matthew; Hotz, Nicholas	Can Blockchain Adoption Mitigate the Opioid Crisis Through More Secure Drug Distribution?
215, 208	Mallala, Bharat, Jyothi Pranavi Devineni	Big Data Analytics on Influencers in Social Networks
219	Syam Sundar Herle	Unsupervised Learning for detecting fake online reviews
224	Rawat, Neha	Big Data Analytics in Detection of DDoS (Distributed Denial-of-Service) attacks
228	Swargam, Prashanth	Big data applications in Indian Premier League
229	ZhiCheng Zhu	Big Data Analytics in Product Development Management

231, 203	Vegi, Karthik, Chandwani, Nisha	Big Data Analytics on Food Products Around the World
232	Rahul Velayutham	Big Data and Hearing Disabilities
233, 204	Wang, Jiaan, Chaturvedi, Dhawal	Big Data in Safe Driver Prediction
234	Weixuan Wang	Big Data Analytics and Applications in the Travel Industry and its Potential in Improving Travel Accessibility
235	Yujie Wu	Big Data analytics in predict house price
236, 218	Weipeng Yang, Geng Niu	Big Data and Its Application in Education
237	Tousif Ahmed	Analyzing everyday challenges of people with visual impairments
301	Gagan Arora	Importance of Big data in predicting stock price
302	Sushant Athaley	Recipe Ingredients Analysis
304	Ricky Alan Carmickle	How Far have Space Walks Walked
hid305	error: yaml	How Far have Space Walks Walked
306, 338	Murali Cheruvu, Anand Sriramulu	Predicting Housing Prices
308	Pravin Deshmukh	TBD
hid309	error: yaml	TBD
310	Kevin Duffy	Gerrymandering Detection Using Data Analysis
311	Matthew Durbin	Big Data in Genomics and Medicine
312	Neil Eliason	Big Data Mental Health Monitoring - A Private and Independent Approach
313	Tiffany Fabianac	The Impact of Clinical Trial Results on Pharmaceutical Stock Performance
314	Sarang Fadnavis	TBD
315	Garner, Jeffry	TBI - A Data Driven Journey Beyond Contact Sports... Putting Data In The Drivers Seat
316	Robert Gasiewicz	Big Data and Edge Analytics in Weather Monitoring and Forecasting
318	Irey, Ryan	None
319	Mani Kumar Kagita	Face Detection and Recognition Using Raspberry Pi Robot Car
320	Elena Kirzhner	Big Data Applications in Real Estate Analysis
323	Uma M Kugan	Plugin to cmd5 That Creates a Docker Swarm Cluster on 3 Raspberry Pis
324	Ashok Kuppuraj	Big Data Analytics in factors affecting Bitcoin
325	J. Robert Langlois	The importance of data sharing and replication, but what about data archiving?
326	Mohan Mahendrakar	None
327	Paul Marks	How Big Data Will Help Improve People's Health Worldwide
328	Dhanya Mathew	Predicting Profitable Customers in Banking Industry
329	Ashley Miller	Big Data and The Customer Experience Journey
330	Janaki Mudvari Khatiwada	Big Data Analytics in Monitoring Outdoor Air Quality
331	Tyler Peterson	Big Data Applications in Predicting Hospital Readmissions
332	Judy Phillips	Big Data Analytics to Reduce Health Care in the United States
334	Peter Russell	The Intersection of Big Data and IoT
335	Sean Shiverick	Using Machine Learning Classification of Opioid Addiction for Big Data Health Analytics
336	Jordan Simmons	None

337, 333	Ashok Reddy Singam, Anil Ravi	IoT and Big Data Analytics for Equipment Predictive Health Management (PHM)
339	Hady Sylla	Diagnosis of Coronary Artery Disease Using Big Data Analysis
340	Timothy A. Thompson	New Approaches to Managing Metadata at Scale in Research Libraries
341	Tibenkana, Jacob	Not submitted
342	Nsikan Udoyen	TBD
343	Borga Edionse Usifo	Income Prediction Using Machine Learning Techniques
345	Ross Wood	Agricultural Data Science
346	Zachary Meier	Big Data Analysis for Wild File Prevention and Tracking
347	Jeramy Townsley	Killings by Police in the United States
348	Budhaditya Roy	Big Data Application in Precision Medicine and Pharmacogenomicsn

Big Data Analytics in E-commerce

Himani Bhatt, Mrunal L Chaudhary
Indiana University
Bloomington, Indiana
himbhatt@iu.edu,mchaudh@iu.edu

ABSTRACT

Humongous amounts of data gets generated every day in the domain of E-commerce industry. With the increasing competition and ever-changing market trends, it is a challenging task for the store owners to strategize business and marketing activities. If the companies are able to predict customer behavior, they can come up with business designs which can help them in making predictions about the customer purchasing patterns and thereby increase their revenue. In this project we have aimed to do analysis on the data of an E-commerce non-store online retail giant based in UK. The dataset, available in the UC Irvine repository by the name of 'Online Retail', consists of the the goods purchased by different customers at a given time. Through this data available to us, we have done customer segmentation on the basis of the type and amount of goods purchased by a customer. We achieved this by doing a thorough exploration of the data, data pre-processing and then running different Machine Learning Classifiers to classify the customers in different categories.

KEYWORDS

HID 202, HID 205, i523, Machine Learning, Analysis, E-commerce, retail, Customer Segmentation, Python, Regression, Boosting, KNN, Random Forest.

1 INTRODUCTION

The E-commerce industry is in constant shifts due to the ever-increasing changes in the technologies used to develop and maintain the E-commerce systems, the services that they are willing to offer, the market strategies which gain popularity at the time, and most importantly- the customer behavior [3]. The online store owners are the ones who are most affected by these changes. And since the competition in the field of E-commerce is fierce, the online store owners need to come up with business strategies and technologies which provide better customer services leading to their satisfaction and earning customer loyalty. To achieve this, they need to address these ever changing issues to survive and thrive in the E-commerce market and come up with better decisions faster. The key to achieving this lies in better understanding of the customer behavior and their purchasing patterns. That is where analytics comes into play. Analysis of customer behavior and purchasing patterns helps in devising better and accurate marketing strategies which can not only help in generating more profits but also in saving both time and efforts that goes into trying and testing different marketing activities [3]. This ability to capture and analyze user data, and then provide useful and in depth insights in it is what Machine Learning empowers us with. In this project, we aim to do analysis on a data set 'Online Retail' from the UC Irvine Machine Learning Repository to determine the customer purchasing pattern by using different machine Learning algorithms like K-Means Clustering, Logistic Regression, Random Forest, Gradient Boosting, etc.

2 BACKGROUND

Before the advent of the World Wide Web, transactions that happened on a day to day basis meant physical presence of customers, the brick-and-mortar setting of a store which offered a limited variety of goods. With the evolution of internet and its application in retail, the field of E-commerce emerged and changed the entire facet of shopping. Since a proper set-up of a store is no longer needed, customers can buy goods at much lower prices, with a wider variety to choose from and that too without the need of physical presence. The online market is expected to grow by almost 56% from the year 2015 to 2020 [16]. In the United States alone, 56% of the population prefers to shop online. The E-commerce industry is growing at an average rate of 23% every year, with 90% of the Americans having done online shopping at some point in their lives [15]. With so many transactions happening over the internet, naturally the amount of data getting generated is humongous. Also, with the constantly changing market trends, strategies to overcome the competition and make profits need to be constantly improved. The key issues therefore are managing the data and drawing insights from them which will help in bettering the business decisions. To store and maintain the magnanimous amounts of data getting generated *everyday* is a huge hassle, because along with the volume, this data gets generated at a break neck speed and in different formats from traditional numeric databases to unstructured text documents [12]. The big data technologies like Hadoop and Spark can be used in addressing these hurdles, namely the volume, velocity and variety.

2.1 The Three V's of E-commerce Big Data

Like other technologies which deal with a humongous amount of data, E-commerce must also respond to the 3 Vs, namely Volume, Velocity and Variety:

Volume Thousand of online transactions happen every day making the data generation a real time process. The integration of Big Data involves collection of relevant data like customer behavior statistics on the basis of their searches, transactions, demography, etc. The challenge here is not only gathering the data but also in analyzing it.

Variety The data from online transactions comes in different varieties, right from structured databases to unstructured text documents, videos, feedback emails and comments, and others. The retailers need to understand this for making the right business decisions by keeping a leeway for possible data fluctuations such as seasonal ad peak loads like Black Friday sales.

Velocity Handling the huge amounts of data which is generated at unprecedented rates is another challenge that needs to be taken care of. It is therefore imperative to do rapid analysis so that timely actions can be taken to sustain in the competition and boost the profit margins

Storing and maintaining the big data is a hassle in itself, but it will provide little value if proper analysis is not done on it. That is where we will be focusing on in this project - making sense of the data. Hence for the scope of this project, we have performed analysis on a small data set of around 45 MB. Machine Learning Algorithms learn from the data. Since we will be using Machine Learning Algorithms, the accuracy of Analysis will only increase with increase in the size of the dataset. We will be discussing this in further detail in the coming sections.

We have now established the fact that the E-commerce companies have a lot of data at their fingertips. Making use of this data is where the challenge lies. Machine learning is an approach by which insights can be drawn from digital data at a rate much faster than any human is capable of doing [7] . Following are some of the biggest challenges that are faced in the field of E-commerce which Machine Learning addresses successfully:

(1) Optimization of the Prices:

Pricing, and in that, online pricing is critically important. Since prices of the competitors are only a few clicks away, it is far easier for the customers to compare prices. Setting up the optimum price, by considering many factors like the prices set by the competitors, the time of the day, the type of the customer and the product's demand therefore is a difficult task. Machine Learning technology can set the prices by considering all these factors at once.

(2) Fraud detection:

The E-commerce industry, like the other industries, is susceptible to fraudulent activities. The consequences of these activities can lead to tarnishing the name of the company forever. Machine Learning helps in detecting and preventing the frauds by processing the repetitive data at a high speed.

(3) Search Ranking:

Machine Learning is capable of pulling information from patterns of search and purchase by considering the factors like preferences, content and search items and come up with a powerful search engine that shows what the customer exactly wants.

(4) Product Recommendations:

Machine Learning is capable of effortlessly quantifying the buying patterns of the customers and developing a recommendation engine which makes relevant product suggestions to them.

(5) Customer Segmentation and Personalization:

In any business, Customer base is the most important factor and therefore providing a satisfactory customer experience is of utmost importance. The biggest challenge that E-commerce systems endeavor to overcome is the separation from their customers. In person, a salesperson can quickly take in what the customers are saying, their economic status, their body language, and behavior to help them find better or desired products. The salesperson thus is able to *segment* customers, and provide them with a *personalized* shopping experience. With online shopping, it is very difficult to make this happen since an in depth understanding is needed of the vast amount of the data to provide tailored choices to the customers, which can result in sale loss.

Machine Learning makes the biggest impact by making it possible to give personalized customer experiences which can boost the sales and thereby increase the revenue.

The type of analysis and Machine Learning Algorithm to be chosen depends solely on the data at hand. The data set we aim to analyze is a transnational data set that has been archived in the UC Irvine Machine Learning Repository under the name 'Online Retail'.

A thorough Exploratory Data Analysis on this data lets us know what kind of Machine Learning Algorithm needs to be used. Also, several models can be applied and the one which gives the best accuracy and precision against the test data can be chosen. To add icing to the cake, we can even combine the results given by the different models to make an ensemble model which gives an accuracy that is better than that of the individual algorithms. Machine Learning Algorithms are mostly classified as supervised and unsupervised Learning algorithms.

In Supervised Learning, each example is a pair of an input object and the corresponding output value, also called the supervisory signal. A supervised learning algorithm analyzed the training data to produce an inferred function which is used to map new, unknown output-value examples [8]. Since there is the output value to *supervise* the learning algorithm, such approach is called 'Supervised Learning'. The most commonly used Supervised Learning Algorithms are Logistic and Linear Regression, Bagging and Boosting Algorithms, Decision Trees and Random Forest. The Logistic Regression algorithm determines the relationship between the input and the output variables and generates a classifier model to predict the category to which a new example belongs to. Thus Logistic Regression is a classification algorithm [2]. Decision Trees are non parametric Supervised Learning Algorithms which create a model by learning simple decision rules inferred from data attributes to predict the value of a target variable. Decision Trees can be used either for Classification or Regression [5]. Bagging is a technique used to reduce the variance in the predictions by combining the result of multiple classifiers modeled on different sub-samples of the same data set. One of the most commonly and widely used implementation of Bagging is Random Forests. In Random forest, there are multiple trees which classify a new sample based on the set of attributes and a new sample is classified to that class which received the maximum 'votes' from the individual trees. In case of doing Regression with the help of Random Forest, the average of the outputs given by different trees is taken [10].

In Unsupervised learning, only the input data is known with no knowledge of the corresponding output variable. The goal therefore of the Unsupervised Learning Algorithms is to model the underlying distribution or structure in the data to understand the data more. Since there is no output available to validate or 'supervise' the answers, such learning algorithms are called Unsupervised. The most common application of unsupervised learning is clustering. Clustering enables to differentiate the data by discovering the inherent groupings of the input. The most common implementation of Clustering is the K-means algorithm. This algorithm works iteratively to assign each data point to one of the k-groups based on the feature similarity.

3 EXPLORING THE DATASET

The dataset taken for the analysis is the 'Online Retail' data set available on the UCI Machine Learning Repository. This is a transnational dataset which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data set consists of 5,41,909 transactions and 8 features which describe each of these transactions. There are missing values present in the dataset. All the attributes are integer and real numbers. The size of the dataset is 43.4 MB.

3.1 Attribute Information

InvoiceNo : This refers to the Invoice number. It is a Nominal, 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
StockCode : This refers to the Product (item) code. It is a Nominal, 5-digit integral number uniquely assigned to each distinct product.
Description : This refers to the Product (item) name. It is of Nominal data type.
Quantity : This refers to the The quantities of each product (item) per transaction. It is Numeric in type.
InvoiceDate : This refers to the Invoice Date and time. It is Numeric,and represents the day and time when each transaction was generated.
UnitPrice : This refers to the Unit price. It is of Numeric type, and represent the Product price per unit in sterling.
customerID : This refers to the Customer number. It i a Nominal, 5-digit integral number uniquely assigned to each customer.
Country : This refers to the Country name. It is of Nominal type, and represents the name of the country where each customer resides.

4 DATA PREPARATION

4.1 Installation Steps

The project has been implemented in Python 2.7 version and we have used the Jupyter Notebook App for the program execution. The Jupyter Notebook Application is an application having server-client architecture which allows editing and executing notebook documents through a web browser. A notebook document is a human readable and machine executable document which can be executed for implementation of data analysis. The Jupyter Notebook Application can be executed on the local host or can be installed on a remote machine accessed via the internet [6].
The Jupyter Notebook can be installed very easily on a machine which has either Python 2 or Python 3 version. Since we have implemented our project in Python 2.7, following commands are to be run in the terminal:
pip install –upgrade pip
The above command will upgrade the Python package manager (pip). pip install jupyter
The above command will install Jupyter in the local machine.
Once the Jupyter Notebook has been installed, it can be run using the following command in the terminal:
jupyter notebook
This command will run the Jupyter notebook in the default browser of the machine on the default port 8888 of the localhost.

Figure 1: Data set Contents

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom

4.2 Packages Installation

Before running the code the following packages were imported/installed in the Python environment.

Pandas Pandas provide a very fast and flexible data structures to make working with relational data easy and fairly intuitive.

Numpy This is a fundamental package for scientific computation with Python and can be used as an efficient multi-dimensional container of generic data.

Sklearn Scikit-learn makes a wide range of supervised and unsupervised machine learning algorithms available in python. We have implemented all the Machine Learning Algorithms using this library.

4.3 Null Value Treatment

Before going ahead with Data Exploration, a quick look through the data showed many missing values. Hence before doing any analysis, it is imperative to treat the missing values. The dataset has almost 25% of the entries that are not assigned to any of the customer i.e. customerID attribute for those entries is null.
Missing value treatment can be done by deleting the columns and/or rows which have missing values beyond a decided threshold, or replacing them with the attribute mean, median or mode. Since the missing values in our case is the customerID, the replacement method cannot be applied. Also, these entries are useless for the analysis since we aim to do Customer Segmentation and without knowing the customerID, it cannot be achieved. Hence we have deleted the rows with missing customerID. After removing these entries , the dataset left is with 4,06,829 transactions.

The content of the dataset appears as shown in the Figure 1.
We also removed the duplicate values present in the dataset. There are 5225 such entries present in the data set that are deleted.

5 EXPLORING THE CONTENT OF VARIABLES

The dataframe has 8 variables and we can draw some inferences by analyzing these variables.

5.1 Countries

From the data we can see that there are 37 different countries from which orders were placed. We can determine the number of orders per country by a 'Chloropeth' map. A Chloropleth map shown in

Figure 2: Distribution of Orders based on Countries

Figure 3: Customer Products Transactions

	products	transactions	customers
quantity	3684	22190	4372

Figure 4: Number of products per Customer

	CustomerID	InvoiceNo	Number of products
0	12346	541431	1
1	12346	C541433	1
2	12347	537626	31
3	12347	542237	29
4	12347	549222	24

Figure 2 uses different colors and shades within predefined areas to indicate quantities in those areas.

The Figure 2 shows that maximum number of orders are placed from UK.

5.2 Customers and products

On observing the number of users, products purchased and number of transactions made; we can see that these are not proportional. This suggests that there were many transactions made for cancelling the orders shown in Figure 3

We can also determine the number of products purchased in each transaction. It shows that some customers purchased goods in bulk whereas some purchased a single product in a transaction. Also the orders with InvoiceNo starting with C are the cancelled orders.The details are shown in Figure 4.

Figure 5: Transactions for Cancellation

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
61619	541431	23166	MEDIUM CERAMIC TOP STORAGE JAR	74215	2011-01-18 10:01:00	1.04	12346	United Kingdom
61624	C541433	23166	MEDIUM CERAMIC TOP STORAGE JAR	-74215	2011-01-18 10:17:00	1.04	12346	United Kingdom
286623	562032	22375	AIRLINE BAG VINTAGE JET SET BROWN	4	2011-08-02 08:48:00	4.25	12347	Iceland
72260	542237	84991	60 TEATIME FAIRY CAKE CASES	24	2011-01-26 14:30:00	0.55	12347	Iceland
14943	537626	22772	PINK DRAWER KNOB ACRYLIC EDWARDIAN	12	2010-12-07 14:57:00	1.25	12347	Iceland

Figure 6: Stock Codes

```
POST            -> POSTAGE
D               -> Discount
C2              -> CARRIAGE
M               -> Manual
BANK CHARGES    -> Bank Charges
PADS            -> PADS TO MATCH ALL CUSHIONS
DOT             -> DOTCOM POSTAGE
```

5.3 Cancelled Orders

Almost 16% (3654) of the transactions are corresponding to the cancelled orders. In the dataset, corresponding to each cancelled transaction we should have an order placed with same quantity of products requested. While checking the same in the dataset, we found the details shown in Figure 5 for some of the orders.

This hypothesis should apply to the complete dataset, but on checking the whole dataset it is found out that there are some cancelled orders without the purchase order (the history of the order) made. This is done by locating the entries that indicate a negative quantity and then checking if there is an order indicating the same quantity (but positive) with the same description and the same customerID. We still get negative quantities. Going deeper in to this suggests that the entries with description 'Discount' have negative quantities associated with that transaction. And hence, to do the verification, we eliminated the 'Discount' entries. But again the initial hypothesis do not match; we still have negative numbers appearing in the quantity.
This can be because the buy orders were performed before December 2010 (the point of entry of the database). We can delete the records where a cancel order exists without the corresponding purchase order or where there is at least one counterpart with the exact quantity (since both records are logically cancelling each other).Total 8795 such records are found and deleted from the dataset.

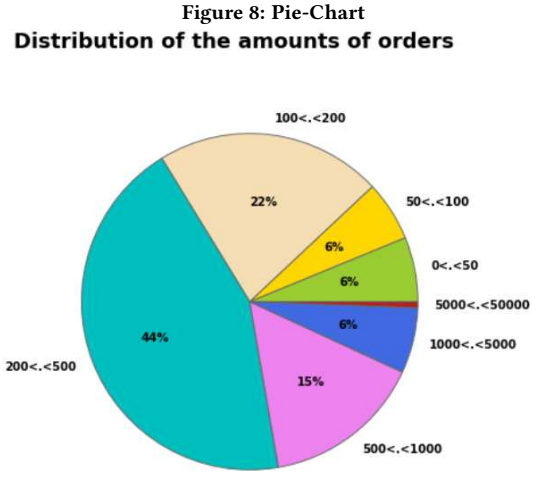
5.4 StockCode

The StockCode variable should ideally contain letters. So we have filtered out the codes with only letters. We can observe from Figure 6, different type of transactions based on these (example D is for discounted transaction).

Figure 7: Basket Price

	CustomerID	InvoiceNo	Basket Price	InvoiceDate
1	12347	537626	711.79	2010-12-07 14:57:00.000001024
2	12347	542237	475.39	2011-01-26 14:29:59.999999744
3	12347	549222	636.25	2011-04-07 10:42:59.999999232
4	12347	556201	382.52	2011-06-09 13:01:00.000000256
5	12347	562032	584.91	2011-08-02 08:48:00.000000000
6	12347	573511	1294.32	2011-10-31 12:25:00.000001280

Figure 8: Pie-Chart



5.5 Basket Price

We have added a new variable to indicate total price of the purchase (by multiplying unit price of each product with quantity purchased). Each transaction corresponds to the prices for a single product. On grouping the records based on a single order, we can see the complete price for that order as shown in Figure 7.

We can visualize the orders distinguished on the basis of total price of the basket. It can be shown as Figure 8 using a pie-chart.

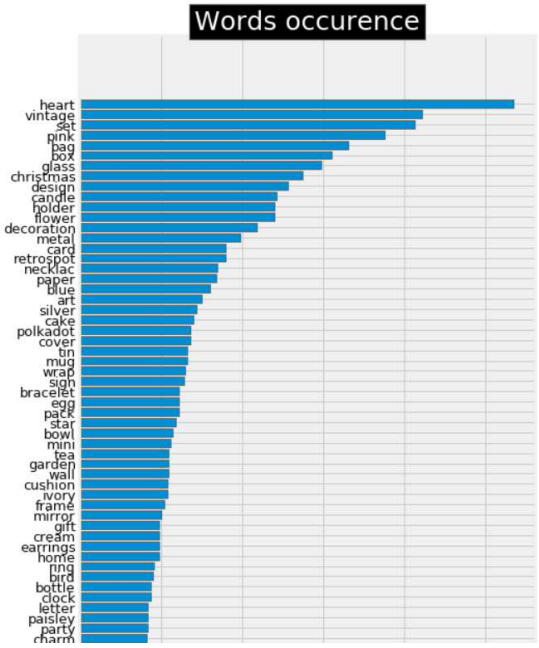
It shows that majority of the orders are the bulk purchases since 60% of the orders have amounts greater than 200 Sterling.

6 EXPLORING PRODUCT CATEGORIES

The dataset contains two variables- Stockcode and Description defining products. We can categorize the products based on the content of the description variable. This can be done in the following way. Firstly, the proper or the common names appearing in the products' description are extracted. Then the root of the word and combining set of names associated with this particular root is extracted. Lastly, the frequency of the word is found in the description variable of the dataframe.

Upon checking, we found that there are 1483 keywords present in the description variable of the dataset. The most common keywords can be determined based on the occurrences. The Figure 9 shows the top word occurrences.

Figure 9: Word Occurrences



6.1 Categorizing Products

We have obtained around 1400 keywords from the above occurrence list , most of which do not make sense. After discarding the keywords that are appearing less than 13 times, we are left with 193 keywords that we will consider for our analysis.

These significant keywords are used for creating categories of the products. The data has been encoded using the principle of one-hot-encoding.

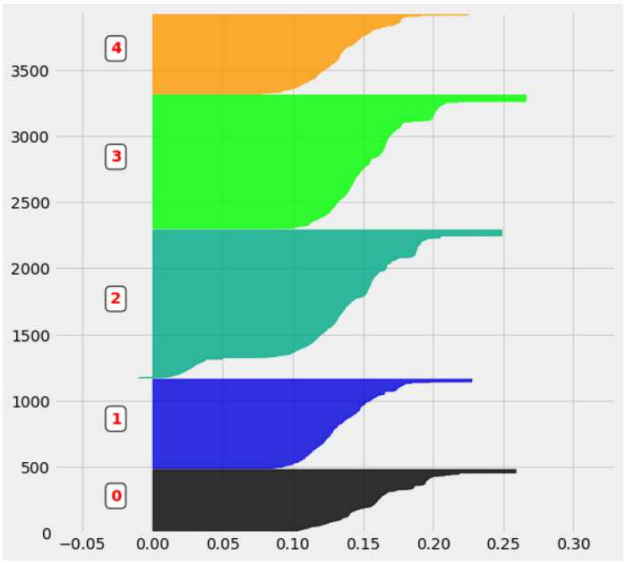
One hot encoding - One hot encoding is a process by which categorical variables are converted into a binary format of 0's and 1's that could be provided to ML algorithms to do a better job in prediction. The words present in the descriptions of the products are encoded. Also price range column is added as it will help in balanced grouping of the products.

6.2 Clustering of products

In the previous step we have created a matrix with encoded version of words present in the description variable. K-means clustering is used for the cluster assignment and since the data is in binary format because of encoding, the most appropriate distance method will be Hamming's metric (other distance functions are euclidean distance, Manhattan distance, binary distance, etc). It basically measures the minimum number of substitutions required to change one string into the other. But since the k-means package available in sklearn uses Euclidean distance by default, we have used it for our analysis.

Selection of optimum K-value:

Figure 11: Silhouette plot



The number of clusters can be selected using silhouette analysis on K-means clustering. It is used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to the points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1]. Silhouette coefficients (as these values are referred to) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

The Figure 10 shows silhouette score for different values of k. These scores do not have significant differences, but since for k value greater than 5, the resulting clusters have very few elements in them, we have taken k as 5.

Figure 10: Silhouette Scores

```
('For n_clusters =', 3, 'The average silhouette_score is :', 0.10158702596012364)
('For n_clusters =', 4, 'The average silhouette_score is :', 0.12680045883937879)
('For n_clusters =', 5, 'The average silhouette_score is :', 0.14553871352885445)
('For n_clusters =', 6, 'The average silhouette_score is :', 0.15122077520906058)
('For n_clusters =', 7, 'The average silhouette_score is :', 0.1463684372259042)
('For n_clusters =', 8, 'The average silhouette_score is :', 0.14764212603720744)
('For n_clusters =', 9, 'The average silhouette_score is :', 0.1397423040247273 7)
```

6.3 Validating Quality of Classification

6.3.1 ***Silhouette Score.*** From the silhouette plot shown in Figure 11 we can see that cluster 1 has more number of elements than the other clusters. But overall distribution of elements in the clusters is comparative. Same can be seen from the Figure 12.

6.3.2 ***Principal Component Analysis.*** The main idea of principal component analysis (PCA) is to reduce the dimensionality of

Figure 12: Cluster Composition

2	1118
3	1009
1	673
4	606
0	472

Figure 13: PCA

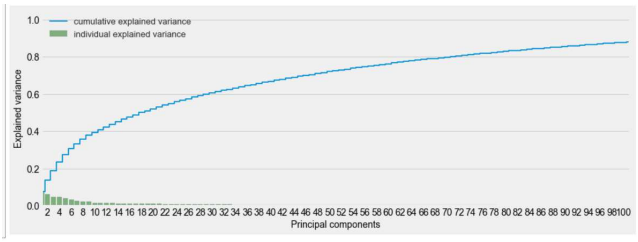
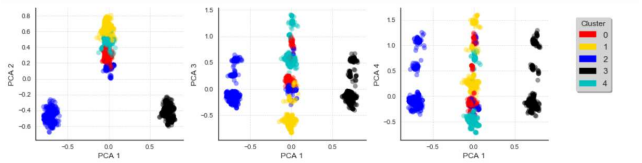


Figure 14: Biplot



a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. The initial matrix has large number of variables and hence, PCA is used for dimensionality reduction. From the Figure13 we can say that we need more than 100 components to explain 90% of the variance in the data.

Another application of PCA is that it sets the indication of ficlusterfi membership. Biplot is the best example that can be provided here to support this idea. Using biplot, we get the indication of number of clusters in a dataset. Below Figure 14 shows these on limited number of components (since it is only for visualizing cluster distribution). We can observe the groupings of points or clusters as expected.

Figure 15: Table

	CustomerID	InvoiceNo	Basket Price	categ_0	categ_1	categ_2	categ_3	categ_4	InvoiceDate
1	12347	537626	711.79	124.44	83.40	23.40	187.2	293.35	2010-12-07 14:57:00.000001024
2	12347	542237	475.39	0.00	53.10	122.59	130.5	169.20	2011-01-26 14:29:59.999999744
3	12347	549222	636.25	0.00	71.10	119.25	330.9	115.00	2011-04-07 10:42:59.999999232
4	12347	556201	382.52	19.90	78.06	41.40	74.4	168.76	2011-06-09 13:01:00.000000256
5	12347	562032	584.91	97.80	119.70	99.55	109.7	158.16	2011-08-02 08:48:00.000000000

Figure 16: Number of Purchases

	CustomerID	count	min	max	mean	sum	categ_0	categ_1	categ_2	categ_3	categ_4
0	12347	5	382.52	711.79	558.172000	2790.86	8.676179	14.524555	14.554295	29.836681	32.408290
1	12348	4	227.44	892.80	449.310000	1797.24	0.000000	0.000000	58.046783	41.953217	0.000000
2	12350	1	334.40	334.40	334.400000	334.40	0.000000	27.900718	23.654306	48.444976	0.000000
3	12352	6	144.35	840.30	345.663333	2073.98	14.301006	3.370331	53.725205	12.892120	15.711338
4	12353	1	89.00	89.00	89.000000	89.00	22.359551	19.887640	44.719101	13.033708	0.000000

7 EXPLORING CUSTOMER CATEGORIES

In the previous section, we have divided products in 5 clusters. We have added a dummy variable categ_product to indicate the cluster to which that customer belongs. Based on the clustering done on products we have created variables categ_0..4 which stores amount spent on each of the product category. And the categ_product variable which we have just created will have initial cluster assignment based on these variables. These can be further grouped on the basis of InvoiceNo as shown in Figure 15.

7.1 Subsetting dataframe based on Time

We have taken 12 months data for the analysis. This can be done on the basis of variable InvoiceDate present in the dataset. Using this data we have developed a model to characterize and anticipate the habits of customers using the site and this, we are doing it from the first visit.

In the previous section we have seen the basket price of each invoices. For further analysis we will combine these on the basis of customerID to analyze the number of purchases made by each customer as shown in Figure 16. A customer category of particular interest is that of customers who make only one purchase. So one objective may be, for example, to target these customers in order to retain them. In the dataset we have almost one-third of the customer base similar to this.

7.2 Categorizing Customers

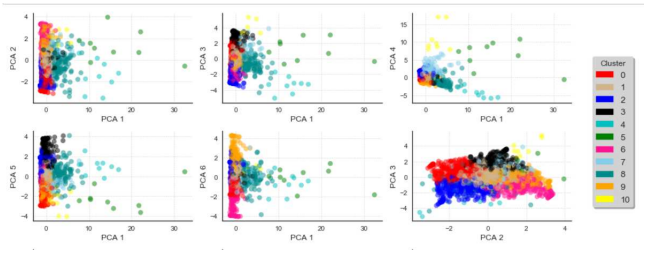
The information transactions per user is used for characterizing different types of customers. Because of different ranges of variations of different variables we have first scaled the data set. As done in the case of product categorization, we have again used K-means algorithm for cluster assignment.

Using the silhouette score, the optimum value of k comes out to be 11. The assignment of customers into different clusters is shown in Figure 17

Figure 17: Number of Purchase

	1	0	2	6	3	9	8	7	5	4	10
nb. de clients	1484	451	371	344	296	284	191	161	9	9	8

Figure 18: PCA



Now we will check validity of the cluster assignment using PCA and Silhouette plot as done in the case of product categorization.

7.2.1 ***PCA***. There is a certain disparity in the sizes of different groups that have been created. So we have validated it using PCA. From the representation shown in Figure 18, it can be seen, for example, that the first principal component allow to separate the tiniest clusters from the rest. More generally, we see that there is always a representation in which two clusters will appear to be distinct.

7.2.2 ***Silhouette Plot***. As with product categories, another way to look at the quality of the separation is to look at silhouette scores shown in Figure 19 within different clusters:

We can see that the different clusters are indeed disjoint.

8 CLASSIFICATION OF CUSTOMERS USING CLASSIFICATION ALGORITHMS

In the previous section, we have made different client categories. In this part we will adjust a classifier so that the consumers can be classified in different client categories. The main aim of this is to enable the Classification on the first visit of the customer. To do this, we have defined a class that will allow interfacing the common functionalities to the different classifiers. Since we are going to classify the client on the basis of his/her first visit, the only parameters that we take into consideration are the contents of the basket and not the frequency of visits or the variation in the basket price over a period of time. Once this is done, we have split the dataset into train and test sets. The classification algorithms which we used to do this are mentioned below.
Before we delve deeper into the Classification Algorithm, some important concepts that need to be addressed are Cross Validation, Bias, Variance, underfitting and overfitting of the model.

Variance Variance essentially means how much the models estimated from the different training sets differ from each other. It measures how much the predictions made for a

Figure 19: Silhouette Plot

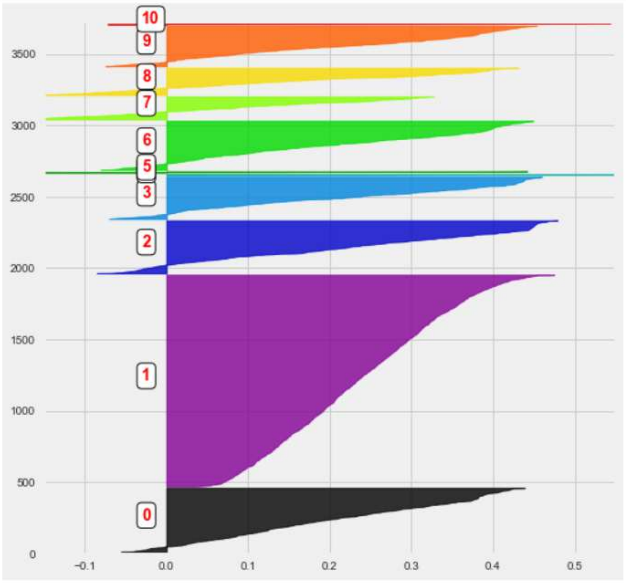
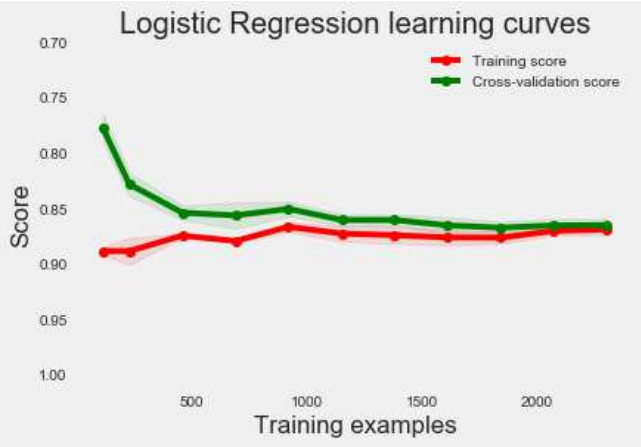


Figure 20: Logistic Regression Learning Curve



given point vary between the different realizations of the model [4]. When the training data tries to fit all the sample points to define the model, even the outlier data points, which are nothing but the noise, affect the model. Usually, the variance increases with increase in the complexity of the model.

Bias Bias essentially means how much the average model over the training sets differ from the true model. Bias usually occurs if the model is over simplified or if some inaccurate assumptions are made. Thus Bias increases with increase in the over simplification of the model [4].

Underfitting Scientific study of mental processes and behaviour.Underfitting occurs when the model is too simple to make relevant classification of the testing data [4].Thus when a model possesses high bias and low variance, we say there is underfitting of the model.

Overfitting Overfitting of a model occurs when the model is too complex and tries to fit in the irrelevant/outlier datapoints from the training set which is nothing but the noise [4]. Thus when a model possesses low bias and high variance, we say there is overfitting of the model.

Cross Validation Cross Validation is a technique for evaluating the predictive models by partitioning the original dataset into a training set to train the model, and a testing set to evaluate the model [11]. There are different ways to implement Cross Validation, the most effective of them all is the K-fold Cross Validation. In this method the dataset is divided into k subsets, out of which one is used as the test data and the remaining $k-1$ are combined together to form training data. This process is done k times, ensuring that every single sample in the dataset gets to be tested exactly one time and gets trained upon exactly $k-1$ times. The variance therefore gets decreased as the k increases [11].

8.1 Logistic Regression

Logistic Regression as mentioned before is a Supervised Learning method which does analysis on a dataset containing two or more independent variables for determining the outcome. This outcome, i.e the dependent variable, is binary in nature, meaning it can have only two possible outcomes [9]. Multinomial Logistic Regression as the name suggests, generalizes the Logistic Regression to multiple classes, meaning the model can be used to predict the probabilities of the different outcomes of a categorically distributed dependent variable [17]. The goal of a Logistic Regression model is to determine a fitting model which best describes the relationship between the dependent variables (output variable) and a set of the input independent variables. Logistic Regression generates the coefficients along with the standard errors and significance levels of the below equation for predicting the logit transformation of the probability of presence of the characteristics of interest in a given sample example [9].

$$logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + ... + \beta_k X_k$$

where p is the probability of the presence of a characteristic of interest. and $logit(p) = log(p/1-p)$ In logistic Regression, the goal is to choose the parameters β in such a way that the likelihood of observing the new sample values is maximized [9].
In the Python code, we have imported the module 'linear_model' from the 'sklearn' package to perform Logistic Regression by using the function 'logistic_regression'. And we have taken the $k = 5$ for k-fold cross validation. While performing Logistic Regression, we created an instance of the Class_Fit class and then ran the model on training data and see how the predictions are made as compared to the real values. The learning curve graph is as shown in Figure 20.

As we can see from the Figure 20, when the number of training examples increases, the cross-validation and train curves almost converge towards the same limit suggesting that the model has low variance. Thus we can say that model is not suffering from overfitting. Also one point to note is that the accuracy is high, which

means that the model has low bias, thus suggesting that it does not under-fit the data. The precision which we got from running the Logistic Regression model on the training data is 88.78%.

8.2 K Nearest Neighbours

KNN is a non parametric algorithm which means that there are no underlying assumptions that are made on the data. Also it is a lazy learning algorithm meaning that it does not do any generalization by using the training data. All the training data is needed during the testing phase [13].

KNN makes predictions using the training dataset directly. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. For regression this might be the mean output variable, in classification this is be the model (or most common) class value. To elaborate on this, KNN makes predictions using the training dataset directly. These predictions are made for a new sample by going through the entire training set to find k such samples which are most similar or which are the 'neighbors' of the the new instance. Once these k instances are found out, the output variable corresponding to these is summarized and in case of Classification, it gives a class value to which the new instance belongs. The k 'neighbors', i.e., the most similar instances from the data set are found by using the distance measure- k such instances whose distance from the new instance is the least [13]. There are many distance functions which can be used, the most popular being the Eucledian distance function, the formula for which is given by:

$EuclideanDistance(x, x_i) = sqrt(sum((x_j - x_i i)^2))$

where x is a new data point and x_i is an already existing point [13].

The optimum value of K can be found by algorithm tuning, i.e. running the algorithm over several values of k and finding out and then figuring out for which k the algorithm gives the best results [14].

The output, i.e the class of the new sample can be calculated as the class which has the highest frequency from the k neighbours. Thus, each of the instances votes for their own class and the class which gets the maximum votes is taken as the prediction value [14].

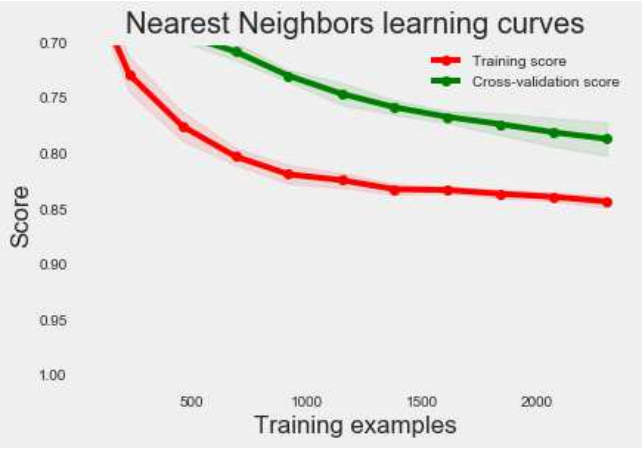
In Python, the 'neighbors' library is imported from the sklearn package which performs the KNN classification through the Kneighborsclassifiers function. The parameters that are used are 'n_neighbors' which represents the number of neighbors to use, in our case we have used the np.arrange method to give sequence from 1 to 49. Also, we run the model using the K-fold Cross Validation with the value of $k = 5$. Once the model is run, we have drawn the learning curve graph which is as represented in the Figure 21.

The precision which we got from running the KNN model on the training data is 80.33%.

8.3 Random Forest

As the name suggests, Random Forest is an ensemble classifier which consists of many classification trees. An ensemble classifier is a multiple classifier algorithms, decision trees in the case of Random Forests, and the final output is the combined output of the all the classifier algorithms. In our case we will be using Random Forest

Figure 21: KNN Learning Curve



Algorithm for classification of the clients into different categories. A Random Forest grows many trees. For classifying a new object from an input vector, each tree in the forest gives a classification and vote for a particular class. And the forest then chooses the class having maximum number of votes over the other classes [1].

The question here that needs to be addressed is, how does the growth of a tree happen?

Each tree is grown as follows:

If the training set consists of N cases, then N cases are sampled with replacement from the original data. This is the training set for growing a tree. Thereafter, a number m¡¡M which is the number of input variables is taken such that the best split obtained on these m is used to split the node. The value of m is constant throughout the forest-growing. Each tree is allowed to grow to the fullest possible extension [1]. In Python, the 'ensemble' library is imported from the sklearn package which performs the Random Forest classification through the RandomForestlassifier function. The parameters given to this function are criterion, n_estimators and max_features. The criterion is used to measure the quality of the split. The Gini is for measuring the Gini impurity and Entropy is for information gain. The max_features are the number of the features that can be chosen when looking for the best split. For 'sqrt', the number of maximum features chosen are square root of the number of the features and for 'log', it is log of the number of the features. And the n_estimators is the number of trees in the forest.Once the model is run, we have drawn the learning curve graph which is as represented in the Figure 22 .

The precision which we got from running the Random Forest model on the training data is 90.17%.

8.4 Gradient Boosting Classifier

AdaBoost Classifier, short for Adaptive Classifier is another example of ensemble classifier. It is a general ensemble method which creates a strong classifier by combining the outputs of the weaker learning algorithms into a weighted sum to finally provide the output of the *boosted* classifier. This is done by building one model from the training set and then building a second one which attempts to rectify the errors from the first model and so on until either the

Figure 22: Random Forest Learning Curve

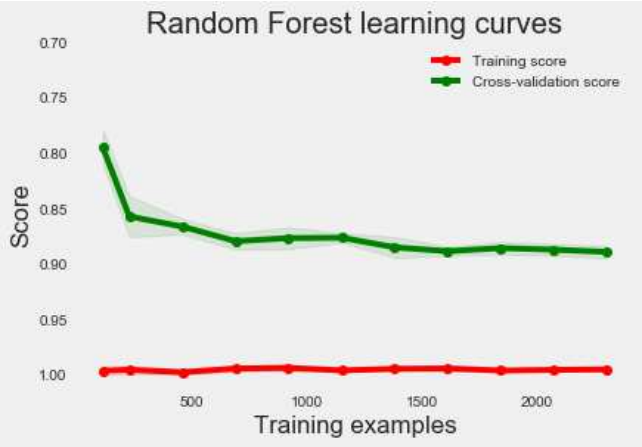
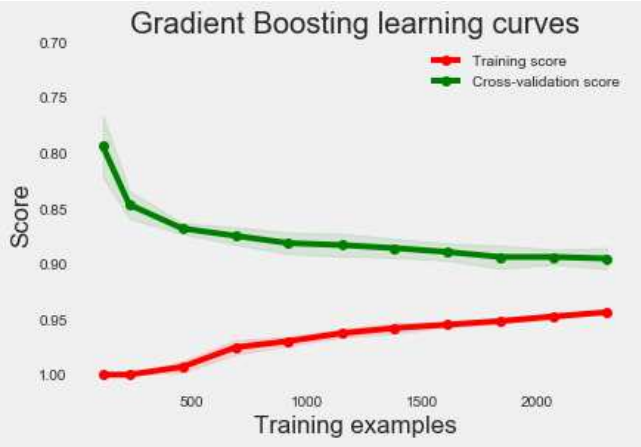


Figure 23: Gradient Boosting Learning Curve



limit of maximum models that can be added is reached or the training set is predicted accurately. AdaBoost is an adaptive algorithm, meaning that the weak learning algorithm can be tweaked to create a stronger classifier [6].

The Adaptive Boosting algorithm was recast into a statistical framework. "Arcing is an acronym for Adaptive Re-weighting and Combining. Each step in an arcing algorithm consists of a weighted minimization followed by a re-computation of [the classifiers] and [weighted input] [6]" This framework is called as Gradient Boosting.

Gradient Boosting involves three elements namely:

A loss function The selection of the loss function depends on the problem at hand. For example if it is a regression algorithm, then squared loss functions are used and if it is a classification algorithm then logarithmic functions are used. The main aim of the algorithm is to optimize the loss function.

A weak learner Regression trees are used as the weak learners in the Gradient Boosting Algorithm since they can output real values for splits which can be added together and the residuals in the predictions can be corrected. The weak learners are used for making predictions. These trees are constructed in a greedy manner usually up to 4-8 levels.

An additive model The additive model is made so as to add the weak learners to minimize the loss function. The trees are added one at a time with no changes to the existing trees in the model. A gradient descent model is used to reduce the loss when adding trees first by parameterizing the tree and then by modifying the parameters of the tree and moving in the right direction by reducing the loss in residuals. This approach is called Functional Gradient descent [6].

This framework was further developed by Friedman and called Gradient Boosting Machines. Later called just gradient boosting or gradient tree boosting [6].

In Python, the 'ensemble' library is imported from the sklearn package which performs the Gradient Boosting classification through the GradientBoostingClassifier function. The parameter given to this function is n_estimators which is the number of boosting stages to perform. Gradient boosting is fairly robust to over-fitting so a large number results in better performance. Learning curve is shown in the Figure 23.

The precision which we got from running the Gradient Boosting model on the training data is 90.86%.

Now that we have the results of all the models, we can combine them using VotingClassifier method that we imported from the sklearn package to improve the classification model. Since we have already found the best parameters for each of the classifiers, we have adjusted the parameters of the classifiers accordingly. So now the best parameters are taken and merged to define a classifier which we then trained on the data. When we created a prediction on this, we got the precision value as 90.44%.

9 TESTING THE PREDICTIONS

Until now we have done all the analysis on the data from the first 10 months. After this we test the model on the set_test dataframe which contains the data of the last two months. The regrouping of the data is done according to the same procedure that we followed while regrouping the training data. But now we have to take into consideration the time difference in between the two datasets and the count(the total number of visits which the client made to the website) and sum(total amount that he/she spent) variables so that we have an equivalence in between the training set and testing set. The dataframe so obtained is now converted to a matrix and we retained only those variables that define the category to which the clients belonged. And just like on the training dataset the method of normalization was called, to maintain consistency, the same method is called on the test set as well.

Each row of the matrix obtained now represents the buying habits of the customers. Now all we have to do is to define the category to which the customer belongs by using these habits. The important point to note here is that this is just the test data preparation step by defining the category to which the consumer belongs for a period of two months through the variables count, min, max and sum. Thus this step *does not* correspond to the classification step itself. The classifier that we defined in the step 5 uses variables that were

defined from the client's first purchase.
So now, we have the data available for two months, and through that we can define the category to which the consumer belongs. The predictions now obtained by running the classifiers on test data can be tested against these categories. The instance of the k-means clustering method that we used in the Customer Categories section is used to define the category to which a client belongs. This contains the predict method which will calculate the distance of the consumers from the centroids of the 11 categories that we deduced, and the category which is closest to the clients' buying habits will define his/her category. Thus all we need after this for the execution of the classifier is to select the variables on which it acts, i.e. on mean, cat_0, cat_1, cat_2, cat_3 and cat_4. After examining the predictions of the different classifiers, we get precision scores as shown in Table 1.

Table 1: Algorithms with their Precision Scores

Algorithm	Precision(%)
Logistic Regression	72.99
KNN	68.44
Random Forest	75.93
Gradient boosting	75.74

And now, like we did in the Section 5, we will use the voting classifier method to merge the results obtained by these individual classifiers and see whether they combined result is better than the individual. It turns out that it is. We get the precision rate for the combined classifier to be 76.48% for the test data set. This concludes the analysis phase.

10 CONCLUSION

E-commerce is one of the emerging fields for Data Analysis since a lot of data gets generated every day at a break-neck speed in many different formats. To sustain in such a business, a very robust and extensive data analysis is needed to keep up with the ever changing markets by implementing different marketing strategies. And since the whole business revolves around the customers, they form the most important aspect of the analysis. We have tried to achieve Customer Segmentation on the basis of the purchasing patterns and frequency of client visits to their online portal. The dataset on which we performed analysis provided details on the purchases made by the consumers over a period of more than a year. Every entry in the dataset contained the purchase of a particular product on a given data by a particular customer. Out of the 591909 entries made in the dataset, approximately 4000 different consumers are present. From the information available for each consumer, we decided to go ahead with Customer Segmentation analysis by developing a classifier that predicts the type of purchase a consumer would make and his/her frequency of visits to the E-commerce website.
In the first step of this classification, we found out the different products sold by the company, and then classified the products into 5 categories of goods by using K-means clustering. In the second step we performed the classification of the customers on the basis of purchasing habits in the first 10 months. The customers were classified into 11 categories on the basis of the types of products they usually bought, the number of visits they made to the website and the amount for which they shopped over a period f 10 months. Once we had the categories of the the consumers, we performed training of the data of the first 10 months using different classifiers namely Logistic Regression, Random forests, KNN and Gradient Boosting algorithms to classify the consumers in these 11 categories, on the basis of their first purchase. The classifiers were based on these variables: the total price of the current purchase and the percentage of the amount spent in each of the 5 product categories. Once the customers were classified in the 11 categories, the quality of the data set was tested on the remaining two months of the dataset. This was achieved in two steps. In the first step, we assigned the category to which each customer belonged to, and then the classifier predictions were compared against these categories. And then we combined the results of the various classifiers by using the Voting Classifier method. The model performed with a 76.48% of precision, that is 76.48% of the times the clients were awarded the right classes.
One bias which we did not consider while doing the analysis is the seasonal fluctuations, like festive and seasonal sales. Since at these times the sales of products may rise and just before and after the sale duration, the sales may drop. Thus the purchasing habits of customers are dependent on the time of the year as well. Hence the seasonal effects may cause the actual sales in the last two months to be quite different from the ones which we extrapolated from the first ten months to the last two months. For overcoming such biases, it would be beneficial if the data were of a larger size and covered a larger period of time.
Knowing the type of a customer is critically important for en E-commerce business. By doing so, the store owners can provide personalized services to the customers, which will yield higher customer satisfaction. Customer satisfaction is directly proportional to the loyalty of the customers, thus Customer Segmentation and Personalization can help the company in increasing their brand name. Knowing the preferences and choices of customers also helps in catering to those needs of the customers which they may not be aware of in the first place. Thus by knowing the purchasing patterns of the customers, we can provide them with tailored suggestions, which can even increase the revenues of the company. Thus through proper implementation of the business strategies and marketing activities, which are motivated by a thorough Analysis of the data available can help the company in attracting loyal customers, increasing the revenue and establishing a better brand value.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Dr. Gregor von Laszewski for giving the opportunity to work on this project. The author would also like to thank the Associate Instructors of the class for their help and for answering questions on Piazza which helped everyone.

REFERENCES

[1] Jason Brownlee. 2016. https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/. (2016). https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/

[2] Jason Brownlee. 2016. Supervised and Unsupervised Machine Learning Algorithms. (2016). https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
[3] Justin Butlion. 2015. An Introduction to Analytics for Ecommerce Websites. (2015). https://blog.kissmetrics.com/intro-to-ecommerce-analytics
[4] Scott Fortmann-Roe. 2012. Understanding the Bias-Variance Tradeoff. (2012). http://scott.fortmann-roe.com/docs/BiasVariance.html
[5] Prashant Gupta. 2017. Decision Trees in Machine Learning. (2017). https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
[6] Antonino Ingargiola. 2015. What is the Jupyter Notebook? (2015). http://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html
[7] Lacie Larschann. 2017. 7 Powerful Applications of Machine Learning in E-Commerce. (2017). https://www.granify.com/blog/powerful-applications-of-machine-learning-in-e-commerce
[8] EILEEN MCNULTY. 2015. WHATfiS THE DIFFERENCE BETWEEN SUPERVISED AND UNSUPERVISED LEARNING? (2015). http://dataconomy.com/2015/01/whats-the-difference-between-supervised-and-unsupervised-learning/
[9] Medcalc. 2017. Logistic Regression. (2017). https://www.medcalc.org/manual/logistic_regression.php
[10] Sunil Ray. 2017. Understanding Support Vector Machine algorithm from examples (along with code). (2017). https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
[11] Jeff Schneider. 1997. Cross Validation. (1997). https://www.cs.cmu.edu/~schneide/tut5/node42.html
[12] Granner Smith. 2017. Big Data: Making It Big For E-Commerce Retailers. (2017). http://www.digitalistmag.com/customer-experience/2017/04/28/big-data-making-it-big-for-e-commerce-retailers-05049637
[13] Saravanan Thirumuruganathan. 2010. A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm. (2010). https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/
[14] Saravanan Thirumuruganathan. 2010. A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm. (2010). https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/
[15] Tracey Wallace. 2017. Ecommerce Trends: 147 Stats Revealing How Modern Customers Shop in 2017. (2017). https://www.bigcommerce.com/blog/ecommerce-trends/
[16] Wikipedia. 2017. E-commerce. (2017). https://en.wikipedia.org/wiki/E-commerce
[17] Wikipedia. 2017. Multinomial logistic regression. (2017). https://en.wikipedia.org/wiki/Multinomial_logistic_regression

Big Data Analytics in Product develop management

ZhiCheng Zhu
Indiana University Bloomington
936 S Clarizz Blvd
Bloomington, Indiana 47401
zhuzhic@iu.edu

ABSTRACT

The success of a new product is to a large extent due to whether the producer making efficient and accurate strategies between the different stages of a product lifecycle. Big Data analytic techniques have the potential to improve the efficiency of product development, making accurate product strategy, channel strategy, pricing strategies and promotional strategies in the different period of a product lifecycle. In this project, I will try to figure out how the data affect the decision making and try to use the relation between different twitter account to describe the potential of the Big data also I will describe decision trees and PageRank and how do these tools produce a good market segmentation and market positioning for a product.

KEYWORDS

i523, hid229, Big data, Product Development, Technology

1 INTRODUCTION: BIG DATA

The advent of technology has resulted in virtually all industries and organizations collecting large volumes of data. The data collected results from diverse source, which include product sales, customer information, historical industry data, and employee information just to mention a few. Computers and the internet in particular have made it easier to collect data of different kinds because they make it easy to create, store, transfer, and analyze data. As a result, data has become a critical asset for many organizations and corporations in their bid to control the markets of the products or services they offer consumers. Furthermore, According to Arora, big data also refers to large and complex dataset (13). This means that it is virtually impossible to use traditional processing applications to organize and analyze this data. Therefore, there are various challenges associated with big data due to its large volume and complexity [1]. These problems include data capture, data storage, data analysis, sharing of the data, making searches on the data and the privacy of the information [**?**]. Information increasingly becomes an important factor in determining the success of a product. A few years ago, manufacturing and the Internet have still belonged to two different separated industries, But as the mainstream consumer groups changed from old generation to Millennial generation, and the use of computers and the establishment and application of the Internet have produced a violent shock to the traditional way of product development, thus resulting in a new product development strategy. I have to say that the relation between the manufacturing and the Internet are getting closer and closer. One of the major changes might cause by using big data as a technique to develop right product and make effective promotion strategy.

information.png

Figure 1: Difference between various generations [10]

2 EXAMPLES OF BIG DATA

For example, organizations with an online presence such as online market places collect data from their clients. This includes different types of data such as customer information, purchases, and time spent on the website. For very large organizations (such as Amazon or eBay) dealing with thousands and probably millions of customers daily, this type of data soon becomes voluminous and difficult to analyze. For effective analysis, such data needs large physical storage space and organization in ways that will make it usable and of benefit to the organization [1]. The main solution to this problem is the use of a data warehouses since traditional databases are too basic to handle the complexities involved with big data. Data warehouses provide the much-needed processing power needed in handling and analyzing big data [1]. Another issue is Outdated decisions and information are bound to create disadvantages when competing with other companies, but the traditional data processing system is obviously not suitable for the era of big data. Corporate decision-makers must adopt new technologies to face the changes of the times and customer. For example, Hadoop technology is a new and widely accepted technology.

- "Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs[9]."

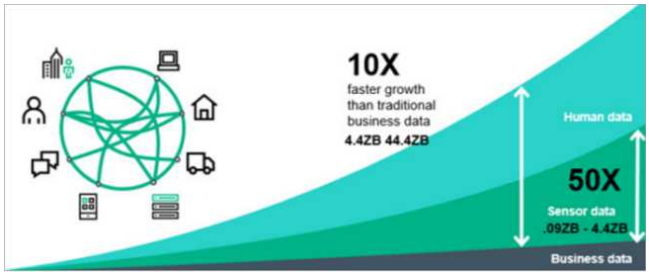


Figure 2: The Exponential Growth of Data [6]

2.1 Application of Big Data in Product Development

how to clean the data and find useful data more quickly in product development is even more important.Quickly identify the characteristics of the target customers and their possible needs for a variety of products. base on the analysis, we can develop some products that more suitable for marketplace needs, or we can be more accurate when we try to find our potential target customer. According to different consumer behavior, we can design different software, for example, during the financial turmoil we found that lots of users are price sensitive, we can design a kind of software that can provide local discount merchandise information in real time.If the target customer is a group which wants higher quality of life, we can push more high value-added product. One of the most vital areas where big data finds use is in product development management. The product development process allows for the design and release of new products into the market. Product development also involves processes such as forecasting, planning, and marketing of new products. The process adopted ensures that first the product developed meets a certain need in the market. Second, the process certifies that the price set reflects the amount consumers are willing to pay. Lastly, it guarantees the organization is making the product in such a way that it will be able to reach the targeted market. Typically, an organization will collect market data from various systems.

- For instance, transaction-processing systems provide invaluable data to organizations [11]. An example of a company that uses big data is the online market place Amazon. The market place has thousands of products stocked by the company and by third party sellers using the platform to sell their products. The checkout system in such an organization will collect very important data such as the products sold, the customers buying the products, the time of purchases and such details [14]. The details of the customers will probably include the age and gender of the customer. When the company analyzes such data, it provides the management with a vital insight into the businessfi activities and performance. Such as SEM analysis, SEM analysis is a research and analysis methods which focus on customer satisfaction. This is a good way to make a classification for our users. One of the most typical examples is some recommender systems such as Pandora use songs or artist properties to create a radio station, all of these songs and artist have similar attributes. User feedback is used to adjust the content of the radio and recommend some music which is more attractive to the listener.

3 ANALYSIS OF CUSTOMER NEEDS AND MARKET DEMAND

3.1 Anticipating Customer Needs for New Products

For instance, using the data collected from the checkout system, the company can tell what type of products are likely to be bought and by which customers [14]. Using information about customers contained in other systems such as customer relationship management (CRM) systems, the company can get information about the people likely to purchase a certain product, at what time they are likely to make the purchases, and the other goods they are likely to purchase with the products [14]. The company can then use such information to make decisions on which products to stock, what price to sell them, which products to suggest to clients as they are making purchases and the time of day, week and even year that such products are in demand. For instance, Walker points out that big data plays an important role in helping Amazon launch new products. In particular, the data collected by the organization on books played a central part in the establishment of the Amazon Kindle product, which is highly successful in the market [14].

3.2 Inventory Management

Therefore, with the help of big data, product developers and manufacturers are able to ensure that the products needed by the customers are available in the right quantities and at the right times [4]. For example, a person purchasing a large flat screen television is also likely to purchase wall brackets for mounting the television set. The company can also get information on what brand and size of TV sets are in demand. The company can then ensure that it stocks such products when they are in demand. Because of the increased competition, profit margins can be very small. This requires the company to ensure efficiency to ensure that it avoids issues such as dead stock, which have negative impacts on profitability. Getting the right amount of products is critical as it ensures that the company meets all the demand and there are no excess items, which are a cost to the organization [11].

3.3 Anticipating Customer Demands

hub node with high in-degree.png

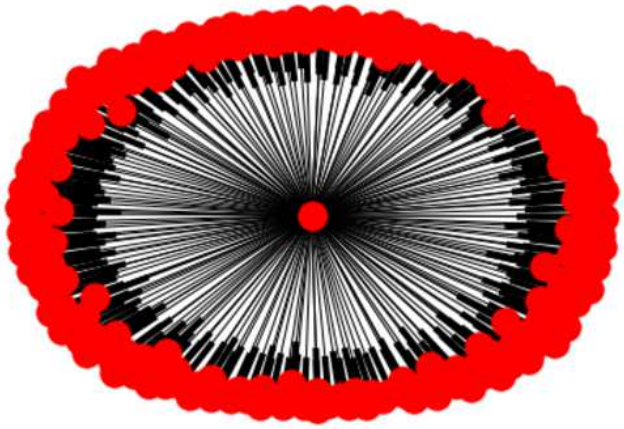


Figure 3: Hub Node in network

In addition, using big data, product developers able to anticipate demand for new products and it can therefore embark on developing such products and ensure that customers get the products. For example, Facebook the social media giant employs big data in making decisions on how to present its services for clients [8]. For

instance, the company collects data on user activity on its platform [8]. Using this information, the company is able to present an interface customized to a specific user. Such data also finds use in targeted advertising, which is one of the key ways in which such companies make majority of their revenues. In 2013, Morabito pegs the revenue drawn by Facebook from advertising at almost 8 billion dollars [8]. The use of big data by the company has made it able to forecast the needs of the market and provide users with products and services that the customers find useful. This has resulted in the company becoming the worldfis most used social media site and maintain its advantage over its rivals, some of which the company has ended up buying out. what is worth to mention is that some Big Data technique can help these company decide which user are more worthy of advertising. According to what I got from my programming, some people with a high degree centrality in their tweet network has a higher value than other people with a sparse network. Because the dense network can spread the information faster with a lower cost.

4 MANAGEMENT OF THE PRODUCT DEVELOPMENT CYCLE

In product development management, a product passes through various stages in the lifecycle. The main stages are product research and development stage, the introduction stage, the growth stage, maturity stage and the decline stage [13]. Big data can help to ensure that a company develops a product that maximizes returns in each of the main stages. Once the product is in decline, the company also has the information needed to make decisions on the products it will introduce that will ensure that the company maintains its competitive advantage on the market.

4.1 Research and Development

In the product research and development stage, an organization will collect large amounts of data from the public, from its own internal systems or data from third parties [2]. Such data will contain information on the type of products favored by the market. The company will analyze this data and come up with information that its management can use to improve the overall decision-making process [2]. Using the example of the online retailer Amazon, the company can use big data to carry out research and development processes to determine that many customers want an easier way of making payments. The organization can obtain such information directly from customers, as well as, based on received feedback. For example, the company could notice trends in customers who browse for items but abandon the purchase when they are required to pay. This may be because the payment methods offered by the company are difficult or customers are not confident in them. Using such information, the company can decide to introduce alternative ways to make payments, which are easy to use.

4.2 Introduction Stage

Big data also proves to be very useful in the introduction stage. For instance, a company selling warm winter clothes would not be successful is such items in summer. The main reason for this is that at the time, customers are not in need of the products. Organizations can derive such information by analyzing data on the type of purchases made by customers in different periods of the year.

4.3 Growth Stage

Furthermore, the growth stage is also vital for any product in the market. During this stage, the product introduced by the company gains a foothold in the market [13]. During this stage, more and more customers learn about the product and they are likely to make a purchase. Using big data can be advantageous to a company as it can help it to organize its marketing appropriately. The purpose of this is to ensure that the products reach the widest coverage [2]. Big data can also help the company in making decisions such as which locations to introduce the product.

4.4 Maturity Stage

Similarly, in the maturity stage of the life cycle of a product, the product has already gained a foothold in the market and its growth slows down [13]. Big data can help the company to make decisions that will enable the product to maintain growth during this stage for the longest time.

4.5 Decline Stage

The decline stage occurs after a product has been in the market for a while and probably newer technologies have become available reducing the usefulness of the product to the market [3]. Big data can help the organization in this stage of the product life cycle by ensuring that the product remains available for people who still need the product. For example, smartphone use has overtaken the use of feature phones in the market. Feature phones offering simple functions have therefore reached the decline stage of the cycle. Despite the noted decline in the global environment, in many developing countries feature phones remain in demand due to factors such as battery life and cost. Using big data, a company manufacturing feature phones can understand its market and therefore be able to ensure that supply to such markets remains available. Moreover, by the decline stage of the product life cycle, the company dealing with the product is likely to have begun the research into the next product that it will offer to clients that will meet their emerging needs. Big data, which will consist of data collected from the lifecycle of the previous product, will prove to be very useful in the development of the new product. As the name suggests, this is a lifecycle and the process should ideally continue indefinitely for the company. This ensures that the company is always ready with a new product that is able to meet the demand of clients in the market. Constant innovation is vital in the modern market where competition is very high. Nokia Corporation, once the largest manufacturer of cell phones is an example of a company that failed because of not innovating constantly. In the mid-2000s when competing firms introduced the first smartphones, Nokia was the dominant player According to Robbins, Rolf, Ian, and Mary by 2007 Nokia controlled 40 percent of the global mobile phone market [3]. However, because of poor forecasting, the company continued to manufacture its previous phones and soon it lost its market share to other companies manufacturing smartphones such as Apple [12]. The use of big data in making decisions during the product lifecycle

can help a company to avoid making such mistakes, which may prove to be catastrophic. From the analysis above, it is evident that during the different stages of the product lifecycle, there are different strategies that organizations can employ. For example, product manufacturers and sellers can utilize different marketing strategies during the different stages of a productfis lifecycle. The marketing strategy that is useful in the introduction stage may not be as effective during the growth stage of the product. The production strategy employed during the introduction stage may be very effective but the same strategy when employed during the growth stage or other subsequent stages may not be very effective. Big data therefore helps an organization to make the best decisions on different strategies based on the information obtained from analyzing the data. This ensures that the strategies selected for various activities during the different stages of the productfis life are the most appropriate which ensures that the company is able to gain the most benefits from its products.

5 BIG DATA ANALYSIS METHODS

In big data, organizations can adopt different analytic methods in order to extract useful information. Data in its raw form is not very useful to an organization. Analysis of the data ensures organizations come up with patterns emerging in the data with the aim of improving the decision making process. The analytical method chosen for analysis of big data depends on various factor such as the type of data available (e.g. qualitative or quantitative), the amount of data available and the result desired. Among the most common analysis methods in big data are Decision tree analysis, PageRank, and kNN algorithm.

5.1 The Decision Tree Analysis

The decision tree analysis is a method of analyzing data that uses a graph in the shape of a tree, hence the name. In this method of analysis, a decision maker considers a decision and its resulting consequences [16]. These consequences include the chance of the consequence occurring, the costs involved when the consequence occurs and how useful the consequence is for the organization. The initial decision represents a node and the possible consequences are the branches [5]. Each consequence then becomes another node and the tree represents consequences in further branches [5]. When using a decision tree to make a decision, the decision maker selects the path that is most likely to lead to the desired solution. Big data requires the analysis of large amounts of data. A decision tree algorithm will analyze the data available and present a decision tree with all the possible paths (the connections between nodes and branches) based on the information available [5]. For example, the decision to select a particular marketing strategy will show the chance of it achieving the result, the costs involved, and the utility to the company, depending on the information obtained from the large data sets. This will therefore present different paths based on the different strategies chosen. Usually, the decision maker will make a decision by selecting the path of least resistance. Usually this path contains the best attributes needed by the organization. For instance, one path may be more costly than a different path based on the strategy chosen but they eventually lead to the same result. The path that costs less will therefore be best suited for the organization since it will allow it to achieve its objective more efficiently. Lastly, because making a decision will lead to different consequences (which in themselves require other decision) the path that best suits the organization results from the final objective through the path of branches and nodes that best meet the needs of the organization [5].

5.2 PageRank

PageRank is an algorithm named after Larry Page who was a co-founder of the search giant and technology company Google. This algorithm finds use in ranking the search results on the search site [17]. This algorithm works by counting the number and quality of links to a webpage. This algorithm helps to determine the quality of the information on the different WebPages with information related to the search queries entered [7]. Based on the quality and number of links pointing to a page, the algorithm is able to determine the quality of the webpage [17]. This then results in the page receiving higher raking in search results. PageRank is very important as it helps the search giant to present the most relevant answers to queries made on its site.

For instance, one webpage may contain very many links concerning the search query but the information contained on the webpage may not be of the highest quality. This means that another webpage with higher quality information even with fewer links can still rank higher than the other webpage [7]. This algorithm works by using data generated from previous searches with similar queries. Such an algorithm makes extensive use of big data to ensure that it presents the most appropriate results for a person making a query.

5.3 k-NN algorithm

The k-NN algorithm is an algorithm used in pattern recognition [15]. It finds use in both classification and regression. This is the simplest form of machine learning as it uses an approximation of values nearest to the value under analysis. For instance, in classification, the desired output is class membership. The algorithm achieves this by looking at the nearest neighbors of the value under inspection. The value receives assignment to the class to which most of its nearest neighbors belong. In regression, the desired output is typically the property value of the object under study. This is obtained by getting an average of the values of the objects that are the immediate neighbors of the object being inspected. This means that the nearest neighbors to an object contribute more towards its value than objects that are located further away from the object inspected. This means that using these algorithms, the values of an object can be predicted accurately, which helps in making complex decisions easier based on the immediate results expected [15].

6 BIG DATA ANALYTICS AND DECISION MAKING

6.1 Big Data and Competitive Advantage

Data as previously mentioned is one of the most important assets for any business. This data needs to be analyzed so as to come up with usable information that help the management of the organization to make decisions that are likely to succeed in the market.

Because of the increased competition in the modern business environment, many organizations have employed big data to help them make decisions such as how to arrange items in stores, what items to stock, the prices that are best for the market and such decisions [14]. Although these decisions might look simple, it is very important for an organization to get them correct. Miscalculations made in such decisions could lead to losses including financial and market share losses. This is because modern customers want value for their money. This means that the organization must offer the best possible services at the lowest cost. Furthermore, this makes them attract more customers for their products or services and in the process enable them to make a higher margin. Because most competitors will use some form of data to make their decisions, it is very easy for an organization or company to lose its competitive advantage to its competitors [14]. Many organizations consider making accurate decisions in an efficient way a critical aspect of their competitive advantage. For instance, a company can release a product before the competition releases their version. Customers will buy the product already in the market allowing the organization to gain a market advantage over the rivals who have not released their product. Big data has made it possible for such organizations to obtain patterns from extremely large data sets, which are more accurate and therefore likely to produce accurate decisions [14]. Another strategy can help our product to attract more customers is using the internet hub node to spread the information about our product and advertising on the network, the network hub node is a node with a number of links that greatly exceeds the average. for example, Donald Trump's twitter is one of a most popular node in twitter network. If someone can lobby Trump to promote the product I believe we will have increasingly more customer growth. These hubs play the same role such as what Macy's used to play. Macy's used to be a big node for people's Holiday purchase. because it attracts most of the people living in that area to enter the store. The famous twitter account attract their follower because they have a fancy lifestyle or fulfill some value their follower admired. all of these are helpful for product promotion and the spread will have a more incredible efficiency and effectiveness.

7 CONCLUSION

In conclusion, despite the use of big data already being widespread in the business world, its importance will continue to grow with time. This is because of the large number of devices that are now generating data. The internet of things has resulted in a situation where even everyday appliances connect to the internet and they have the ability to collect large amounts of data. This results in more data that organizations can analyze to discover patterns in the market that organizations can exploit. Organizations are also overcoming the challenges facing big data with the collection of data now largely automated from different systems internal and external to the organization. Moreover, specialty companies have come up with the sole purpose of collecting and analyzing market data. Manufacturers are offering solutions to the technological challenges involved in big data by developing larger storage devices that are able to store increasing amounts of data efficiently. The security of such information has seen significant improvements with more secure communication and storage channels. The use of big data is also set to increase as current analytic methods become efficient. New analytic methods are also likely to be developed that will ensure that the data collected from independent systems and the market are analyzed better in order to develop information that can be acted upon more readily in the market. This point to a future where big data will be more important than ever in ensuring that companies come up with products and strategies that enable them to be more competitive in the market and therefore increase their chances of survival in the market.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper as well as TAs' helpful suggestions on this paper..

REFERENCES

[1] Ritu Arora. 2016. Conquering Big Data with High Performance Computing. (2016).
[2] Ron Basu. 2016. *Managing Projects in Research and Development.* Abingdon, Oxon : Routledge, 0.
[3] Louis E Boone. 2013. *Essentials of Contemporary Business. Hoboken.* New Jersey : John Wiley & Sonse, 0.
[4] Jeanne G. Harris Robert Morison Jinho Kim Davenport, Thomas H and D J. Patil. 2014. *Analytics and Big Data.* Boston Massachusetts : Harvard Business Press, 0.
[5] Mohammed Guller. 2015. *Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large-Scale Data Processing, Machine Learning, and Graph Analytics, and High-Velocity Data Stream Processing.* New York: Springer, 0.
[6] insideBIGDATA. 2017. The Exponential Growth of Data. Web page. (February 2017). https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data
[7] Amy N Langville and Meyer C D. 2006. *Google's Pagerank and Beyond: The Science of Search Engine Rankings.* Princeton, N.J: Princeton University Press, 0.
[8] Vincenzo Morabito. 2015. *Big Data and Analytics: Strategic and Organizational Impacts.* 0.
[9] The power to know. 2007. What is hadhoop? (Nov. 2007). https://www.sas.com/en_us/insights/big-data/hadoop.html
[10] Media Insight Project. 2015. How Millennials Get News: Inside the habits of Americafis first digital generation. Web page. (March 2015). https://www.americanpressinstitute.org/publications/reports/survey-research/millennials-news/single-page/
[11] Foster Provost and Tom Fawcett. 2013. *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking.* Sebastopol, CA: O'Reilly Media, 0.
[12] Rolf Bergman Ian Stagg Robbins, Stephen and Mary Coulter. 2014. *Management Vs.* Sydney: Pearson Education Australia, 0.
[13] John Stark. 2015. *Product Lifecycle Management: Volume 1.* Cham: Springer, 0.
[14] Russell Walker. 2015. *From Big Data to Big Profits: Success with Data and Analytics.* NY: Oxford University Press, 0.
[15] Jia Yingmin and Du Junping. 2017. *Proceedings of 2017 Chinese Intelligent Systems Conference: Volume I.* S.l.: Springer Verlag, 0.
[16] Jie Lu Zhang, Guangquan and Ya Gao. 2015. *Multi-level Decision Making: Models, Methods and Applications.* Heidelberg: Springer, 0.
[17] Albert Y Zomaya and Sherif Sakr. 2017. *Handbook of Big Data Technologies.* Switzerland : Springer, 0.

Big Data in Safe Driver Prediction

Jiaan Wang
Indiana University Bloomington
3209 E 10 St
Bloomington, IN 47408
jervwang@indiana.edu

Dhawal Chaturvedi
Indiana University Bloomington
2679 E 7th St
Bloomington, Indiana 47408
dhchat@iu.edu

ABSTRACT

For years, people have been trying to reduce their automobile insurance bills. Insurance companies claim that price will be reduced for good drivers and raised for bad ones. However, inaccuracies in their data predictions lead to the exact opposite. The data-set being used is released by Porto Seguro, an auto and homeowner insurance company from Brazil. It consists of information from several hundred thousands of policyholders. The goal is to predict the probability an auto insurance policyholder files a claim the next year using classification algorithms. A good prediction with decent accuracy can correctly adjust prices for policyholders.

KEYWORDS

i523, HID233, HID204, Big data, Classification, Safe Driving, Predictive Analytics, Neural Networks

1 INTRODUCTION

Everyday, people die from car accidents and it should come as no surprise that automobile accidents are one of the most common causes of death in the United States [11]. As reported by the CDC, Centers for Disease Control, approximately over 40,000 people lose their lives to fatal automobile accidents each year. It should be clear that we need to enhance road safety for drivers all over the states [7]. However, as we are currently in the age of big data, these automobile accidents could be prevented by using modern technologies and methods such as artificial intelligence and predictive analytics.

Big data describes large quantities of data that are impossible to analyze using traditional data analysis methods. It includes structured and unstructured data. Structured data can be SQL database stores and unstructured data can be videos, images, social media feeds, etc. In industries, data analytics is often performed on big data in order to find specific patterns or anomalies that could prove useful for business decisions and choices. The amount of data is usually irrelevant in these cases. For example, smart cars utilize big data to improve their safety features and systems. They collect data such as driving patterns and routes as they travel from point A to B. This information is then sent to the computers onboard and gets transferred to the company servers where the data undergo analysis. The result is then collected and stored to enhance smart-car systems [13].

Big data is also helpful in providing insights for product development. It can find the causes for issues and problems in products through data analytics, which can then be used to improve the design. For example, the driver assistance feature on a Mercedes-Benz car not only has safety features but also collects data on driving habits. If large amount of drivers speed through intersections or break hard during traffic hours, the company can obtain these information and use them properly to enhance their systems for better road safety. They could add a detector with GPS data to spot intersections or traffic jams. Furthermore, another data that are useful to collect are driving routes. Google Street View utilizes big data from driving routes to update their maps and display views of different places [13].

Aside from smart cars, we also need to master data collection in order to know how automobile accidents happen. For example, the technology behind the famous black box, which tracks planes and cockpit communications to determine the reason behind crashes, is getting used in cars as well. It is not expensive nor complicated to apply this technology onto the majority of vehicles out there. By recording the precise time, locations, speed and other variables, this technology can definitely help us collect valuable data and information in car accidents. The result from data analysis can also assist us in a deep understanding of causes behind automobile collisions in order to save more lives by preventing future accidents. The first country who thought to implement this technology on cars was South Korea. As a result, in the following year, a 14 percent decrease was saw in the number of car accidents along with a 20 percent decrease in the number of injuries and deaths of fatal automobile accidents [7].

Predictive analytics is a powerful method in big data analytics to help predict future events or outcomes based on current and historical data. It usually utilizes big data techniques such as data mining, predictive modeling and statistics. It uses a wide range of predictive models which depends on the type of event we are predicting. For example, most predictive models produce a number called a score where a higher score indicates a higher chance of that outcome happening in the future. It is a very useful tool for making business decisions and assessing potential risks in many industries such as insurance, retail, etc. Predictive analytics does not inform users about things that has happened before today. It tries to predict for a particular driver the probability that he or she may be involved in an car accident in the near future or any other chose time as accurately as possible [11].

For example, in order to find high-risk drivers, it is not enough to just have driving records, automobile incident reports or traffic tickets. We also need something called telematics. Telematics is defined as the combination of telecommunications and informatics. It collects, stores, sends and receives data and information via transmission-enabled devices. For example, the use of the car black box technology mentioned previously to collect and obtain information on driving behaviors or patterns is called vehicle telematics. With telematics data, companies can determine the possibility of a driver in a future car accident along with the expenses coupled with it. They can also take actions such as putting high-risk drivers

into training schools to correct their bad driving behaviors before an incident happens [11].

By studying the telematics data from a specific driver, we can learn his or her driving behaviors and create a report that details the potential danger this driver may inflict and use these data to correct those bad driving habits. The probability of a driver being involved in a car accident in the future can be used to categorize drivers into different groups. With these probabilities, companies can create a safety score to provide them with suggestions on which drivers to deal with first. The fundamental role for a safety score is to identify drivers with high risks before an accident happen to give the driver a chance to correct those bad driving habits and prevent incidents from happening [11].

However, just like other countless programs on risk evaluation, predictive analytics is not supposed to be flawless. In companies such as UPS, FedEx, USPS or any other services that use a large amount of drivers and vehicles called fleet vehicles, predictive analytics is just the first process and it requires the total cooperation and commitment from the company, drivers and fleet staff to achieve the highest efficiency. Companies that have numerous successful fleet operations are those who plan ahead and bring together all fleet personnel into the action. The most effective way to adjust the accuracy and preciseness of predictive models is to test the models every few days or few weeks or any other time that is suitable for the operations. Due to the fact that most operations have tight schedules, this leeway time will give companies enough room to call in safety personnel to step in to either train drivers to correct their driving habits or repair fleet vehicles beforehand to avoid major system failures [11].

Predictive analytics and telematics data are being used in almost all fleet companies as more of them start to see the value in predictive analytics. With the help of predictive analytics, fleet companies can be actively engaged in making better decisions about their fleets and companies by improving road safety, reducing expenses and risks or decreasing work load time [11].

2 BIG DATA IN THE INSURANCE INDUSTRY

For a long time, auto insurance companies have calculated insurance rates based on personal mileage through out the years [5]. Traditional auto insurance companies categorize users by demographics such as gender and other factors such as education. They then make predictions based on past statistics about their chances of getting involved in future car accidents. This means that the monthly payments insurance companies charge you are only calculated according to the information they have on you and these information has nothing to do with your driving behaviours. As a result, the premiums you pay every month is usually based on past data from people who have identical demographics as you. While a few of the factors are actually helpful in determining your risk score - for example, if you have had multiple accidents in the past, you are likely to be involved in a new one in the future - other factors such as how many cars you previously owned have little to do with your actual risk of being in an accident but yet they still matter when calculating the price. And no one ever use the most important factor in determining monthly premiums which is driving behaviors [9].

Major auto insurance companies have connection to large amount of information as well as data processing power in order to calculate risk scores and monthly premiums. It is no easy task for them to combine several different factors into one single price. Still, big data is not fully utilized. Even though these companies use a variety of different models to calculate monthly rates for drivers, not all of their methods are optimized. In a study conducted earlier this year, it was found, with the help of data mining, that there exists predictive models that have higher accuracy in categorizing drivers into high and low risk groups. Among those models was one that combined 16 factors which produced an extremely high accuracy in risk assessment [9].

Big data can help insurance companies in a big way to make better and smarter business decisions.

- Financial fraud has always been a big problem for both insurance companies and their clients ever since the invention of insurance. The expense that insurance fraud inflicts each year is more than 40 billion dollars as reported by the FBI. However, insurance companies now can put big data into good use as a brand new approach in detecting insurance fraud. Coupled with immense computing power and complex mathematical algorithms, insurance companies are now able to analyze their data to find abnormalities which might indicate possible fraud. For example, a variety of applications and computer software now have the ability to detect outliers automatically in the data. However, the anomalies detected do not always turned out to be fraud situations. There could be other reasons or explanations for them but this new approach certainly makes insurance companies a lot easier to detect potential fraud [3].
- Another benefits for insurance companies to use big data is that big data analytics applications such as Apache Hadoop are engineered to be simple to use with office software such as Excel. As a result, it is much easier to write reports in Hadoop which is intended to work with Excel. Insurance staff can now access huge amounts of data swiftly to obtain the information they need and produce a report in the form they already know how to use [3].
- Two years ago Google released a tool for residents of California to compare rates among different auto insurance companies. Since then, the competition has been increasing in the industry. However, there is no need for this useless competition because big data can help insurance companies find better ways to provide their customers with a good price while still earning profits. By collecting data and customer information from various sources such as social media, insurance companies can utilize these big data to accurately predict which customers are likely to file claims in the future and then try to bring in more of these customers [3].
- Aside from auto insurance, big data also has some interesting applications in the industry of health-care. With the accurate and effective collection of data on medical records, insurance companies can provide better health insurance plans for people so that they can have longer and better lives. As a result, people with better health insurance plans

file less claims which means insurance companies spend less money and earn higher profits. For example, insurance companies can advertise wearable technologies or devices such as apple watch to track customers well-beings in order to provide them with incentives to exercise or obtain better lifestyles [3].

- The insurance industry is continually changing. Insurance companies that can not match the pace will lose profits and resources. However, using big data, insurance companies can study and learn real time data such as social media feeds to obtain more information about customer styles and preferences. This can help insurance companies to design better marketing strategies, adapt faster to customer feedback and construct products that are more attuned with their customers' tastes [3].
- Last but not least, big data can also help insurance companies to provide insurance plans that are tailored to their customers. Every year, millions of money is lost because insurance companies do not have the means to personalize their insurance plans. However, with the help of big data analytics tools, employees in insurance companies now have the ability to gather more precision information on every one of their customers with ease so that they could create insurance plans according to each individual's needs. These tools and software can also give insights and advice based on the collected data to provide better support for employees to make decisions. This in turn will increase their customer satisfaction and lead to more customers in the future [3].

However, powered with advanced technology and big data analytics, insurance companies now-days have access to customers' driving behaviours for more personalized insurance rates. They collect specific data on how often you drive every day, how long you drive each time, how often you speed, how often you break hard and so on to determine the probability of you being in an accident in the near future. With these precise data on each individual customer, insurance companies can assess risk scores for everyone and use that information to calculate your monthly rates [5]. For example, an auto insurance company called Root is one of the first mobile auto insurance companies that are intended to help you on the go. They promise to only insure the good drivers to make sure they get the best rates. Their methods are simple. Download the app, take a test drive with the app on-board for several weeks and Root will send a personalized premium plan based on your driving behavior. Then you can just select the plan you want to purchase and buy via your phone [9].

These new and innovative mobile auto insurance plans are called UBI, short for Usage-Based Insurance and they calculate monthly premiums mostly based on driving habits. Applications on smartphones and on-board diagnostic devices along with in-car tracking technology from manufacturer are used to record mileage and driving behaviours. These mobile insurance programs tend to give discounts for good drivers as a reward for their good driving behaviors. They are even adding new rewards such as roadside assistance on top of discounts for drivers who have maintained good driving behaviors for long periods of time. By collecting big data on driving habits, auto insurance companies can promptly discover mistakes when accidents happen by knowing the exact positions of each car and driving habits data such as speeding or braking as well as environmental data such as weather or road conditions [10].

On the surface, these Usage-Based Insurance plans appear to be plausible and feasible. An application or a sensor is installed on your phone or car to track your driving behavior instead of estimating costs based on factors such as age, gender, education, traffic records, accident reports and so on. Several programs such as *Drivewise* from Allstate insurance and *Snapshot* from Progressive insurance have been released to the public for a couple of years in some states, completely based on customers' choices. You do not have to install them if you do not want to. However, the majority of drivers have been embracing these monitoring devices since it has no apparent downside. As long as your driving behaviors are considered to be safe, such as slow braking and accelerating or no driving around midnight, your should receive discounts like 5 to 10 percent or even up to 20 percent on your monthly premiums. On the other hand, customers are starting to worry about their privacy but we still do not know what the worst thing that might happen if we keep letting insurance companies monitor our driving behavior. According to the Wall Street Journal, these Usage-Based Insurance plans are growing exponentially. The biggest auto insurance company in United States, State Farm, announced their plans to expand their *Drive Safe and Save* program to the entire country soon. Their major advertising strategy is that by enrolling in the program, consumers can get discounts in their insurance premium by proving that they drive safely [12].

For example, one of the earliest Usage-Based Insurance programs available to the public was released about 10 years ago by Progressive and General Motors. This particular program, with the help of GPS, applied discounts based on customer mileage. Many of the Usage-Based Insurance programs these days still implement this strategy but many improvements have been made. Insurance companies nowadays know everything about the way you drive from where you drive to when you drive as well as how you drive. There are also a variety of options to choose from such as *pay as you drive* and *pay how you drive*, thanks to telematics. The advantage of having telematics in these insurance programs is to improve efficiency such as reducing response time for accidents. In addition, Usage-Based Insurance data can be analyzed using on-board diagnostic devices which are often plugged in via the on-board diagnostic 2 port on cars. These diagnostic devices do not have the ability to track car positions but they do generate more precise and meticulous data about car usage. Although telematics is the typical way to record driving habits, new creations in the future will possibly use smart-phone's location services or GPS abilities to track bad driving behaviors such as speeding and hard braking. Liberty Mutual and State Farm both tested their new tracking technology via smart-phones or other smart devices on-board cars in 2015. By 2020, the majority of auto insurance companies will be using Usage-based Insurance programs coupled with telematics data. It is no doubt that Usage-Based Insurance programs will continue to grow and achieve even higher precision and availability [4].

3 CURRENT APPLICATIONS

Predictive analytics are being utilized with telematics data to enhance and improve road safety. Telematics have been used for a long time in insurance industry to monitor driving behaviors such as speeding to identify high-risk drivers. Now coupled with predictive analytics, these data are being analyzed to predict the likelihood of a driver being involved in future accidents. SmartDrive Systems, a transportation safety and intelligence company, employs even more interesting ways to predict accidents. They record and gather video feeds from dashboard cameras in cars which are then integrated with telematics data. This way, they can improve their predictive analytics on driver safety and eventually leads to better predictions [1].

SmartDrive Systems uses a private cloud to provide their clients with predictive analytics solutions. All the data from their clients are collected by the company and stored on their cloud. The data includes telematics and video feeds from millions of clients with more than 4 billion mileage. SmartDrive constantly improves their predictive models because the agreements SmartDrive has with their clients permit them to study all the data they gather. The usage of both telematics data and video feeds is a great idea which enables SmartDrive researchers to better understand the data and interpret the results. By combining what they see through the video feeds and the results from telematics data analysis, the researchers can draw conclusions such as making a U-turn on a narrow road within some fixed radius is dangerous [1].

Indiana State Police came up with a different way to predict incidents and are making their predictive analytics methods open source. In their approach, they produced something called *Daily Crash Prediction Map* which finally completed in November. It contained data such as accident reports from all the police departments in Indiana going back to 2004 as well as data on daily weather, historical traffic amount and so on. This map highlights where potential accidents may happen categorized by their probabilities. It also features information about past accidents such as locations, dates, causes, fatalities and so on [2].

Liberty Mutual is the nation's third largest property and casualty insurance company. Last year, Liberty Mutual partnered up with Subaru. In doing so, customers was granted access to *Starlink*, Subaru's multimedia and navigation system, which can track and notify drivers via an app if they are speeding or braking too hard. By enrolling in the *RightTrack* program provided by Liberty Mutual, drivers can get up to 30 percent discounts for good driving behaviors [5].

4 DATA ANALYSIS

The data we are going to use for our analysis is of Porto Seguro. It is one of Brazil's largest auto and homeowner insurance companies. Inaccuracies in car insurance company's claim predictions raise the cost of insurance for good drivers and reduce the price for bad ones. The task is to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year. An accurate prediction will allow them to further tailor their prices, and hopefully make auto insurance coverage more accessible to more drivers.

4.1 Approach

We will be mainly discussing about the Exploratory data Analysis we have performed on the data. We will be using the help of both R and python environment and supporting packages to perform the necessary statistical analysis. Along with this, we will discuss about the Machine or Deep Learning algorithms or models that we will be using to achieve the near solution for the problem.

4.2 Feature Information

Dimensions of the data [Rows x Features] : [595212, 59]

The data-set constitutes different varieties of features.

Binomial Features [Count : 17] : ps_ind_06_bin, ps_ind_07_bin, ps_ind_08_bin, ps_ind_09_bin, ps_ind_10_bin, ps_ind_11_bin, ps_ind_12_bin, ps_ind_13_bin, ps_ind_16_bin, ps_ind_17_bin, ps_ind_18_bin, ps_calc_15_bin, ps_calc_16_bin, ps_calc_17_bin, ps_calc_18_bin, ps_calc_19_bin, ps_calc_20_bin.

Categorical Features [Count : 14] : ps_ind_02_cat, ps_ind_04_cat, ps_ind_05_cat, ps_ind_01_cat, ps_ind_02_cat, ps_ind_03_cat, ps_ind_04_cat, ps_ind_05_cat, ps_ind_07_cat, ps_ind_06_cat, ps_ind_08_cat, ps_ind_09_cat, ps_ind_10_cat, ps_ind_11_cat.

Integer Features [Count : 16] : ps_ind_01, ps_ind_03, ps_ind_14, ps_ind_15, ps_ind_11, ps_calc_04, ps_calc_05, ps_calc_06, ps_calc_07, ps_calc_08, ps_calc_09, ps_calc_10, ps_calc_11, ps_calc_12, ps_calc_13, ps_calc_14.

Floating Features [Count :10] : ps_reg_01, ps_reg_02, ps_reg_03, ps_calc_01, ps_calc_02, ps_calc_03, ps_car_12, ps_car_13, ps_car_14, ps_car_15.

The remaining two features constitutes **id** and the **output (or) target**. All the features has been clearly represented using post script, **_cat** for categorical data, **_bin** for binomial data.

The **missing values** in the features are represented by -1.

4.3 Data Pre-processing

As shown in 1, missing values are found in 14 of the 58 columns. There are 6 features with more than 5000 missing row values. Owing to the shear size of the unavailable data, we have not performed any missing value treatment and removed these features from consideration. Of the remaining data, across rows, data is unavailable in almost 500 (smaller than 1 percent of the whole data-set) rows and these are promptly removed.

4.4 Distribution of the target variable

As shown in 2, target variable claims is a binary variable with a skewed distribution of classes. 96 percent of the customers did not make any claims. We wish to consider this distribution in measuring classification accuracy. Area under the ROC curve, recall and precision would be relevant metrics in this case.

4.5 Numerical Predictors (vs) Target Variable

Average value of most of the numerical predictors is higher when the claims are filed. This is a unique phenomenon and we intend to use this contrast in predicting the target variable.

4.6 Artificial Neural Networks

Artificial neural networks (ANNs) are computing models which are based on biological neural networks that constitute human brains. The idea of ANNs is based on the belief that working of human

Features	Missing Values Count
ps_ind_02_cat	216
ps_ind_04_cat	83
ps_ind_05_cat	5809
ps_reg_03	107772
ps_car_01_cat	107
ps_car_02_cat	5
ps_car_03_cat	411231
ps_car_05_cat	266551
ps_car_07_cat	11489
ps_car_09_cat	569
ps_car_11	5
ps_car_12	1
ps_car_14	42620

Figure 1: Missing Data

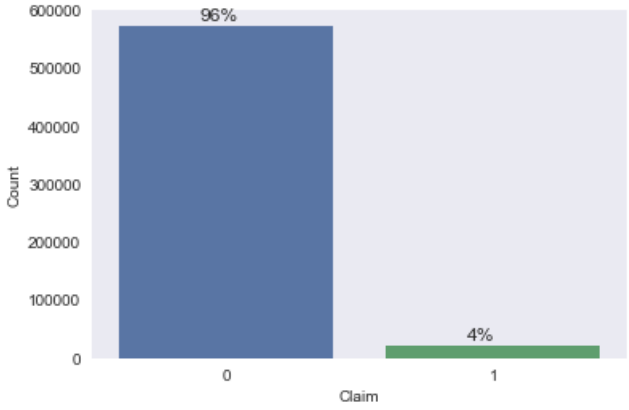


Figure 2: Target Variable Distribution

brain can be imitated for computers by using silicon and wires as living neurons. Such systems learn by progressively improving their performance to do tasks by considering examples, generally without task-specific programming. "The human brain can be considered as a complex network of nerve cells called neurons(about 86 billion)" [8]. They are inter-connected to other millions of cells by Axons. These neurons then react to stimulation from external environment or inputs from other organs. A neuron can then send the message to other neuron to handle the issue or does not send it forward.

ANNs also try to imitate biological neurons of human brain. The neurons are connected by links and they interact with each other. The nodes can take input data and perform simple operations on the data. The result of these operations is passed to other neurons. The output at each node is called its activation. Each node is assigned with weight. ANNs are capable of learning, which takes place by altering weight values. If the network generates the desired output, there is no need to adjust the weights. However, if the network generates an undesired output or an error, then the system alters the weights in order to improve subsequent results.

4.7 Types of ANNs

4.7.1 Feedback ANN. In this type of architecture, the output goes back into the network to achieve the best-evolved results internally. The feedback network feeds information back into itself and is well suited to solve optimization problems. Feedback ANNs are used by the Internal system error corrections [6].

4.7.2 Feed Forward ANN. A feed-forward network is a neural network which consists of an input layer, an output layer and one or more hidden layers of neurons. By evaluating its output by changing its input, the efficiency of the network can be noticed based on group behavior of the connected neurons and the output is decided. The main advantage of this network is that it learns to evaluate and recognize input patterns [6].

5

4.7.3 Radial Basis Function Neural Network. The RBF neural network is the first choice when interpolating in a multidimensional space. The RBF neural network is a highly intuitive neural network. Each neuron in the RBF neural network stores an example from the training set as a "prototype". Linearity involved in the functioning of this neural network offers RBF the advantage of not suffering from local minima [6].

4.7.4 Kohonen Self-Organizing Neural Network. "Invented by Teuvo Kohonen, the self-organizing neural network is ideal for the visualization of low-dimensional views of high-dimensional data. The self-organizing neural network is different from other neural networks and applies competitive learning to a set of input data, as opposed to error-correction learning applied by other neural networks. The Kohonen self-organizing neural network is known for performing functions on unlabeled data to describe hidden structures in it" [6].

4.7.5 Recurrent Neural Network. The recurrent neural network is a neural network that allows a bi-directional flow of data. The network between the connected units forms a directed cycle. Such a network allows for dynamic temporal behavior to be exhibited. The recurrent neural network is capable of using its internal memory to process arbitrary sequence of inputs. This neural network is a popular choice for tasks such as handwriting and speech recognition [6].

4.7.6 Classification-Prediction ANN. It is a subset of feed-forward ANN and the classification-prediction ANN is applied to data-mining scenarios. The network is trained to identify particular patterns and classify them into specific groups and then further classify them into patterns which are unique for that network [6].

4.7.7 Physical Neural Network. This neural network aims to emphasize the reliance on physical hardware as opposed to software alone when simulating a neural network. An electrically adjustable resistance material is used for emulating the function of a neural synapse. While the physical hardware emulates the neurons, the software emulates the neural network [6].

4.8 Data Analysis Using Neural Networks

Rather than beginning our inquiry into the data-set with more traditional methods like regression we straight away tried to learn Artificial Neural Networks. Logistic regression itself, can be thought of as a special case of a neural network with a single neuron (perceptron).

After studying and learning the theory behind ANNs we proceeded to learn how to implement them – by ourselves at first, and later using TensorFlow. So far we have tried quite a bit of different models and learned some lessons about the data.

- **Data Cleaning :** As pointed out by the EDA above, there were a few columns which had a lot of missing data. For columns which had > 1000 values missing (6 columns), we disregarded them altogether. For the remaining data points, we disregarded the rows which had any one particular value missing. We started with a simple data cleaning strategy so as to not complicate it too much at the initial stages, but we will probably want to look at it again as we go along.
- The first thing that we tried is using a simple **perceptron**. The input layer had 51 nodes (after removing the id, target and 6 other columns in data cleaning) and the output layer had a single perceptron with a sigmoid activation function. The best score that we got when we uploaded our code to Kaggle was 0.03 whereas the leader-board is hovering around 0.290 so this is not too impressive.
- However, now we added more hidden layers and nodes to see if we get a better job of fitting the data. To start off, we only consider the continuous variables so that we don't have to worry about handling binary/categorical data. We have 24 nodes in the input layer. The final layer has one node since it is a classification problem. We kept all activation to be logistic and experimented a bit with the number of hidden layers and nodes to get a best score of **0.211** with this simple approach.
- **Issued Encountered** Looking at the results from the neural network there is one major issue. Whenever we add too many hidden layers (> 3) the outputs for the test data are all ≈ 0 and the score drops. After some trial and error, we have diagnosed the issue to be the biased nature of the training data (96% of the training data are $0's$). So the ANN sees too many zeros and consequently predicts mostly zeros. We aim to explore different sampling methods to train the neural network to improve the results further (along with cross validation which we haven't implemented).

5 OTHER TECHNIQUES THAT CAN BE USED

Among the machine Learning algorithms that are used in practice, gradient tree boosting is one technique that shines in many applications.Tree Boosting has been shown to give many state of the art results for many standard classification problem.

The most important factor for the success of XGBoost is its ability to scale in all the scenarios. The XGBoost algorithm run ten times faster than the existing popular solutions on a single solution and scales to billions of example in distributed or memory-limited settings.

The Porto Seguro data-set is clearly an classification problem, the data will be having only two outputs either the car insurance holder is going to claim or not.

While domain dependent analysis and feature engineering plays an important role in defining or modeling the solutions, the fact that XG Boost is the consensus choice for learners shows the impact and importance of our system in tree boosting.One problem with the Porto Seguro data-set we do not have have much information about the Features and all the feature must have to under go through strict statistical treatment to build an optimal solution.

6 CONCLUSION

Reducing insurance rates has always been a difficult task in the past. However, armed with advanced technology, insurance companies now have the ability to track personal driving behaviours to provide better suited personalized insurance plans. We performed Neural Networks algorithm on clients data from Porto Seguro, one of

Brazil's largest auto insurance companies in order to predict the probability of drivers filing claims in the next year. Our analysis proved to be a success and our model yielded a high accuracy.

7 APPENDIX

7.1 Links to iPython notebook:

https://github.com/bigdata-i523/hid233/blob/master/project/Shallow_Neural_Nets.ipynb

7.2 Links to iPython notebook pdf version:

https://github.com/bigdata-i523/hid233/blob/master/project/Shallow_Neural_Nets.pdf

7.3 Work Contribution

Jiaan Wang - Sections: Abstract, Introduction, Big data in the insurance industry, Current application, Data analysis (10 percent) and Conclusion

Dhawal Chaturvedi - Sections: Data analysis (90 percent), Other techniques that can be used and Conclusion as well as Jupyter notebook

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.

REFERENCES

[1] Steve Banker. 2016. Using Big Data And Predictive Analytics To Predict Which Truck Drivers Will Have An Accident. Web Page. (Oct. 2016). https://www.forbes.com/sites/stevebanker/2016/10/18/using-big-data-and-predictive-analytics-to-predict-which-truck-drivers-will-have-an-accident/#7fd6888b1cb0 HID: 233, Accessed: 2017-11-28.

[2] Jenni Bergal. 2017. Troopers Use fiBig Datafi To Predict Crash Sites. Web Page. (Feb. 2017). https://www.huffingtonpost.com/entry/troopers-use-big-data-to-predict-crash-sites_us_589c88ebe4b0985224db5e19 HID: 233, Accessed: 2017-11-28.

[3] Robert Cordray. 2015. 6 Ways Insurance Companies Can Tap The Power Of Big Data. Web Page. (Aug. 2015). https://www.digitalistmag.com/industries/insurance/2015/08/13/insurance-companies-can-use-big-data-advantage-03281426 HID: 233, Accessed: 2017-11-30.

[4] Crosley Law Firm. 2016. Benefits and Concerns About Usage-Based Insurance. Web Page. (Nov. 2016). https://crosleylaw.com/blog/big-data-behind-bad-driving-insurers-use/ HID: 233, Accessed: 2017-11-28.

[5] Brian Fung. 2016. The big data of bad driving, and how insurers plan to track your every turn. Web Page. (Jan. 2016). https://www.washingtonpost.com/news/the-switch/wp/2016/01/04/the-big-data-of-bad-driving-and-how-insurers-plan-to-track-your-every-turn/ HID: 233, Accessed: 2017-11-28.

[6] Naveen Joshi. 2017. Six types of neural networks. Web Page. (April 2017). https://www.allerin.com/blog/six-types-of-neural-networks HID: 204, Accessed: 2017-11-28.

[7] Mikkie Mills. 2017. 4 Ways How Big Data Will Improve Road Safety. Web Page. (May 2017). https://datafloq.com/read/4-ways-big-data-will-improve-road-safety/3127 HID: 233, Accessed: 2017-11-28.

[8] Tutorials Point. 2015. Artificial Intelligence-Neural Networks. Web Page. (April 2015). https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_neural_networks.htm HID: 204, Accessed: 2017-11-28.

[9] Cristol Rippe. 2017. Big Data, Better Rates: Why Current Car Insurance Rate Calculations are Unfair. Web Page. (Jan. 2017). https://blog.joinroot.com/big-data-better-rates-why-current-car-insurance-rate-calculations-are-unfair/ HID: 233, Accessed: 2017-11-28.

[10] Jonathan Shafer. 2016. How Big Data Analytics is Changing the Competitive Auto Industry. Web Page. (Aug. 2016). https://pentaho.com/blog/2016/08/26/how-big-data-analytics-changing-competitive-auto-industry HID: 233, Accessed: 2017-11-28.

[11] Grace Suizo. 2015. Using Predictive Analytics to Improve Fleet Decisions. Web Page. (Sept. 2015). http://www.automotive-fleet.com/channel/gps-telematics/article/story/2015/10/using-predictive-analytics.aspx HID: 233, Accessed: 2017-11-28.

[12] Brad Tuttle. 2013. Big Data Is My Copilot: Auto Insurers Push Devices That Track Driving Habits. Web Page. (Aug. 2013). http://business.time.com/2013/08/06/big-data-is-my-copilot-auto-insurers-push-devices-that-track-driving-habits/ HID: 233, Accessed: 2017-11-28.

[13] David Walker. 2017. How Will Big Data From Self-Driving Cars Influence Road Safety. Web Page. (July 2017). https://technofaq.org/posts/2017/07/how-will-big-data-from-self-driving-cars-influence-road-safety/ HID: 233, Accessed: 2017-11-28.

Big Data Analytics and Applications in the Travel Industry and Its Potential in Improving Travel Accessibility

Weixuan Wang
Indiana University Bloomington
Bloomington, Indiana 47405
wangweix@indiana.edu

ABSTRACT

Big data applications and analytics have been influencing and improving tourists' experience. Travel accessibility refers to provide access for people with disabilities or limited mobility (such as seniors), who represent a growing market in the travel industry by spending billions on leisure and business trips. This report explored the implementation of bid data analytics and applications in tourism, disabilities related studies and assistive technologies for people with disabilities. This report explored the potentials of big data applications and analytics in understanding the needs and travel experience of people with disabilities and improving travel accessibility and quality of life for people with disabilities.

KEYWORDS

i523, HID234, Big Data Analytics, Travel Accessibility, People with Disabilities, Quality of Life

1 INTRODUCTION

People with disabilities represented a large neglected tourism market. According to Amadeus annual report, 15 percent of worldwide population (around 1 billion people) lives with some forms of disability [3]. According to United Nation, people with disabilities are the largest minority group in the world [4, 16, 19]. Notably, the number of people with disabilities is expected to increase as a result of extension of human life-span, decreases in communicable diseases, the improvement of medical technology, and decrease of child mortality [42]. While some forms of disabilities might be genetic, but temporary or permanent disabilities can happen to anyone, such as spinal cord injury after car accident, or limited mobility at later stage of life [19].

Population aging trend also signifies that disability will be a more common and urgent issue in the future [22]. The World Health Organization estimates that by 2050, 21.5 per cent of the global population will be aged over 65 [3]. As a large and fast growing minority group worldwide, people with mobility limits and accessibility issues faces a large ranges of barrier when traveling, and travel and tourism demand of this group is often underestimated or completely ignored [3]. According to the Open Door Organization (ODO) market report in 2015, people with disabilities spend 17.3 billion dollars annually for their own travel [33]. Because people with disabilities usually needs a care giver or family member to accompany them when traveling, the potential economic impact could double [33].

Accessible travel or accessible tourism refers to the inclusive travel activities that enable people with access requirements, including mobility, vision, hearing and cognitive dimensions of access, to function independently, with equity and dignity through the delivery of universally designed tourism products, services and environments [3]. However, the travel experiences for people with disabilities are more than access issues. In order to achieve travel accessibility, which means provide travel activities for people with disabilities, a variety of aspects for travel needs must be taken in consideration. An accessible destination and appropriate accommodation only lay the foundation for a particular travel experience to happen for people with disabilities [33]. More aspects that need to consider for people who are traveling with disabilities, such as accessible transportation, accessible online booking [3].

The ultimate aim for those involved in supporting accessible travel is to empower every individual to plan and travel independently, at their own will [50]. However, the task is not a easy one. Making the whole travel chain accessible, including the information and booking procedures, as well as the infrastructure and processes become a important task for travel accessibility [3].

The development of information communication technologies especially the creation and distribution of user-generated content (UGC) or consumer-generated content (CGC) has successfully changed how people travel and how people gather information for travel [12]. Big data application and analytics has become a trending topic for the tourism industry and tourism studies [12]. The fast development of information and digital technology has changed many people's lives, especially the life of people with disabilities has also been improved by technology [10]. People with poor visions can using cell phones to contact others, access information online with screen readers. People with hearing problems can text other people with their cell phone. The use of big data for disabilities related research, disability informatics and developing assistive technology has been studies to improve the quality of life for people with disabilities [22].

Although big data is becoming an important topic in both tourism studies and disability related studies. There are a gap in the literature about how big data can be used in accessible tourism practice and studies, the potential of big data analytics and applications for improving travel accessibility has not been discussed before. Travel for business and leisure, especially travel independently and with dignity, constitute an essential needs for people with disabilities, and plays a fundamental part in the quality of life for people with disabilities. This study is trying to explore the use of big data applications and big data analytics in tourism and disability related practice and research, illustrating and discovering the potential of using big data applications and analytics for accessible travel and tourism practice and studies.

2 TOURISM AND BIG DATA

Information Communication Technologies (ICTs) have been transforming tourism business globally and revolutionizing the world of Tourism. It transforms tourism from a labor-intensive to an information-intensive industry [45]. Tourists influence by the developments in search engines, network speed and capacity have been using use technologies for better planning and experiencing their trips [47]. In addition, ICTs enable travelers to access reliable and accurate information and make reservations faster, cheaper and more convenient than the traditional way [12]. The development of ICTs also enables Internet users to both create and distribute information (especially multimedia information), which is called user-generated content (UGC) or consumer-generated content (CGC) [12].

Big data is a new and trending topic in the tourism industry and tourism studies, however, it is not unfamiliar to tourist activities. Most activities in the tourism industry had been generating a huge amount of data for several years. Booking flight tickets, reserving a hotel room and renting a car all leaves a data trail [40]. These data could add up to more than hundred of terabytes or petabytes structured data in the conventional databases [2]. Discussions of travel planning on online travel community such as the Lonely Planet Community, status updates and posts on social media like Facebook and Twitter, compliments and compliant on review websites like TripAdvisor and Yelp, recording and sharing travel experience on travel blogs constructs more challenging and live unstructured data that arrives at a much faster pace than a conventional database [2]. Tourism practitioners and tourism scholars are trying to understand tourists' behavior by accepting and analyzing these big data [40].

Tourists in the digital age often use a variety of tools to access information that the tourism industry or other users have provided [46]. A tourist produces a high volume of data when they are searching for travel websites, reporting issues on mobile applications, sharing traffic information in the cities, searching and posting on social media, taking and sharing photos, reporting experience on travel websites and social media, documenting their trips on blogs [2, 40]. All these data that are produced constantly can demonstrate tourists' motivation, interests, and their planning patterns and so on [47].

Previous studies have demonstrated several different usage and formats of big data in the travel and tourism industry [47]. Social media is one of them that has a huge effect on the tourism industry. Social media includes social networks, review sites, blogs, media sharing, and wikis [46]. The exceptional growth of these data sources has inspired companies and institutions to come up with new strategies to understand the socio-economic phenomenon in various fields [40]. Discussions and information sharing on social media are considered as electronic word-of-mouth (eWOM) that has in some degree substituted tradition face-to-face word-of-mouth (WOM) for information exchange of tourist experience [12].

Most tourism research utilizing big data are focusing on CGC or UGC, especially online reviews for a hotel. A recent study conducted by Guo, Barnes and Jia used data mining approach and linguistic analysis to extract meaning from 266,544 online reviews for 25,670 hotels [24]. They mined their customer review data from TripAdvisor using a web crawler [24]. Through their linguistic analysis of their data and cross-comparing with perceptual mapping of the hotels, they found 19 controllable dimensions that are important for hotels to manage their interactions with visitors (such as the price for value, check in and check out) [24].

Photo post on photographic sharing website also can also provide extensive information on the tourists. Previous studies have connected photos posted on Panoramio, Flickr, and Instagram [10, 29]. Because when a tourist post pictures on these websites, their photo is tagged with geographic locations and ordered chronologically. Therefore analyzing photos posted by tourists can provide a photo density map to better understand tourists' behaviors, and potentially provide opportunities to detect atypical tourists behavior and characterize communities behaviors [29]. However, the study also has its own limitation because of the limitation of technology to better exploit the data [10]. Another study focused on the sequence of locations in shared geotagged photos by tourist to identify and recommend travel routes which helped the travel recommender system to generate personalized recommendation according to interests and time available [29].

Overall for tourism industry and tourism research, big data has becoming more and more popular. Both tourism practitioners and tourism researcher has recognized the influence of big data and big data sources for tourism development. Big data in the tourism industry are generated by tourists directly, compared to traditional data sets that are gathered from surveys, they have argued that these direct data from tourists themselves can better represent their true travel experience [24]. Therefore, big data presented us opportunities to better understand tourist behavior, their motivations, and interests.

3 DISABILITY AND BIG DATA

There are many different definition of disabilities from different organizations. The most cited official definition is the 1976 definition of the World Health Organization [4]: "An impairment is any loss or abnormality of psychological, physiological or anatomical structure or function; a disability is any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being; a handicap is a disadvantage for a given individual, resulting from an impairment or a disability, that prevents the fulfillment of a role that is considered normal (depending on age, gender and social and cultural factors) for that individual". While people with disabilities are those people who have limitations in their actions or activities resulting from physical, sensory or cognitive impairments, however, there are many types and levels of disabilities and their actions and activities are affected differently by their disabilities [4]. The complexity of disabilities presents difficulties and challenges to accommodate the different needs of people with disabilities and improve their qualities of life [35].

The number of individuals living with some sensory or cognitive impairment or assisting an affected person is enormous [38]. Researchers has been using disability informatics to better understand people with disabilities. Disability informatics is a sub-specialty of health informatics that is defined as "any application that collects, manages, and distributes information that are related to people

with disabilities, as well as to caregivers (including familiar members and health care providers) and rehabilitation professionals" [4]. Disability informatics is closely related to other health informatics areas such as medical informatics, public health informatics and consumer health informatics, because people with disabilities usually have some secondary medical condition such as poor health status and increased personal health care needs.

Gather medical and health information can help to better understand and accommodate people with disabilities [38]. A study from the early 2000 has identified the potential of public health informatics for prevention at all vulnerable points in the causal chains leading to disability and proposed that applications should not be restricted to particular social, behavioral, or environmental contexts, but in a more global context [48].

Another previous research has designed and deployed an extended version of Artemis system (a cloud system designed to acquire data and store physiological data of clinical information for real-time analytics) in a hospital. They have identified that high speed physiological data produced at intensive care units as big data, and the proper use of such data can promote health, reduce mortality and disability rates of critical condition patients and create new cloud-based health analytics [26]. Research also has shown that many disabilities are genetic, therefore, bioinformatics has implications in the education of genetic screening and gen therapy treatments in the future [4].

People with disabilities usually need some assistive technology in their daily life. These technology that assist them to perform basic physical and social functions. The use information technology and assistive applications in disability informatics are categorized into three areas: virtual, personal, physical.

- Virtual environment refers to use of digital technologies like website and the Internet [4]. The digital revolution had and will continue to have a profound positive impact on the life of people with disability by empowering them with the help of digital technologies [4]. However, there are still access issues in the digital world. One of the barrier is the use of the World Wide Web (WWW or Web). Therefore, virtual environment for people with disabilities is usually discussed regarding to web accessibility.
- Personal Environment refer to having a safe personal environment for people with disabilities,which includes personal management and health monitoring [4]. Safety monitoring and health monitor devices are essential in this personal environment, which enables a safer personal environment and also provide health information for their medical care providers [4]. However the ethic of such health monitor devices are always in debate, some believe it can be an invasion of privacy and a restriction of personal freedom, others hold the ground that its main purpose is to help people with disease or disabilities, since it can alert their caregiver if the individual are exposed to harm (such as a person with mental disability and has a history of self-harming, these device can prevent unwanted behavior) [15].
- Physical environment refers to the actual living space, traveling environment for people with disabilities. People with mobility disabilities, visual impairment or cognitive limitation all need special help in their physical environment. Since the American Disabilities Act passed in the 1990s, the accessibility of physical environment has been improved in a great degree. However, people with disabilities still would meet some barrier and problem, one of them is the lack of curb cuts. Assistive information technologies has been developed in an effort to solve this problem. One of them is MAGUS, which is a project using geographical information system to inform users about wheelchair accessibility in urban areas [4]

The contribution of Big data and cloud computing have been recognized and accepted by researchers in health informatics [44]. The potential of big data and cloud computing for disability informatics and for people with disabilities has been explored by a few researchers and organizations. Data-Pop Alliance is one of the organization has recognized the big data and potential for study and help people with disabilities for disability informatics and people with disabilities [35]. Their research has categorized three type of big data source used across disability research: exhaust data (mobile-based data, financial transaction, transportation and online trace), digital content (social media and crowed-sourced/online content), and sensing data (physical and remote) [35]. They also provided the potential for some of these data sources, for example, researchers can use transaction data to compare cost, availability, and use of services that offer accessible options (such as accessible hotel listings) [35]. They also suggested that researcher can use social media data to represent people with disabilities as a network of interaction and using crow-sourcing to map the locations of accessible businesses and public places [35]. The organization has also identify four functions of big data on disability:descriptive, predictive, diagnostic and engagement. Descriptive function of big data is to describing and presenting the collected information such as using location data to map workplaces that are accessible to people with disabilities [35]. Predictive function is making inferences based on collected information such as discovering trends in the growth of number of accessible businesses in a certain urban area, while the diagnostic function means establishing and making recommendations on the basis of causal relations such as showing what can help increasing accessible business in a certain area [35]. Finally, the engagement function refer to shaping dialogue within and between communities and with key stakeholders through communication of data [35].

Cloud computing in combined with big data can also provide great opportunities for research and improvement of quality of life for people with disabilities [7]. The term cloud "refers to everything a user may reach via the Internet, including services, storage, applications, and people" [25]. Depending on the type of using, the "cloud" can be use for different purpose, such as for companies, the cloud could be used for hosting services so as to avoid the costs and difficulties associated with hosting onefis own servers and software and for individuals, the could is often used as information storage [26]. Regardless of the types of usages for cloud, the end using must still access the information and services residing in the cloud through device like a smart phone or computer [25]. Cloud computing has been used to provide more accessible virtual environment,

especially Web access through project like WebAnywhere, which is a cloud based tool for blind using to access Internet [25].

Cloud computing and big data analytics can also be helpful in health monitoring. The Artemis project mention earlier provide a example of big data analytics and cloud computing usage in health monitoring, by creating new cloud-base health analytics solutions [26]. Previous researchers have developed a mobile app to collect motion data of Parkinson's disease (PD) which is a disease resulting in mobility disorder using the smart phone 3D accelerometer and to send the data to a cloud service for storage, data processing, and PD symptoms severity estimation, which provide an user-friendly and economically affordable system to monitor and assess the condition of PD [34]. Although this system is not for people with disabilities, but it provided potentials for similar systems to be developed for different kind of disabilities.

Another application of cloud computing and big data in assistive technology is the CloudCast platform, which is a cloud-based speech recognition services that can be used for many assistive technology application for people with speech difficulties and hearing impairment, it also facilitate the collection of speech data required for the machine learning techniques [15]. Similar to Alexa Voice Service, it provide reliable speech recognition which can be used with assistive devices for people with hearing impairments, but CloudCast platform also provide customization for assistive technology applications benefiting users with speech impairment [15]. This research provided a great example of using big data and cloud computing in combine to solve a certain problem for people with disabilities (in this case it is barriers for speech impairment).

The development of information technology and assistive technology has improved the life of people with disabilities. The use disability informatics and health informatics can help researchers and service and technology providers to better understand the needs and wants for people with disabilities. Studies has discussed and proposed the great potentials to use big data source to better represent people with disability and identity and study issues and propose actions and solution to the challenges faced by people with disabilities. Using Cloud computing and big data also helps improving assistive and information technology that are now used to help people with disabilities. To improve the quality of life for people with disabilities, travel as a necessary needs and right for human cannot be ignored and the travel demands from the population with disabilities have to be addressed. Therefore, it will be beneficial and necessary to study travel accessibility with the help of big data and digital technologies, in order to improve the quality of life for people with disabilities. By reviewing previous literature, this study is exploring the potential relationship between big data and travel accessibly as shown in Figure 1.

4 TRAVEL ACCESSIBILITY AND BIG DATA

To explore the potential of big data applications and analytics in improving travel accessibility, the complexity of travel accessibility have to be addressed. Accessible travel includes not only the point-to-point transportation (such as air travel, flights), but also the accessibility of destination [3, 16, 30]. For people with disability to actually make the trip, they will also require booking for transportation and hotel reservation to be accessible. This study is going

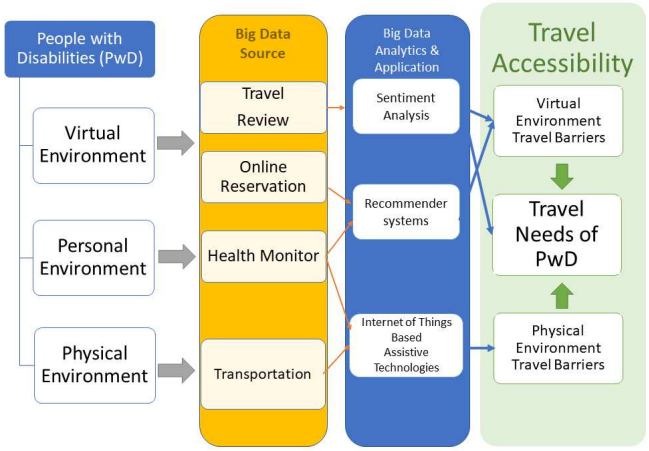


Figure 1: The Relationship Between Travel Accessibility and Big Data.

to explore the use of big data application and analytics in different aspect of travel such as reservation, long distance transportation (in the form of airline travel), and destination transportation, the existing evidence and potentials for using these big data to improve travel accessibility..

4.1 Big Data and Online Reservation

Nowadays, the majority of the travel planning process happens online. Tourists would use a variety of tourism website, search engines, and reservation domains. The Internet also plays an important part at during and post travel stage, as tourists report issues on mobile applications, share traffic information in the cities, search and post on social media, take and share photos, report experience on travel websites and social media [2, 40]. These activities seems mundane and easy to complete for the general public, but for people with disabilities they can huge hassles or barriers.

Previous studies in relation to disability informatics demonstrated the profound positive impact of that the digital revolution on the life of people with disability by empowering them with the help of digital technologies [4]. However, there are still access issues in the digital world, the most urgent one is the use of the World Wide Web (WWW or Web). The Web has always had a strong awareness and been advocacy for accessibility since early on in its evolution [4]. The World Wide Web Consortium (W3C) had passed the Web Accessibility Initiative (WAI) and Web Content Accessibility Guidelines in the late 1990s [4].

A number of assistive technologies were designed to help people with disabilities to use the Web. For example BBC Education Text to Speech Internet Enhancer (BESTIE) is a CGI Perl script that can help people with disabilities who are using text-to-speech systems for Web browsing to modified the web page removing images, Java and Javascript code that may cause difficulties to understand the BBC web page content [18]. However, the limitation of BESTIE is that it is only compatible with BBC website. Other researchers also came up with Personalizable Accessible Navigation (PAN), which is a set of edge services designed to improve Web pages accessibility

which allow personalization and the opportunities to select multiple profiles, making it compatible for web as well as mobile devices [34].

The online sector of the tourism industry has quickly adopted big data applications to better understand the need of customer and to improve online experience for customers [2]. The online sector of the industry include meta-search engines (like Google), online travel agencies (like Expedia) and some information website companies that distribute tourism information (TripAdvisor)[29]. Amadeus, a tourism company known for its global distribution system, has developed a program "Amadeus Airline Cloud Availability" that can generated special result and increase search for its customers [40].

Travel domain companies like Marriott, Southwest airline, and Amtrak also developed assistive devices specially for people with disabilities to use when browsing their websites. These assistive technologies can help people with disabilities to navigate online reservation website, and help them to independently booking their travel reservations for hotels, restaurants, airplane tickets and attraction passes.

These assistive devices was design to help people with disabilities to have access to the Web. However, current web accessibility standards do not respect disability as a complex and culturally contingent interaction. The needs and demand of people with different types of disabilities are certainly not the same and this make it hard to understand and pinpoint the real needs for people with disabilities to access the Web. Researchers have proposed to use big data to better understand of the relation between disability and technology and recognized the difference of disabled people in the "Global South" where different contexts constitute different disabilities and different experiences of web access [4].

4.2 Accessible Transportation and Big Data

An inaccessible transport network prevents many people from going to school or studying, working, going to the doctor, meeting friends, going shopping or to the cinema and other activities that are taken for granted. However, for people with disabilities, an inaccessible transport network would left them dependent and confined in their own home [3]. More importantly an inaccessible transport network would also prevent people with disabilities to travel for business or leisure [30]. Older adults, people with disabilities, individuals in low-income households, especially those living in rural areas can face significant mobility challenges [32].

Concerns about getting into an accident, congestion, price of travel, access to transit, and lack of walkways are important issues for a large percentage of the population, but they tend to be more important for people with disabilities [32]. For today's travel and transportation businesses, it is important to address the issue of inclusion, which is the potential to enable a broader range of people to use transportation infrastructure regardless of their individual abilities or disabilities [30]. Accessibility transportation is essential for travel accessibility, because it represent two aspect of travel accessibility: first is long distance transportation accessibility, and the second is accessible destinations. For a destination to be accessible, it have to have an accessible transport network allow people with disabilities to navigate with the destination.

4.2.1 Airline and Big Data. The airline industry is very familiar with big data use in their daily operations and market research. Airline companies have been using their big data which is the large volume of structured information that has been produced internally [29] to analyze prices of plane ticket. Moreover, airlines have optimized the details of planning for the crew and routing [40].

Previous studies in airline network used Big Data mined from the U.S. air transportation system over the years from 1998fi?!2014 to characterize the network's behavior and determine what internal and/or external drivers result in structural changes to the airline network [14]. Airline delay patents has also been studied with the help of big data by identifying by the number of late arrivals as a percent of total operations [43].

In another previous study, researchers used data from 2006 to 2008 in order to provides the result about the total flight delay for a specific period of time caused due to climate, security, carrier, National Aviation System, Arrival and Departure based on total number of flights getting delayed over in the given period of time [43]. In the study, the authors used time series analysis along with the integration of heterogeneous database to identify and achieve the Airline Seasonal Delay which is implemented and visualized in R, they were able to identify a trend line to provide the insights for the aviation industry to take future measures to avoid delays and manage them [43].

Airline studies have also used big data analytics on passenger reviews data. The advance development of social media and mobile helped the passengers to post reviews in a ubiquitous way, allow them post real time feedback over Facebook, Twitter on airports, airlines, and other travel providers [11]. However, passengers' review can be really complex since travel activities usually involve multiple parties, therefore, the travel domain application systems are also typically managed by different stakeholders like airlines, airports, travel agencies, security and other services providers like cars, bus, trains, hotels, events. In order to provide a holistic approach to manage complex passenger reviews with data gathering, processing and disseminating, a previous study has proposed a reference architecture to manage passenger reviews where multiple stakeholders are involved by using data lakes, which can store, manage and analyze structured and unstructured data with cheaper cost, well-distributed, open sourced and powerful set of tools.

Even though previous studies on airline using big data have not address the issues of accessibility. These previous studies still show the potential for using big data source from the airline industry and passenger reviews to study the interests, motivation, needs and demands from people with disabilities.

4.2.2 Transport Network Accessibility and Big Data. Accessibility in the transport network studies are different defined than travel accessibility, since the urban accessibility is focus on provide access of transportation and transit to general public, not specifically people with disabilities [32]. However, since transport network accessibility usually are connected to urban transport network, which represent an important part of physical environment for people with disabilities it can provide some insight for accessible destinations especially urban destinations. Accessibility has always been a key concept in urban and regional planning for its capacity to link

the activities of people and businesses to the possibilities of reaching them effectively. The accessibility of the transport network is already challenging for general public, for people with disabilities who requires special needs, it become a tremendous and difficult barrier, because they are required to make rapid, real-time decisions that are especially difficult for special needs populations. Therefore, previous studies using big data on urban accessibility can provide some potentials for travel accessibility studies [5].

For Urban an regional planning, accessibility represent traffic capacity to link the activities of people and business to the possibilities of reaching them effectively [32]. Accessibility is defined as "a dynamic attribute of locations that varies over time due to changes in the transport network and in the attractiveness of destinations for certain activities" [32]. Since the emergence of big data generated by social media, smart phones, satellite navigation system and other technologies, the information on transport networks has improved conclusively in recent years [5]. Navigation companies such as TomTom; websites and applications like Bing Maps, Google Map; collaborative projects like Open-Street-Map; the public availability of Transit Feed Specification and data from other transit authorities opens up a rapid growing field of research on real time and time-of-day variations in private and public transit accessibility [32]. These companies and institutions have increasingly detailed systems with information on the features of roads and public transport networks, and their databases include information on speed variations on the roads and the frequencies of passage in public transport networks, all of which contribute a more efficient and dynamic vision to urban accessibility studies [32].

Researchers have just started using these new information sources in studies on urban accessibility. Previous studies utilized data obtained from global positioning system devices to calculate speeds, congestion levels and accessibility conditions at different times of day (morning, midday, evening,), other studies analyzed data from Be-Mobile system (which provided the geo-located positions of 400,000 vehicles equipped with tracking devices) to calculate car travel times [32].

Previous studies also have gathered and analyzed information from web services (such as Google map API) to calculate travel times between origins and destinations. These studies were able to use new information source and big data provided by the development of technologies to retrieve information about local locations and traffic condition for local facilities like groceries, malls, restaurants, banks, recreation centers and others, and estimate accessibility by car, walking, cycling and public transit options [32]. Previous research also used data from social media such as Twitter in combined with data from satellite navigation system (like TomTom) to provide a dynamic approach and obtain profiles that highlight the daily variations in accessibility in urban cities, identify real time influence of congestion and population location changes, by providing different accessibility profiles from different transport zone, the researchers was able to analyze the relationship between the performance of the transport network and the attractiveness of the destination [32]. Although this study is not design to provide information for people with disability, but such dynamic approach using real time big data can also benefit people with disabilities and help them identify places to go in the city and the most accessible route to the attractions.

Another study also proposed and developed a travel assistance device for people with disabilities by using real time data from global positioning system. According to this study, recent advancements in mobile technology enabled smart phones with global positioning system provide real-time location-based services and its related data. The researchers for this study designed, implemented and tested a travel assistance device (TAD) that is designed to help transit riders with special needs using public transportation [5]. This device is a navigation software program designed to prompt individuals via a cell phone to exit the bus at a pre-set location. This travel assistance device provides the people with disabilities customized real-time audio, visual and tactile prompts for exiting the transit vehicle by announcing "Get ready"and "Pull the cord now!", based on its real-time assisted GPS data provided by the embedded GPS chip in the cell phone [5]. Once the software is downloaded and installed to the cell phone, parents, travel trainers, or other authorized individuals can access the web management page to schedule bus routes to be transmitted to the cell phone [5]. The system also provides alerts to riders, their caretakers and travel trainers when the rider with disabilities deviates from the planned route. With a website allowing easy access for the design and planning of new trip itineraries, the device allows authorized personnel (usually caregivers, family members) to monitor the riderfis location in real-time from any computer [5]. The travel assistance device was catered to the needs of people with disabilities, increasing their level of independence and their care-takers security. This travel assistance device represented beneficial practice for people with disabilities [5].

However, there are still many challenge for the development of such device or services. One of the main challenge is that different needs of people with disabilities making it difficult to complete satisfy and assist people with different disabilities. Since the device proposed by this study is still at experimental stage and their test sample are limited, although through the test, the device was proved to be easy to use, it still might pose as a challenge for other people with disabilities, especially people with cognitive limitations [5].

5 PROMISES OF BIG DATA IN TRAVEL ACCESSIBILITY

As mentioned above, big data has some potentials in studies and research that are intended to enable and improve the ability for people with disabilities to travel independently. This section are going to explore different big data related technologies, applications and analytics that can help people with disabilities to travel with ease and researchers to better understand the demand and need for people with disabilities.

5.1 Internet of Things and Travel Assistive Technology

5.1.1 Internet of Things Assistive Technology and Potentials for Travel use. Assistive devices or technology (AT) for people with disabilities are not a new concept. Assistive devices refers to "any item, piece of equipment, or product system, whether acquired commercially off the shelf, modified, or customized, that is used to increase, maintain or improve functional capabilities of individuals with disabilities", for examples, canes, crutches, walkers,

wheelchairs, and shower chairs, hearings aids, visual aids, other hardware, and software that improve ICT access or communication capacities are all assistive devices [5, 41]. People with disabilities are usually seen as depend, and in need of help from caregiver or family members [49]. However, with the technological growth that has been seen in the last 20 years, a wide array of devices have been adapted, created, and utilized with the potential to create independence for individuals with disabilities. Virtual reality, sensor monitoring devices, smart phone have all been used to assist those who require assistance in their day to day lives [32].

Older adults and individuals with physical, sensory, and mental/cognitive disabilities encounter many barriers to inclusion and accessing various opportunities and services that the society has to offer [41]. In order to overcome these barriers, people with disabilities needs some forms of assistive technologies or devices. Traditional technologies has many challenges, however, with the development of technology, the Internet of Things (IoT), smart homes, smart buildings, smart cities, and other smart environments can overcome some of these challenges due to their prevalence and diverse capabilities [9].

One of human's fundamental needs is the mobility and capability of independently traveling around, including for people with physical disabilities and even blind or visually impaired individuals [5]. In order for people with visual impairment to independently travel or moving around requires indoor and outdoor navigation capabilities, the blind and visually impaired people may rely on some type of AT to supplement their navigational abilities [41]. Previous studies has proposed using the advanced sensors of the smart phones in providing meaningful interactions with the environment for individuals with different abilities [41]. A typical modern smart phone has more than twenty sensors, which include GPS-based systems that can be useful for outdoor navigation, however in general, these sensors in smart phone still may lack the required precision and reliability for use by the blind or partially sighted individuals [41].

Previous research has proposed using Bluetooth beacons and audible instructions delivered through an interface device for navigation by the blind and partially sighted people is based on the use of such as bone conduction earphones or smart phones [5, 41]. A research team has designed and tested a system "with 16 Bluetooth beacons providing pin-point accurate indoor location mapping" for unassisted mobility in the London Underground [41]. The system is based on a mobile app, Wayfnder, which is a open platform that has the promise of promoting future development on assistive devices for people with visual limitations [41]. Microsoft, Guide Dogs, UK and several other organizations have also embarked on expanding similar concepts to respond to the challenges that people with sight loss face while navigating the cities [41]. These companies and organizations has developed technology and application that allow users to start a trip and they have a "Look Ahead" and "Find the way" mode that can help people with vision limitation to explore the city and let them stop at any point and check that they are heading the right direction [5].

5.1.2 The Big Data Challenge of IoT Based AT. While the growth and progress the IoT and smart environments, technologies such as sensor technologies for comprehensive monitoring and surveillance progress and advance unbelievable fast, nevertheless, there still existed many challenges for these technologies [41]. One of the most important challenges is data availability. Because these technologies generated an enormous amount of data that surpasses the processing and use capabilities [41]. For instance, real-time localization and navigation systems that are designed to assist people with visual impairment to travel around, face two major related challenges: one is the allocation of computational resources that can process the large amounts of data coming from multiple sensors and cameras, fast enough in a real-time and synchronized manner, so that they can provide real-time guidance for people with special needs [41];the second issue relates to quick and real-time access to dynamic data sets through interfaces that are appropriate for the user [41].IoT devices typically have the issues of energy constrained, with small memory, limited processing power, and restrictive communication capabilities [41]. One positive aspect of this challenge is, these dynamic data obtained from IoT based AT device are extremely valuable and could help researchers and companies to better understand the needs of people with disabilities and analyzing such data can help researchers and companies to design better product for people with disabilities [17].

Another issues that AT adoption faces is its ability of meeting the usersfi needs and desires [41]. AT has been criticized because although AT devices "fimay have technical merit, and may solve obvious problems, but still fail to address the complex interplay of issues at work and to take the most appropriate approach to dealing with these matters. Furthermore, it is important to acknowledge that there may not even be a firightfi problem to tackle. Flexibility cannot be overvalued" [41]. Due to the complexity of the needs and wants for people with different disabilities, it can be challenging to develop an assistive device that can accommodate most people, however, a holistic understanding of the intended users is required [41]. It is important for researchers and AT device engineers to understand the wants and needs of people with disabilities, to be able to design an AT product that actually can fulfill what people with disabilities want, instead of just assuming what people with disabilities needs [5].

Some people may feel intimidated by the newer technologies such as those of the IoT-based AT [15]. First these device required some sort of learning and adapting period, and for people with disabilities, it might too longer time than for "normal" people, which can be taxing for people with disabilities and give them extra pressures [41]. For people who are used to being in control of their devices, some automate processes of IoT-based AT, which was intended to provide support for people with disabilities, ironically, may pose potential stress for operating and adapting to the devices [41]. A similar issue arises from the fast pace of the development of such advanced technologies. With the rapid advancement of technology, products and service become obsolete really quickly as the newer improved version become available. As mentioned above, people with disabilities do have a learning curve and need adapting process for IoT based assistive technologies, the constant and multiple upgrades of new version can make it harder for people with disabilities to adapt. Therefore, the elderly or people with disability or dementia may miss out on obtaining the full benefits of these devices or services [5, 15, 41]. The costs, learning curves,

or simply a lack of awareness can potentially prevent these people to use new technologies at all.

With the the emergence of new technologies such as Internet of Things and large scale wireless sensor system, IoT based AT emerged as potential solution and promise for improving the quality of life for people with disabilities. They provide new opportunities and can aid people with disabilities in their travel, and help them overcome travel barriers. However, there is still manage challenges for people with disabilities, especially when it come to complex situations that they are going to encounter during their travel. There is a distinct gap in IoT based AT research: the lacking of holistic understanding of the needs and wants from people with disabilities. More studies needs to be conducted on the opinions and users experience of IoT based AT.

5.2 Sentiment Analysis on Online Reviews

The popularity of social media, especially review sites like TripAdvisor and blogs and wikis, leads to an enormous amount of personal reviews for travel-related information on the Web [37]. More importantly, the information in these reviews is valuable to both tourists and travel and tourism practitioners for various understanding and planning processes [49]. These UGC cames from all kinds of tourists with different demographic background, within with also has reviews from people with disabilities. Therefore, analyzing hotel reviews on various website and platform that are posted by people with disabilities can help us better understand the needs of this population group. One of the most common analytics method for large amount of review data is sentiment analysis [37].

Sentiment analysis, which is also called opinion mining, is one of the most active research areas in natural language processing [37]. The aim of sentiment analysis is to define automatic tools able to extract subjective information from text in natural language, and to create structured and actionable knowledge to be used by either a decision support system or a decision maker [23, 49] . The sentiments of reviews, online reputation or online documents are usually categorized in positive, negative and (in some studies) neutral sentiments [21]. The main goal of the sentiment classification is to extract "the global sentiment based on the subjectivity and the linguistic characteristics of the words within an unstructured text" [21]. Therefore, sentiment analysis provided a framework to transform unstructured text to structured data, which make it strongly applicable to both the academic field [8]. Because of the importance of sentiment analysis to business and society, it has spread from computer science to management science and the social sciences [36]. As a social science field and business industry, tourism an travel studies have already been using sentiment analysis in the research.

Previous studies have identified two primary approaches for sentiment analysis: methods based on the combination of lexical resources and Natural Language Processing (NLP) techniques; and machine learning approaches [21]. Since 2009, researchers have been using machine learning methods in the natural language processing (support vector machine (SVM), Nave Bayes, and the N-gram model) to do sentiment analysis on TripAdvisor reviews [49]. Their study analyzed online reviews related to travel destinations, using different supervised machine learning algorithms The algorithms to evaluate the reviews about seven popular travel destinations in Europe and North America [49].

Thee etBlogAnalysis project developed a combined crawler /sentiment extraction application for the tourism industry, which used a simple and robust linguistic parsing methodology with information and terminology extraction methods in order to determine relevant utterances on expression level [37]. It will also provide a warning for tourism operator such as a hotel, if too many negative entries have been generated by their reviewers [21].

In tourism studies, sentiment analysis has been compared to traditional qualitative analytic methods. A previous study compared three alternative approaches for mining consumer sentiment (manual content coding, corpus-based semantic analysis, and stance-shift analysis) from large amounts of qualitative data found in online travel reviews [9, 13]. They applied three different approaches to study consumers' reaction to farm stays in order to demonstrate how large volumes of qualitative data can be analyzed quantitatively in a relatively efficient and reliable way [21]. Manual content coding is the same as traditional the content analysis approach involving two researchers collaborated in a manual coding process designed to extract consumer likes and dislikes from the qualitative data [20]. According to the comparison, computer generated sentiment analysis such as stance-shift analysis processing on both syntax and lexicon assures the coding maintains the statement's context identifying what is important to the informants by the way they express their comments. Most importantly, stance-shift analysis does not categorize what the researcher thinks is important in reviewer's words [9]. The study suggested by combining different approaches in sentiment analysis such as using stance-shift analysis first identifies the significant word segments then using corpus-based semantic analysis detects key themes in those segments helps uncover narrative themes of consumer experiences in large qualitative databases [9].

Sentiment analysis will help researchers to better understand people's travel experience, however, there are few studies have been done to identify demographic information of the reviewer and compare the sentiment analysis result across different demographic [23]. A recent invention present the possibility of identifying demographic characteristics while conducting sentiment analysis. The invention consist of a product or service review to determine demographic information of the reviewer [6]. A sentiment text analysis is performed on the product or service review, wherein the sentiment text analysis examines the product or service review to determine a sentiment of the product or service review. The sentiment of the product or service review is categorized based on the demographic information of the reviewer [6]. This invention presents the promise of using sentiment analysis on the travel experience of people with disabilities. However, challenges still remain for research of UGC generated by people with disabilities, such as the challenge presented by privacy concerns of personal data online [27].

5.3 Recommender System

Nowadays tourists faces a very challenging task of trip preparation because of the huge amount of information available on the Web about tourism and leisure activities [1]. Recommender systems

becomes essential for tourists and tourism operators. For tourists, recommender systems can be a useful tools to help them make decision for travel planning, such as the choices of destinations, attractions, accommodations and restaurants. As for tourism operators, it can be a great marketing opportunities for them to reach a variety of targeted potential consumers. Complex problems such as automated planning, semantic knowledge management, group recommendation or context-awareness have by now been heavily studied in this area [31].

There are already several tourism recommender system available for general public. TIP and Heracles systems provide recommendation service through mobile devices for tourism, through implement hybrid algorithms to calculate tourist preferences, using the defined tourist profile and location data [31]. Crumpet system provides new information delivery services for a variety of different tourist population based on location aware services, personalized user interaction, accessible multimedia mobile communication that uses Multi-Agent Technology [39]. CATIS is a Web based tourist information system using context-awareness, which include context elements such as location, time of day, speed, direction of travel and personal preferences. This system provided information to tourists relevant to his or her location and time [39]. TravelWithFriends using group recommendation service, the first step is to build a recommendation list for each user and to merge them to obtain a destinations shortlist. Afterwards, each group member rates all these options and a Borda count is used to determine the best five destinations to be recommended [31].

Classical recommender systems filter the domain items according to a particular user, using his or her demographic data, past ratings or purchasing history [28]. This approach are used to recommend specific items such as books, songs or films [28]. However, it may not be suitable for travel activities, since most of time travel is an activity that involves a group of people (such as family members, friends). Therefore, it is necessary to take into account the different preferences of all members of travel group when providing recommendations [31]. Previous studies and technology reports have identified two primary options for group recommendation: the first one is to merge the lists of items recommended to each group member, or creating a group profile with everyonefis preferences and then compute a single list of group recommendations [20]. The second optionfis first step is the same as the previous option, by constructing of a list of recommendations for each group member. In a second step though, an automatic consensus-reaching process is applied, in which individual preferences are continuously updated until a high degree of agreement between all the group members is reached [20].

The use of semantic domain knowledge in the recommendation process has heavily increased in recent years. Previous studies have defined the semantic similarity between two concepts is as "the ratio between the number of different ancestors and the total number of ancestors of both conceptsfifi [31]. The items to be recommended are clustered according to this semantic similarity and the recommendation procedure selects the best item from random clusters [39]. Previous study has shown that this procedure keeps the accuracy and increases the diversity of the results [31]. Semantic information can also be used to determine the items to be recommended in a personalized visit to a museum or destinations, by using a shortest-path semantic distance to determine which museum objects or attractions should be recommended to the user [31].

Previous study also proposed a hybrid tourism recommendation system for persons suffering from physical or intellectual limitation. This proposed recommendation system is not simply trying to improve experience, but to create and increase the confidence of users that despite of their limitations they can visit and experience certain places without being afraid, and to help them to truly live a touristic experience. As shown in Figure 2, the system models a user stereotype profile, by identifying the userfis functionality and point of interest (POI) accessibility level, , which represent userfis related knowledge which is layered with several knowledge representation structures and models and produce an accurate touristic recommendation plan [39]. The study represent itself as a opportunity to provided needed information to people with disabilities through a hybrid tourism recommendation system.

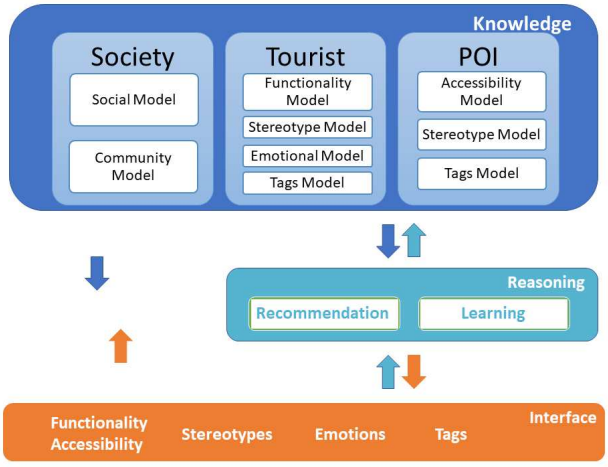


Figure 2: Hybrid Tourism Recommendation System[39].

6 CONCLUSION

This study has explored the big data applications and analytics in tourism industry and research, and disability related research. This study illustrated the importance of improving travel accessibility by recognizing the underestimated market for travel of people with disabilities. The lack of research on big data application and analytics in travel accessibility was identified. By recognizing the complexity of travel accessibility, this study present the potential of using big data analytics and application to better understand the need of people with disability in two travel accessibility aspects: online reservation, and accessible transportation. Although there are few studies on big data and accessible online reservations and accessible transportation directly, this study illustrate big data utilization in web accessibility, airline studies and urban accessibility. These previous studies show promises of using big data analytics and application to address accessibility issues and the needs of people with disabilities in these aspects. This study also explored the promise of big data in travel accessible by exploring:

- Potentials of Internet of Things (IoT) based assistive technology (AT): help people with disabilities overcome travel challenge presented by physical environment.
- Recommender systems: help people with disabilities to get more needed information online, and make it easier for them to navigate the virtual environment.
- Sentiment analysis on online reviews: help researchers and practitioners to better understand the needs and behaviors of people with disabilities.

However, there are still a lot challenge faced by researchers and organizations interested in improving the quality of life for people with disabilities. The most dominated challenge is the different needs for people with different disabilities types and function levels. Future studies could use sentiment analysis of reviews online generated by people with disabilities to better understand their needs and identify the differences between different disabilities groups. Future studies should also analyze dynamic data generated by the sensors on the assistive devices for people with disabilities to better understand their travel patterns and to provide more appropriate products for people with disabilities.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski and TAs for i523 for his support and suggestions to write this paper.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2015. *Context-Aware Recommender Systems.* Springer US, Boston, MA, 191–226. https://doi.org/10.1007/978-1-4899-7637-6-6

[2] Rajendra Akerkar. 2012. *Big Data and Tourism.* Technical Report. Technomathmatics Research Foundation.

[3] Amadeus. 2017. *Voyage of discovery.* techreport. Amadeus, Madrid Spain. http://www.amadeus.com Accessed 2017.

[4] Richard Appleyard. 2005. *Disability Informatics.* Springer New York, New York, NY, Chapter chapter 11, 129–142. https://doi.org/10.1007/0-387-27652-1-11

[5] S. J. Barbeau, P. L. Winters, N. L. Georggi, and M. A. Labrador. 2010. Travel assistance device: utilising global positioning system-enabled mobile phones to aid transit riders with special needs. *IET Intelligent Transport Systems* 4, 1 (March 2010), 12–23. https://doi.org/10.1049/iet-its.2009.0028

[6] D.A. Bhatt. 2014. Sentiment analysis based on demographic analysis. (May 15 2014). https://www.google.com/patents/US20140136185 US Patent App. 13/675,653.

[7] Ann Cameron Caldwell. 2011. *Untapped Markets in Cloud Computing: Perspectives and Profiles of Individuals with Intellectual and Developmental Disabilities and Their Families.* Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter Chapter 30, 281–290. https://doi.org/10.1007/978-3-642-21663-3-30

[8] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. 2013. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems* 28, 2 (March 2013), 15–21. https://doi.org/10.1109/MIS.2013.30

[9] Antonella Capriello, Peyton R. Mason, Boyd Davis, and John C. Crotts. 2013. Farm tourism experiences in travel reviews: A cross-comparison of three alternative methods for data analysis. *Journal of Business Research* 66, 6 (2013), 778 – 785. https://doi.org/10.1016/j.jbusres.2011.09.018 International Tourism Behavior in Turbulent Times.

[10] G. Chareyron, J. Da-Rugna, and T. Raimbault. 2014. Big data: A new challenge for tourism. In *2014 IEEE International Conference on Big Data (Big Data).* IEEE, Washington, DC, USA, 5–7. https://doi.org/10.1109/BigData.2014.7004475

[11] Cynthia Chen, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies* 68, Supplement C (2016), 285 – 299. https://doi.org/10.1016/j.trc.2016.04.005

[12] Jin Chung and Dimitrios Buhalis. 2009. *Virtual travel community: bridging travellers and locals.* IGI Global, USA. 130–144 pages.

[13] W. B. Claster, M. Cooper, and P. Sallis. 2010. Thailand – Tourism and Conflict: Modeling Sentiment from Twitter Tweets Using Naïve Bayes and Unsupervised Artificial Neural Nets. In *2010 Second International Conference on Computational Intelligence, Modelling and Simulation.* 89–94. https://doi.org/10.1109/CIMSiM.2010.98

[14] E. Clemons, R. Jordan, and T. Reynolds. 2016. Airline network and competition characterization using big data approaches. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC).* IEEE, Sacramento, CA, USA, 1–10. https://doi.org/10.1109/DASC.2016.7777957

[15] Stuart Cunningham, Phil Green, Heidi Christensen, JJ Atria, A Coy, M Malavasi, L Desideri, and F Rudzicz. 2017. Cloud-Based Speech Technology for Assistive Technology Applications (CloudCAST). *Harnessing the Power of Technology to Improve Lives* 242 (2017), 322.

[16] Simon Darcy. 2010. Inherent complexity: Disability, accessible tourism and accommodation information preferences. *Tourism Management* 31, 6 (2010), 816 – 826. https://doi.org/10.1016/j.tourman.2009.08.010

[17] G Dewsbury, K Clarke, M Rouncefield, I Sommerville, B Taylor, and M Edge. 2003. Designing acceptable 'smart' home technology to support people in the home. *Technology and Disability* 15, 3 (2003), 191 – 199. http://proxyiub.uits.iu.edu/login?url=https://search-ebscohost-com.proxyiub.uits.iu.edu/login.aspx?direct=true-db=ccm-AN=106746102-site=ehost-live-scope=site

[18] Ugo Erra, Gennaro Iaccarino, Delfina Malandrino, and Vittorio Scarano. 2007. Personalizable edge services for Web accessibility. In *Universal Access in the Information Society (W4A)*, Vol. 6. WWW2006, ACM, Edinburgh, UKfiff, 285–306.

[19] Lex Frieden. 2015. Why Disability Informatics? (02 2015). https://sbmi.uth.edu/blog/feb-15/021115.htm

[20] Inma Garcia, Laura Sebastia, Eva Onaindia, and Cesar Guzman. 2009. *A Group Recommender System for Tourist Activities.* Springer Berlin Heidelberg, Berlin, Heidelberg, 26–37. https://doi.org/10.1007/978-3-642-03964-5-4

[21] Aitor Garca, Sean Gaines, and Maria Teresa Linaza. 2012. A Lexicon Based Sentiment Analysis Retrieval System for Tourism Domain. *E-review of Tourism Research* 10, 2 (2012), 35 – 38. http://proxyiub.uits.iu.edu/login?url=https://search-ebscohost-com.proxyiub.uits.iu.edu/login.aspx?direct=true-db=hjh-AN=84339713-site=ehost-live-scope=site

[22] Jan Grue. 2016. The social meaning of disability: a reflection on categorisation, stigma and identity. *Sociology of Health and Illness* 38, 6 (2016), 957–964. https://doi.org/10.1111/1467-9566.12417

[23] Dietmar Grbner, Markus Zanker, Gnther Fliedl, and Matthias Fuchs. 2012. Classification of Customer Reviews based on Sentiment Analysis. In *19th Conference on Information and Communication Technologies in Tourism (ENTER).* Springer, Helsingborg, Sweden, 460–470.

[24] Yue Guo, Stuart J. Barnes, and Qiong Jia. 2017. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management* 59, Supplement C (2017), 467 – 483. https://doi.org/10.1016/j.tourman.2016.09.009

[25] Jeffery Hoehl and Kaleb August Sieh. 2010. *Cloud Computing and Disability Communities: How Can Cloud Computing Support a More Accessible Information Age and Society?* Technical Report. Silicon Flatirons Center, Colorado, US. https://doi.org/10.2139/ssrn.2285526

[26] H. Khazaei, C. McGregor, M. Eklund, K. El-Khatib, and A. Thommandram. 2014. Toward a Big Data Healthcare Analytics System: A Mathematical Modeling Perspective. In *2014 IEEE World Congress on Services.* IEEE, Anchorage, AK, USA, 208–215. https://doi.org/10.1109/SERVICES.2014.45

[27] Jonathan Lazar, Michael Ashley Stein, and Judy Brewer. 2017. *Disability, human rights, and information technology.* Philadelphia : University of Pennsylvania Press, [2017], Pennsylvania, USA. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true-db=cat00016a-AN=inun.16424800-site=eds-live-scope=site

[28] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. 2015. Recommender system application developments: A survey. *Decision Support Systems* 74, Supplement C (2015), 12 – 32. https://doi.org/10.1016/j.dss.2015.03.008

[29] Shah Jahan Miah, Huy Quan Vu, John Gammack, and Michael McGrath. 2017. A Big Data Analytics Method for Tourist Behaviour Analysis. *Information and Management* 54, 6 (2017), 771 – 785. https://doi.org/10.1016/j.im.2016.11.011 Smart Tourism: Traveler, Business, and Organizational Perspectives.

[30] Milo N. Mladenovifj. 2017. Transport justice: designing fair transportation systems. *Transport Reviews* 37, 2 (2017), 245–246. https://doi.org/10.1080/01441647.2016.1258599 arXiv:https://doi.org/10.1080/01441647.2016.1258599

[31] A Moreno, L Sebastiá, and P Vansteenwegen. 2015. Recommender Systems in Tourism. *IEEE Intelligent Informatics Bulletin* 16, 1 (Dec. 2015), 1–2. http://www.comp.hkbu.edu.hk/~iib/

[32] Borja Moya-Gómez, María Henar Salas-Olmedo, Juan Carlos García-Palomares, and Javier Gutiérrez. 2016. Dynamic accessibility using Big Data: The role of the changing conditions of network congestion and destination attractiveness. *Networks and Spatial Economics* 1, 7 (2016), 1–18.

[33] Open Door Organization. 2015. Open Doors Organization Market Study Press Report. (2015). http://opendoorsnfp.org/market-studies/2015-market-study/ accessed 2017.

[34] Di Pan, Rohit Dhall, Abraham Lieberman, and B. Diana Petitti. 2015. A Mobile Cloud-Based Parkinson's Disease Assessment System for Home-Based Monitoring. *JMIR mHealth uHealth* 3, 1 (26 Mar 2015), e29. https://doi.org/10.2196/mhealth.3956

[35] Gabriel Pestre. 2016. Big Data and Disability, Part 1. Data Pop Alliance. (March 2016). http://datapopalliance.org/big-data-and-disability-part-1/ Accessed 2017.

[36] Federico Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu. 2016. *Sentiment Analysis in Social Networks*. Elsevier LTD, Oxford, Cambridge, MA.

[37] V. B. Raut and D. D. Londhe. 2014. Opinion Mining and Summarization of Hotel Reviews. In *2014 International Conference on Computational Intelligence and Communication Networks*. IEEE, Bhopal, India, 556–559. https://doi.org/10.1109/CICN.2014.126

[38] Paraskevi Riga and Georgios Kouroupetroglou. 2013. Indoor Navigation and Location-Based Services for Persons with Motor Limitations. In *Disability Informatics and Web Accessibility for Motor Limitations*. IGI Global, Greece, 202–233. https://doi.org/10.4018/978-1-4666-4442-7.ch006

[39] Filipe Santos, Ana Almeida, Constantino Martins, Paulo Moura de Oliveira, and Ramiro Gonçalves. 2018. Hybrid Tourism Recommendation System Based on Functionality/Accessibility Levels. In *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference, PAAMS 2017*, Fernando De la Prieta, Zita Vale, Luis Antunes, Tiago Pinto, Andrew T. Campbell, Vicente Julián, Antonio J.R. Neves, and María N. Moreno (Eds.). Springer International Publishing, Cham, 221–228. https://doi.org/10.1007/978-3-319-61578-3-23

[40] S. Shafiee and A. R. Ghatari. 2016. Big data in tourism industry. In *2016 10th International Conference on e-Commerce in Developing Countries: with focus on e-Tourism (ECDC)*. IEEE, Isfahan, Iran, 1–7. https://doi.org/10.1109/ECDC.2016.7492979

[41] Seyed Shahrestani. 2017. *Internet of Things and Smart Environments*. Springer-Verlag GmbH, Cham, Switzerland.

[42] Ralph W. Smith. 1987. Leisure of disable tourists: Barriers to participation. *Annals of Tourism Research* 14, 3 (1987), 376 – 389. https://doi.org/10.1016/0160-7383(87)90109-5

[43] M. Sornam, M. Meharunnisa, and Parthiban Nagendren. 2017. Big Data Analytics on Aviation data for the prediction of Airline Trends in Seasonal Delay. *International Journal of Advanced Research in Computer Science* 8, 5 (2017), 2248. http://proxyiub.uits.iu.edu/login?url=https://search-ebscohost-com.proxyiub.uits.iu.edu/login.aspx?direct=true-db=edb-AN=124636583-site=eds-live-scope=site

[44] M. Viceconti, P. Hunter, and R. Hose. 2015. Big Data, Big Knowledge: Big Data for Personalized Healthcare. *IEEE Journal of Biomedical and Health Informatics* 19, 4 (July 2015), 1209–1215. https://doi.org/10.1109/JBHI.2015.2406883

[45] N.L. Williams, A. Inversini, N. Ferdinand, and D. Buhalis. 2017. Destination eWOM: A macro and meso network approach? *Annals of Tourism Research* 64 (2017), 87–101. https://doi.org/10.1016/j.annals.2017.02.007 cited By 0.

[46] Zheng Xiang, Zvi Schwartz, John H. Gerdes, and Muzaffer Uysal. 2015. What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management* 44, Supplement C (2015), 120 – 130. https://doi.org/10.1016/j.ijhm.2014.10.013

[47] Karen L. Xie, Kevin Kam Fung So, and Wei Wang. 2017. Joint effects of management responses and online reviews on hotel financial performance: A data-analytics approach. *International Journal of Hospitality Management* 62, Supplement C (2017), 101 – 110. https://doi.org/10.1016/j.ijhm.2016.12.004

[48] WA Yasnoff. 2000. Public health informatics: improving and transforming public health in the information age. *Journal of public health management and practice* 6, 6 (11 2000), 67–75.

[49] Qiang Ye, Ziqiong Zhang, and Rob Law. 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications* 36, 3, Part 2 (2009), 6527 – 6535. https://doi.org/10.1016/j.eswa.2008.07.035

[50] Ye Zhang and Shu Tian Cole. 2016. Dimensions of lodging guest satisfaction among guests with mobility challenges: A mixed-method analysis of web-based texts. *Tourism Management* 53 (2016), 13–27.

Importance of Big data in predicting stock returns and price

Gagan Arora
Indiana University
2709 E 10th St
Bloomington, Indiana 47401
gkarora@iu.edu

ABSTRACT

In this project, we will discuss the importance of big data in finance industry in predicting financial stock values. We will be using python libraries to fetch financial data from yahoo finance and will further predict the stock price returns of few selected technology companies such as Amazon, Yahoo depending on the historical data of x[16] years. Similarly, we will predict the returns based on y[10] years of data. The prediction will be based on SP 500 market return and market risk volatility. Here y is greater than x and then we will compare the predicted returns with the current returns. For the comparison we will be using the testing time frame as mentioned in the project later. This project will help us understand if more historic data helps in predicting the stock price returns or it adds noise. We will be using statistical approach and CAPM [capital asset pricing model] to predict stock price. Analysis will be done on the jupyter notebook

KEYWORDS

HID-301, Stock Price prediction, stock returns, SP500, risk free market, CAPM model, root mean square analysis, stock beta, Finance, Statistics,mean, variance, market premium, python,yahoo finance, i523

1 INTRODUCTION

By its nature of the business, the finance industry is always driven and dominated by data. The existence of Big data in the finance industry has exposed the big opportunity of growth and value extraction but at the same time imposed the various new challenges, which demand new skill set. [5] suggests that finance experts believe there is a huge potential in terms of value extraction from the financial big data. They also believe that finance industry can benefit more than any other industry. Historically, data was always there in some format either non-digital or digital. However, with digitization, this data has fallen into the prevalence of high volume of information, which we call as Big Data. Dominant drivers for the actuality of big data in the finance industry are mainly customer call logs, social media, news feed, regulatory data etc. Call logs, news feed and etc. fall into the category of unstructured data which is identified as an area where we can extract vast amount of business value.

[4] talks about the three V of big data in finance industry: volume, velocity and variety. [6] clearly depicts the amount of financial data pouring in the daily basis. TechNaviofis forecast (Technavio 2016) predicts data will grow at a CAGR [compound annual growth rate] of 61 percent over the period of 2017-2021. According to the IDC financial insight 2016, every second there is around 10,000-payment card transaction and this number is expected to double by the end of this decade. The Capgemini/RBS Global payments study for 2012 suggests there was about 260 billion transactions in 2012 and is expected to grow between 15 and 22 percent for developing countries. Main drivers contributing to the big data in the finance industry are Data growth, increasing scrutiny from regulators, digitization of financial products, changing the business model and increased customer insight platforms such as customer service. [4] shows 76 percent of banks say the business driver for embracing big data is to enhance customer engagement, retention, and loyalty and seventy one percent of banks say that to increase their revenue, they need to better understand customers and big data will help them to do so.

Thinking about the data strategy, the financial industry has taken the business-driven approach to a big data. According to the IBM report, all financial organizations are not keeping the same pace as peer industry is keeping. Today because of increased competition, customers always expect more personalized banking service and at the same time, there is increased regulatory surveillance which in result creates big pressure on finance industry to better utilize the value of Big data. To achieve better-personalized experience, many banks have started the initiative to utilize the information gained from the vast ocean of data to offer better-personalized products and gain competitive advantage. Despite the fact that financial industry is data-driven, there is a gap in the amount of initiative financial industry has taken to extract the value out of big financial data. Technavio 2016 report has shown only 26 percent of financial organizations has focused on understanding the principal notation of Big data and most of those 26 percent are still struggling to define the clear roadmap. This clearly concludes that finance industry lag behind their cross-industry peers in using more varied data types. A good example to support this fact is that there are very less research and domain knowledge in extracting value out of retail bank call logs.

Big data technologies not only help in extracting the effective business value but analysis of unstructured data in conjunction with a wide variety of data set also helps in extracting commercial value. Big data in finance industry does not necessarily decode to valuable or actionable information. The real benefit lies in developing the technologies, which can be used to extract business and commercial value. [15] talks about what all advantage we can extract from the big data in the finance industry. Few examples are: Detection of false rumors that try to manipulate the finance market, Assessment of exposure to a reputational risk connected to consulting service offered by banks to their customer and Discover

topic trends, detect events, or support the portfolio optimization or asset allocation. Big data based pattern recognition can also help in enhanced fraud detection systems and prevention capability systems. Other benefits of utilizing big data include building a machine learning based algorithm to achieve higher performance and accuracy in the trading algorithm and Enhanced market trading analysis. There has been proven research [12] which states more data increases accuracy and precision of simulations which is the backbone of financial modeling based analytic. This research [12] states modern modeling techniques are data hungry.In this project we will extract inference if more financial data can be used to have better prediction.

2 USE OF STRUCTURED FINANCIAL DATA

This reflects the data which has a higher degree of an organization such as a relational database where information/data is easily searchable and we can easily apply standard algorithm to extract patterns out of it. In this project we will be using Yahoo finance structured data. Examples of such data set include yahoo financial data, trading applications, enterprise finance resource planner, Retail banking systems, Credit history database systems and other financial applications that use legacy application systems. Structured data always has a big advantage of being easily entered, stored, queried and analyzed. Most of the personal banking financial statements are stored in a structured way. Structured dataset combined with the distributed systems can be leveraged to achieve structured big data set on which we can run optimized SQL queries to retrieve patterns. [9] discusses various SQL based ways to specify information quality in data which can be used to filter out the noise. In this project we will be using structured data.

3 VARIOUS CHALLENGES UTILIZING BIG DATA VALUE IN FINANCE INDUSTRY

There are multiple challenges and constraints in extracting value out of big financial data. The biggest challenge is old IT culture and infrastructure. The much financial organization still uses old IT infrastructure which is not compatible with the big data application thus fail to take advantage of big data. Other challenges include lack of skill set and data privacy and security. With the emergence of digitalization, customer data is saved persistently because of which there has been continued concern regarding the customer privacy. Regulatory bodies guidelines on customer data are always ill-defined because of which is there is always a concern regarding the use of customer data. In this project we will use standard python libraries to fetch financial data from yahoo finance. Analysis will be done on the jupyter notebook.

4 STOCK RETURNS PREDICTION - LITERATURE REVIEW

Authors of [1] discuss the importance of stock price and returns prediction based on the data extraction of historic data. This research [1] also shows historic financial data has definitive predictive relationship to the future value of stocks. Stock prediction always help investors to decide perfect timing of buying or selling stocks. There are various data mining, artificial neural networks and machine learning techniques available for the stock price prediction based on the value extraction from the historic financial data. Based on the complexity of stock price matrix, pricing mechanism is essentially a non linear complex system. Authors of [14] and [13] state many predictive algorithm is based on the fundamental analysis of macroeconomics and company fundamentals. [11] states problem with the fundamental analysis is that it is too much focused on the intrinsic and lacks the quantitative aspect of the historic financial data. On the broad category we can define stock prediction analysis is based on two types of analysis: qualitative and quantitative. Choice of analysis is mainly based on the fact if we want to have short term analysis or long term analysis. In this project we have have ten and sixteen years of training data and used close to one year of testing data. Since our analysis is based on the historic data we have chosen to do quantitative analysis. Quantitative analysis is based on the pattern extraction, fact that history repeats and future financial drivers can be extracted based on the historic data. Advantage of using quantitative analysis is that we can use statistical confidence interval to validate the analysis.

There is a huge benefit of using machine learning algorithms in predicting stock prices. These algorithms made easy to cope up with the various financial events such as mergers acquisitions, bankruptcy, fraud, political changes, market crashes, housing bubble, dot net bubble and etc. In this project we have used hybrid approach of combined CAPM [Capital asset pricing] model and machine learning algorithm to mine data of sixteen and ten years of data and used close to 1 year of testing data. These machine learning algorithms can further be used to predict various financial events. Other approaches such as neural networks algorithms, SVM, logistic regression and multiple discriminant analysis can also be used to predict financial events. Example, [2] in their research they proved neural networks algorithms performs better in predicting financial events as compared to multiple discriminant analysis. There are other applications which use these algorithm to find predicted credit rating of a company. Credit rating plays a very important role doing qualitative analysis of the financial health of a company. On the other hand, accuracy of these algorithm is a big challenge because of the amount of huge data which it uses as input. Typically, accuracy of these algorithms is validated based on square root method.

In this project we have used several years of data for analysis which involves more than hundred thousands of rows with multiple columns. Then this data is analyzed two dimensionally with the same set of market return rows. Since this analysis is calculation intense, In the end we also have performed root mean square analysis.

Over the past few years there has been drastic changes in the way stock market operates. With the emergence of advance web services, there has been powerful enhancement in the data communication between various financial application. Because of which there is ocean of real time data is available, thus machine learning algorithm, neural networks algorithms, SVM, logistic regression and multiple discriminant analysis needs to be smarter. Forecasting stocks and financial parameter is of great interest to the investors. As discussed earlier these algorithms needs to modified depending on the fact if we want to have short term profit or long term profit.

[8] has shown the very interesting analysis of comparing the prediction of stock market with the random walk hypothesis. Author

of [8] ran an experiment in which he tossed a coin and recordes the results and mapped head with the company profit and tail with the company loss. Then result of this experiment was shown to the investors pretending these are the actual market profit and loss. Looking at the result graphs, investors believed it as a actual prediction. This research has shown the altogether different outlook which states stock price prediction and forecast can be fooled and stock prices are perfectly random in nature. On this theory many researchers have classified profit based on three hypothesis:

- Weak form Efficient Market Hypothesis: The weak form of the hypothesis states one can not generate profit by just looking at patterns and trends of stock market.
- Semi Strong Efficient Market Hypothesis: The semi strong form of the hypothesis states only possible way of generating profit is via inside trading.
- Strong form Efficient Market Hypothesis: The strong strong form of the hypothesis states its not possible to generate profit since stock market behaves in perfect random way.

However, if we are running root mean square analysis we can surely compare the accuracy of various algorithm and arrive at conclusion which algorithm is viable for prediction.

5 FINANCIAL DATA EXTRACTION

In this section, we will discuss various technical requirements needed to achieve value extraction from the big data in the finance industry. There are various technical requirements such as data Acquisition, data quality, data extraction, data integration, decision support. In order to fulfill requirements, a hybrid approach combining computer science, algorithms, statistics, data mining, machine learning and pattern recognition study needs to be adopted. To explore the advantage of big data there have been initiatives like data virtualization, multi-document summarization, pattern recognition from LOGS and many start-ups have been emerged. All big companies such as Microsoft, Google, IBM and Amazon are investing heavily in this field to leverage business and commercial value out of it. There has been changed in the industry pattern where financial industry is resorting big data to strategize their business. According to [6] with a very rapid pace, the financial industry is utilizing big data advantage in investment analysis, econometrics, risk assessment, fraud detection, trading, customer interaction analysis and behavior modeling. If we look at the Big promise the Big data holds in the finance industry, progress in this field is still in nascent stage and we expect more growth in upcoming years. In this project we will discuss jupyter notebook based solution for Data extraction.

In this project we have used jupyter notebook and rich python libraries to fetch financial stock data. Later in this paper we will discuss the stock data extraction in detail. Later we will also discuss what are different ways to fetch stock data and will discuss few important functions which python libraries

6 FETCHING FINANCIAL STOCK DATA

Fetching structured precise data is always a challenge. There are different ways to fetch the stock market data. In this project we will be fetching data from yahoo finance via python libraries which internally makes remote web service call to the yahoo web server. There are also other ways to fetch data such as:

- Direct download of csv files from yahoo finance or google websites.
- Make web api call to download the data in the json/XML format
- Use python libraries to download data, which internally makes remote web service call to the yahoo web server. This is preferred way of doing since it allows you to save data to system variables directly.
- Call yahoo or finance web service from the application.
- Calling VBA function in excel to fetch yahoo stock data
- Quandl best for using core financial data and this website also includes access to rich python libraries.
- Google sheet has feature to fetch real time stock prices
- Install stocks macros in excel

In this project we have exhaustively used python for data manipulation. Reasons for using python are:

- Syntax is super easy which comes with very level of readability as compared to other programming languages.
- It is free and supports cross platform as python code can be called from any version of machine.
- Python has strong community support so if any problem is encountered, support is available online.
- Python has powerful tools available such as statsmodels, matplotlib, Pandas,Numpy and SciPyfor calculation intense projects

Since we have exhaustively used the `get_data_yahoo` function from the `pandas_datareader` python library we will briefly discuss the parameters it takes. Please note we utilized only those arguments which are relevant to the project requirements. From [10] parameter list as listed below:

- symbols : string, array-like object (list, tuple,Series), or DataFrame Single stock symbol (ticker), array-like object of symbols or DataFrame with index containing stock symbols.
- start : string, (defaults to '1/1/2010' Starting date, timestamp. Parses many different kind of date representations (e.g., 'JAN-01-2010', '1/1/10', 'Jan, 1, 1980')
- end : string, (defaults to today) Ending date, timestamp. Same format as starting date.
- `retry_count` : int, default 3 Number of times to retry query request.
- pause : int, default 0 Time, in seconds, to pause between consecutive queries of chunks. If single value given for symbol, represents the pause between retries.
- session : Session, default None requests.sessions.Session instance to be used
- `adjust_price`: bool, default False If True, adjusts all prices in hist data ('Open','High', 'Low','Close') based on 'Adj Close' price. Adds 'Adj Ratio' column and drops 'Adj Close'.
- `ret_index` : bool, default False If True, includes a simple return index 'Ret Index' in hist data.
- chunksize : int, default 25 Number of symbols to download consecutively before intiating pause.

[3]

- interval : string, default 'd' Time interval code, valid values are 'd' for daily, 'w' for weekly, 'm' for monthly and 'v' for dividend.

In our analysis, for the symbol parameter we are passing ticker symbol one at a time. Though, we have an option to pass multiple tickers as an array argument. We are using `get_data_yahoo` function and utilizing only first three paramters: symbols, start and end. This function returns YahooDailyReader object which can further be manipulated to get Open, High, Low, Close, Adj Close and Volume stock values. Since default number of retry count is three we will be using this default value. Default value of pause which is zero is also good with respect to our requirement so we will not pass this as argument. Session argument should be used when we are handling multiple request in parallel in the code since our project we just need one session so we will not use this argument. `adjust_price` is not required in our analysis since we are interested only in returns which can be fetched using `pct_change()` function. Since return index is of no use in calculating the returns, we will not use this argument. Argument chunk size is used to modify number of consecutive downloads of stocks since we are just using single ticker so this argument is of no use. This function uses interval also as a parameter since we are only interested in daily values and daily value is the default interval so we didn't pass this argument in the function call.We could also use the contemporary google function which is `get_data_google`. Arguments which goes to the `get_data_google` are symbols, start, end, `retry_count`, pause, chunksize and session since we are not using `get_data_google` function in our project we will not discuss these in detail.

7 INTRODUCTION TO CAPM MODEL

CAPM [Capital asset pricing model] model was developed by William Sharpe and John Lintner in 1964. This model is considered so powerful that it is being used in current prediction models. There are few advantages of using CAPM model as compared to other pricing models:

- This model is a single dimensional model and easy to use, still powerful to model capital asset pricing.
- Since this model is based on the market portfolio and risk free rate, this model removes unsystematic risk.
- We can run root mean square algorithm to validate the algorithm.
- This model provides a flexibility to utilize various risk free rates and run model for various time range.
- This model can be applied to various financial objects such as stocks, put option, call option, bonds, and etc

This model can be used to evaluate the theoretical expected return on a security, security can be any financial object such as stocks, put option, call option, bonds, and etc. In CAPM model we evaluate how much financial object is sensitive to the market using statistical analysis. Then this sensitivity which is also known as beta is used to find the expected return on security. This expected return can be on daily basis, weekly basis, monthly basis or yearly. Here is the formula to evaluate expected return:

$$E(R_i) = r_f + \beta_i(E(r_m) - r_f)$$

Where

- $E(R_i)$ is expected return
- r_f is risk free interest rate example: Government bond
- $E(r_m)$ is return on market example SP 500
- β_i is sensitivity of stock with respect to market

β_i can further be defined how much stock is sensitive to the stock market. Example if β_i for a particular stock is two it means if market goes up by five percent then stock will go up by ten percent and if market goes down by two percent then stock will go down by ten percent. In terms of statistics β_i is defined as:

$$\beta_i = \frac{Cov(R_i, r_m)}{Var(r_m)}$$

Where covariance and variance are defined as

- $cov_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$
- $var^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$

β_i matrix can be used to illustrate β_i in a following way:

$$\begin{array}{c} \\ Row1 \\ Row2 \\ Row3 \\ Row4 \end{array} \begin{array}{c} \begin{array}{ccc} \beta_i & MarketReturn & ExpectedReturn \end{array} \\ \begin{pmatrix} +2 & +5\% & +10\% \\ -2 & +5\% & -10\% \\ +0.5 & +4\% & +2\% \\ +0.5 & -4\% & -2\% \end{pmatrix} \end{array}$$

Above matrix suggests how expected returns can be correlated with the the β_i. Example if for certain company has β_i of +2 and market returns is +5 % then company's expected returns can be predicted as +10 %. Please note β_i can be positive as well as negative.

8 PROPOSED ANALYSIS

In this project we will utilize structured data and use CAPM [Capital asset pricing model] to statistically find the expected daily return of selected technological stocks: Amazon and Yahoo. This daily expected return can be used to predict next day stock value given the condition we have current stock price. Following formula can be used to predict next day stock price:

Next Day stock price =: Today stock price * (1+Daily expected return)

Daily expected return will be calculated using CAPM model. Daily expected return sensitivity in CAPM terminology is also known as beta. In this project beta will be calculated based on two time frames:

Time frame 1: [01/01/2000 to 12/31/2016] 16 years of data
Time frame 2: [01/01/2006 to 12/31/2016] 10 years of data

Thus we we will have 2 β_i:

$$\beta_1 = \frac{Cov(R_1, r_{m1})}{Var(r_{m1})}$$

$$\beta_2 = \frac{Cov(R_2, r_{m2})}{Var(r_{m2})}$$

Where

- β_1 is β based on time frame 1
- β_2 is β based on time frame 2
- R_1 is actual return based on time frame 1
- r_{m1} is a mean market return based on time frame 1
- R_2 is actual return based on time frame 2
- r_{m2} is a mean market return based on time frame 2

Above two time frames will be our training data set.We will run two analysis: one on training time frame 1 and other on training time frame 2 to arrive at predicted CAPM variables. Then we will use this training data set to predict stock returns for test data set which will comprise of time frame:

Test data time frame: 01/01/2017 to 11/16/2017

Then we will run the statistically analysis on the test data to evaluate if 16 years of training data produced more accurate result or else it added noise compared to 10 years of training data. Please note this is purely a quantitative analysis not qualitative. Actual returns can also be impacted by a qualitative factors such as mergers acquisitions, bankruptcy, fraud, political changes, market crashes, housing bubble, dot net bubble and etc.

9 PROPOSED ALGORITHM

Code is written purely in python language and used the powerful rich python libraries such as statsmodels, matplotlib, Pandas,Numpy and SciPyfor. We have used jupyter notebook as interpreter tool to python. Code is started by importing above mentioned rich python libraries. Since we are interested only in technological stocks: Amazon and yahoo we need to initialize they stock ticker with the python variable. In CAPM model we need to know the market return in order to know the stock sensitivity we will also initialize market ticker with SP 500 index. As discussed above we will be using `get_data_yahoo` function from the `pandas_datareader` and in this project we will be only utilizing only first three parameters which is stock ticker, start date and end date. For first iteration we will be using `get_data_yahoo` to fetch stocks and market returns for time frame 1. For having better understanding of how the data looks when fetched using `get_data_yahoo` function, we will have amazon financial data matrix calculated like:

$amazonData = dr.$`get_data_yahoo`$(amazon,$ `start_date`$,$ `end_date`$)$

Where

- dr is `pandas_datareader`.data class
- amazon is stock ticker for amazon which is 'AMZN '
- `start_date` is start date of time frame 1: 01/01/2000
- `end_date` is end date of time frame 1: 12/31/2016

and synopsis of above amazon data looks like:

	Open	High	Low	Close	Adj	Volume
23 − Dec − 16	764.54	766.50	757.98	760.59	760.59	1976900
27 − Dec − 16	763.40	774.65	761.20	771.40	771.40	2638700
28 − Dec − 16	776.25	780.00	770.50	772.13	772.13	3301000
29 − Dec − 16	772.40	773.40	760.84	765.15	765.15	3153500
30 − Dec − 16	766.46	767.40	748.28	749.86	749.86	4139400

Similarly using `get_data_yahoo` we will fetch Yahoo and market returns. Since we are interested in daily return, we fetched the daily data from yahoo finance which is evident from the above result data set. Now lets find the percentage change on the daily Close value to get the percentage change array which in finance terminology will be daily return on stock. For finding the percentage change we are using `pct_change()` function on the close column of result set. This function can be elaborated as follows:

`return_amazon` = $amazonData.Close.pct_change()[1:]$

`return_yahoo` = $yahooData.Close.pct_change()[1:]$

`return_market` = $marketData.Close.pct_change()[1:]$

`return_amazon`, `return_yahoo` and `return_market` are two dimensional arrays and we need to convert them to single dimensional array in order to run statistical analysis. We can use dot values method to extract single dimensional array out of 2 dimensional array. This operation can be elaborated as follows:

`X_amazon_actualReturns` = `return_amazon_testing`$.values$

`X2_yahoo_actualReturns` = `return_yahoo_testing`$.values$

`Y_market_actualReturns` = `return_market_testing`$.values$

Please note these are actual returns - fetched from yahoo finance. Now in order to evaluate expected return for the testing period based on the calculated beta we need to calculate the risk free rate r_f as mentioned above in the CAPM formula. Please note `get_data_yahoo` formula will fetch the annualized rate but here we are dealing with the daily returns so this needs to be normalized to daily rate. Here we are using Treas Yld Index-10 Yr Nts bond. Ticker symbol for Treas Yld Index-10 Yr Nts bond is `^TNX`. Please note `get_data_yahoo` will return columns: Open, High, Low, Close, Adj Close and Volume. Dot values will convert to 2 dimensional array and then used index [0][4] to fetch annual rate. Detailed code with comments is mentioned on jupyter notebook.

Conversion of annualized return to daily return can be done using following formula:

$$riskFreeDailyRate = (1 + riskFreeAnnualRate)^{(1/365)} - 1$$

Now we need to copy the content of `X_amazon_actualReturns` to new array `X_amazon_predictedReturns` and initialized each

5

element in X_amazon_predictedReturns using CAPM model as discussed above in Introduction:

$$\texttt{X\_amazon\_predictedReturns} = list(\texttt{X\_amazon\_actualReturns})$$

We will do the same for Yahoo stocks:

$$\texttt{X2\_yahoo\_preddictedReturns} = list(\texttt{X2\_yahoo\_actualReturns})$$

In the code we have run the while loop and each element of X_amazon_predictedReturns and X2_yahoo_preddictedReturns is assigned the value based on CAPM model. Now we have two returns arrays for amazon stocks based on sixteen years of data:

- X_amazon_actualReturns are the actual returns
- X_amazon_predictedReturns are returns based on the CAPM model.

Similarly we have two returns arrays for yahoo stocks based on sixteen years of data:

- X2_yahoo_actualReturns are the actual returns
- X2_yahoo_preddictedReturns are the returns based on the CAPM model.

Now we can utilize mean_squared_error function from the sklearn.metrics python library to find how predicted returns are deviated from the actual returns. We will run this function on both stocks, amazon and yahoo:

$$a1 = \texttt{Y\_market\_actualReturns}$$
$$a2 = \texttt{X\_amazon\_predictedReturns}$$
$$y1 = \texttt{Y\_market\_actualReturns}$$
$$y2 = \texttt{X2\_yahoo\_preddictedReturns}$$

$$\texttt{rms\_amazon} = sqrt(\texttt{mean\_squared\_error}(a1, a2))$$
$$\texttt{rms\_yahoo} = sqrt(\texttt{mean\_squared\_error}(y1, y2))$$

Here is the root mean square values for both the stocks under sixteen years of data case:

- Root mean square error for Amazon stocks analysis based on 16 years of data 0.0013770 or 0.137 percent
- Root mean square error for Yahoo stocks analysis based on 16 years of data 0.0014313 or 0.143 percent

Now we run the same analysis as discussed above for the ten years of data and will validate how much predicted stocks returns based on the ten years of data are deviated from the actual returns using root mean square method. Please note testing data set remains the same we are just using different training data set. This will let us compare if sixteen years of data is of more worth in predicting stock returns or it added noise to the analysis:

- Root mean square error for Amazon stocks analysis based on 10 years of data 0.0005310 or 0.053 percent
- Root mean square error for Yahoo stocks analysis based on 10 years of data 0.0014910 or 0.149 percent

Above analysis is purely quantitative and does not include any elements of qualitative analysis. It shows predicting yahoo stock price or its returns based on the sixteen years of data or ten years of data - both resulted in almost same results. However things are totally different for the amazon stocks, recent ten years of amazon stocks data produced more accurate results as compared to using recent sixteen years of data. Author of [7] agrees with the fact that most recent financial data are the better predictors of the future price returns. Though, in the [7] author has used the neural networks and support vector machine for prediction. Author also stressed that neural networks algorithm produced better accuracy than other machine learning algorithms.

10 THREE PARADIGMS OF PREDICTION

Data prediction and analysis done in this project is purely quantitative. However there are other paradigms of predictions also which we will discuss here. Example in above analysis we totally missed the qualitative aspect of the data. This is why it explains recent data on amazon stocks produced better results. Here are the other prediction paradigms explained by the author of [3] :

- Quantitative research based prediction. This is a method where we utilize statistical tools to arrive at predictive value based on data
- Quantitative research based prediction. This is a method where we utilize conceptual knowledge to arrive at predicted value. Example of such study would be prediction of stocks based on the events such as mergers acquisitions, bankruptcy, Fraud, political changes, market crashes, housing bubble, dot net bubble and etc.
- Mixed research based prediction. This is a hybrid method where we utilize both qualitative and quantitative results to predict result.

In this project we used quantitative based approach to validate the fact if more data is good for prediction or it adds noise. Result of this project also showed there is a importance of recent data in predicting results. This is also validated by the research done under [7]

11 LIMITATION

In this project, analysis is based on two technological stocks: yahoo and amazon. We can extend our study to more diverse portfolio by including more stocks from various industries. Technological stocks tend to be more volatile than other stocks. Since this project is purely quantitative based prediction we deliberately chosen the technological stocks to leverage their volatility. More accurate prediction could also be made by encapsulating qualitative based prediction in the analysis which is more like a hybrid approach. Such hybrid approaches includes assigning weight to each predictions and taking cumulative result. As the part of future work we can also compare results across industry and arrive at conclusion which industry is more stable in prediction. Comparison can based on the root mean square analysis which is discussed in this project report.

12 CONCLUSION

Main objective of doing this project is to know the importance of big data in predicting financial variables. Analysis of this project is based on two stocks: amazon and yahoo. We started this project report with the discussion of importance of big data in financial industry. In the introduction we discussed how various industries are investing in the Big Data to attain higher standards in terms of quality and customer satisfaction. Then we discussed what are

the various types of data available: structured and unstructured. Since in this project we utilized only structured data so it was discussed deeply. This project report also touch base with various challenges financial industry takes in utilizing the value of big data. As the part of literature review we reviewed various researches done in the field of stock returns prediction. In the financial data extraction section we reviewed various technical requirements need for financial data extraction. As the part of data analysis for this project we discussed what are the various ways to fetch live stock data from yahoo or google server. In this project we used the rich financial python libraries for the analysis so we discussed them in details in this report. Financial model which we chose for the prediction is the CAPM model which is explained theoretically in this report. There are two different section where we discussed the proposed analysis and proposed algorithm. We finally concluded the report by discussing three paradigms of prediction. In the end we also mentioned what further can be done under future work section.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and TAs for their support and suggestions to write this paper. TAs and professor are very good in terms of providing valuable guidance and suggestion in a very prompt fashion.

REFERENCES

[1] Qasem Al-Radaideh, Adel Abu Assaf, and Eman Alnagi. 2013. Predicting Stock Prices Using Data Mining Techniques. (12 2013). https://www.researchgate.net/publication/281865047_Predicting_Stock_Prices_Using_Data_Mining_Techniques

[2] A. F. Atiya. 2001. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks* 12, 4 (Jul 2001), 929–935. https://doi.org/10.1109/72.935101

[3] Adam Chu. 2017. Quantitative, Qualitative, and Mixed Research. (2017). https://www.bcps.org/offices/lis/researchcourse/images/lec2.pdf

[4] Daniel D. Gutierrez. 2014. *Big Data for Finance.* Technical Report. Dell & Intel. https://whitepapers.em360tech.com/wp-content/files_mf/1427803213insideBIGDATAGuidetoBigDataforFinance.pdf

[5] Kazim Hussain and Elsa Prieto. 2015. *Big Data in Finance.* chapman and chapman and hall/crc, https://www.cs.helsinki.fi/u/jilu/paper/bigdataapplication04.pdf, Chapter 17, 329–356.

[6] Kazim Hussain and Elsa Prieto. 2016. *Big Data in the Finance and Insurance Sectors.* Springer, Cham, "https://link.springer.com/content/pdf/10.1007/", Chapter 12, 209–223.

[7] Hui Lin. 2014. Stanford. (2014). https://pdfs.semanticscholar.org/56f0/59ea400f31b60bfde4d59aea71bd7b411553.pdf

[8] Burton G. Malkiel. 2015. *A Random Walk Down Wall Street: The Time-Tested Strategy for Succesful Investing.* Recorded Books on Brilliance Audio. https://www.amazon.com/Random-Walk-Down-Wall-Street/dp/1501260375?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=1501260375

[9] A. Parssian, W. Yeoh, and M. S. Ee. 2015. Quality-Based SQL: Specifying Information Quality in Relational Database Queries. *Computer* 48, 9 (Sept 2015), 69–74. https://doi.org/10.1109/MC.2015.264

[10] Kevin Sheppard. 2017. daily.py daily.py. (2017). https://github.com/pydata/pandas-datareader/blob/master/pandas_datareader/yahoo/daily.py

[11] Philip M. Tsang, Paul Kwok, S.O. Choy, Reggie Kwan, S.C. Ng, Jacky Mak, Jonathan Tsang, Kai Koong, and Tak-Lam Wong. 2007. Design and implementation of NN5 for Hong Kong stock price forecasting. *Engineering Applications of Artificial Intelligence* 20, 4 (2007), 453 – 461. https://doi.org/10.1016/j.engappai.2006.10.002

[12] Tjeerd van der Ploeg, Peter C. Austin, and Ewout W. Steyerberg. 2014. Modern modelling techniques are data hungry a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology* 14, 1 (22 Dec 2014), 137. https://doi.org/10.1186/1471-2288-14-137

[13] Martin Walker and Mamoun Al-Debi'e. 2000. Fundamental Information Analysis: An Extension and UK Evidence. *ACS Biomaterials Science & Engineering* 31 (02 2000).

[14] Muh-Cherng Wu, Sheng-Yu Lin, and Chia-Hsin Lin. 2006. An effective application of decision tree to stock trading. *Science Direct* 31 (08 2006), 270–274.

[15] Sonja Zillner, Tilman Becker, and Munn. 2016. *Big Data-Driven Innovation in Industrial Sectors.* Springer International Publishing, Cham, Chapter 4, 169–178. https://doi.org/10.1007/978-3-319-21569-3_9

A HID 301:GAGAN ARORA

- Identified Project topic.
- Collected the python financial libraries.
- fetched data from yahoo finance
- Studied, designed and reviewed CAPM model
- Implemented CAPM model using python libraries
- Created project report

B CODE REFERENCE

All code, notebooks and files for this project can be found in the github repository: https://github.com/bigdata-i523/hid301/blob/master/project/finalProject.ipynb

Predicting Housing Prices

Murali Cheruvu, Anand Sriramulu
Indiana University
3209 E 10th St
Bloomington, Indiana 47408
mcheruvu@iu.edu,asriram@iu.edu

ABSTRACT

In United States, more than 6 million residential homes sold in 2017. With ever-increasing demands, real estate is challenged with complex analysis of homes to provide accurate appraisals and predicting market fluctuations to react accordingly. Big data analytics helps mining the real estate data to provide valuable business insights. In this project, we have planned to analyze housing data to predict sale prices. Using well established datasets, with lots of exploratory variables, we could apply thorough exploration of the data, feature engineering and implement various advanced supervised learning algorithms, such as XGoost, Ridge, Lasso, Random Forest and Neural Network to predict accurate sale prices.

KEYWORDS

i523, hid306, Exploratory Data Analysis, Supervised Learning Algorithms, Ensemble Modeling

1 INTRODUCTION

Real estate, with $235 million dollar yearly revenue, is a continued growing industry in United States. With more than 200,000 residential and commercial brokerage firms, there are millions houses getting sold every year [9]. In recent times, Big Data has changed the way real estate is getting operated and bringing the importance of data analysis to become major factor in the decision making process. The goal of our project is to predict the sale prices of residential homes listed in the test dataset as accurately as possible. Training dataset contains sale price of the homes, and using this training data set, how accurately we can predict sale prices of the homes in the test dataset by applying data preprocessing and thorough data analysis. In this project, we have applied various exploratory analysis techniques and engineer the features before applying a few advanced supervised learning algorithms such as SVM, XGoost, Ridge, Lasso, Random Forest and Neural Network, to create more accurately predicted models.

2 HOUSING DATA ANALYTICS

Traditional real estate forced buyers to have physical presence to see the homes and meet the realtors. Analyzing the sale price was challenging and require extensive understanding of the neighborhood; highly depend on knowledge of the realtor about recent homes being sold in the surroundings. Assessing the sale price is a daunting task even with a good understanding of the features of any specific home. The true value of mining the real estate data and analyzing it lies in making context-aware relevant data and converting the result to enterprise-grade, tangible and *actionable* business insights. In this project, we would like to predict *sale prices* of housing prices using two datasets - training and testing, each with 79 exploratory variables describing almost every aspect of residential homes in city of Ames, Iowa state. However, the datasets, we have got, are snapshots taken in 2010. As a result, these datasets may not reflect the latest trends in the housing sale prices but the analytical approaches taken in this project are generic and can easily be applied to newer datasets. The key to achieve this lies in getting better handle on the housing data and the trends in sale patterns. With proper housing analytics, not only the realtors get benefit in getting predicted appraisals but also help buyers analyze houses with accurate sale prices within their budget. Machine Learning is empowered with all the capabilities to analyze and provide in depth business insights. Interconnectivity between the economy and housing prices is vital motivating factor in doing this project.

Big Data is defined by *four Vs*: volume, variety, velocity and veracity [5]. (a) Volume: Millions of houses that are in the market for sales will generate high volumes of data. (b) Variety: Housing data comes in various formats: structured, semi-structured and unstructured. Structured data usually come from standard datasets collected at various sources. Video and housing pictures are examples of unstructured data. Traditional relational databases (RDBMSs) will not be suitable for scale out distributed processing to handle such volume and variety. Alternatives like *Hadoop ecosystem*, with Distributed File System, Map, Reduce, etc. aspects, allows complex data processing. (c) Velocity: Data can come in batches, near-real time and real-time. During the housing sale seasons, there will be very high velocity in getting the housing details and the sale transactional data. (d) Veracity: Housing datasets are going to have lots of noise and outlier data. Data mining will address these concerns using *data cleansing* and *normalization* techniques. Various types of analytics can be done using machine learning algorithms and data visualizations to see the classification and predicting model patterns.

Big Data Analytics, in the context of real estate, mainly refers to various analytical activities including population growth, buyer and seller profile matching and neighborhood public schools, using statistical tools and techniques with business acumen to explore hidden information from the available public data across United States. It applies data mining and machine learning algorithms to volume of data coming from multiple sources with various types of data formats. Typical data analytics work-flow include: gathering structured and unstructured data, cleaning the data before modeling, evaluating and visualizing to make them usable for business decisions. Data modeling is, the heart of analytics, to better understand, quantify using statistical algorithms and then visualize the model to comply to the business context. Exploratory data analysis and predictive analytics are two major groups of tasks in the data modeling. Exploratory data analysis uses various techniques to provide useful textual and visual summaries of the characteristics

of the data. Predictive analytics focuses on classification and numerical regression tasks. We will apply predictive analytics in this project, to model the predictions of the *sale prices*.

The real estate industry is tied with Big Data in many ways. Various real estate servicing companies providing advanced insights to buyers and realtors using big data analytics. These companies collect various types of high volume data, such as geographic, census and housing data for rent and sale. Just by using zip code or neighborhood information, one can easily analyze and get the information around potential value of neighborhood properties and trends in the sale. Real estate analytics can tap into *smart cities* data to provide in depth analysis of neighborhood health conditions and energy efficiencies. Banks are using big data sources to analyze and set the prices of foreclosure or short sales in the given neighborhood than offering some lower price which may not correlate with the surrounding similar homes [6]. Big data analytics is going to drive various housing aspects including: buyer identification, accurate pricing and geographic targets [10] along with connecting national and local real estate agents. Social networking datasets can help linking the buyers and sellers [12].

3 DOMAIN KNOWLEDGE

To predict accurate *sale price*, we will need to understand the domain well. We need to build the intuition around all the exploratory variables in the dataset and focus on which factors could influence the target variable: *sale price*. If we do not find all these factors, perhaps, we need to add new features to address the gaps in dataset describing the domain. Some of the factors which, we think, can directly influence house prices are:

- What is the overall Size or area of the house?
- How good is the location of the house - closer to highways?
- How good is the neighborhood?
- How old is the house?
- What is the quality of the construction?
- How many garages are there in the house?
- What are the floor plans?
- How many number of bedrooms are there in the house?
- How many number of bathrooms are there in the house?
- What is the size of living area?

4 EXPLORATORY DATA ANALYSIS

We can start the process with exploratory data analysis. There are 1460 rows in the training data set and 1459 rows in the test dataset. Out of the 80 variables, 23 are nominal, 23 are ordinal, 14 are discrete, and 20 are continuous. The nominal variables are related to material, garage, dwelling, and environmental conditions. All the 20 continuous variables are related to the area dimensions. The ordinal variables rate various items within the property. The home listing includes only few quantified variables like typical lot size and total dwelling square footage, but this data set has more specific variables. There are individual category variables derived from basement, main living area and porch based on quality and type. We have combined training and testing datasets for easier analysis. We excluded Id attribute as it does not add value in the modeling. We also removed Sale Price, the target variable, from the training dataset. All the variables are listed in the appendix section as a reference. We applied univariate, bivariate and multivariate analytical techniques to analyze numerical and categorical variables. Various statistical and data visualizations were applied on each type of variable. The primary goal of exploratory data analysis is to amplify the insights of analysts onto given input dataset to analyze the aspects, such as:

- Good fitting of the model
- Analyzing impact of the outliers
- Missing value analysis and imputation
- Feature engineering and ranking
- Algorithm selection and tuning for optimal predictions

4.1 Analyze Missing Values

First part of the analysis was to check for any missing values in the training and testing datasets as shown in Figure (1). Using the bar plot shown in Figure (2), we have identified that there are 5 variables: *pool quality*, *miscellaneous features*, *alley*, *fence* and *fire place quality*, having the most missing data.

```
# python code - check for null values
train = pd.read_csv('../data/train.csv')
test = pd.read_csv('../data/test.csv')

#combine the data sets
alldata = train.append(test)
na = alldata.isnull().sum()
.sort_values(ascending=False)
```

Figure 1: Code - Null Checks

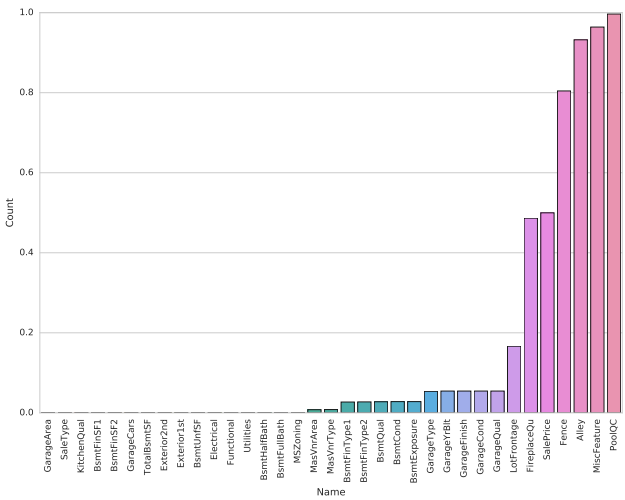


Figure 2: Graph - Missing Values

All the missed values of numeric variables are analyzed further to decide whether we need to delete the instances of all the data with missing values or impute them with something meaningful. There are various ways to find the estimate to replace the missing value including:

- Mean: Replace missed value with the mean value of the corresponding variable
- Regression: Some predicted value by regressing missing variable on all the other variables
- Interpolation and extrapolation: An estimated value from other observations of the same variable

4.2 Analyze Numerical Variables

There are 37 numerical variables after excluding the *Id* variable. List of numerical variables are: MS-Sub-Class, Lot-Frontage, Lot-Area, Overall-Qual, Overall-Cond, Year-Built, Year-Remod-Add, Mas-Vnr-Area, Bsmt-Fin-SF1, Bsmt-Fin-SF2, Bsmt-Unf-SF, Total-Bsmt-SF, 1st-Flr-SF, 2nd-Flr-SF, Low-Qual-Fin-SF, Gr-Liv-Area, Bsmt-Full-Bath, Bsmt-Half-Bath, Full-Bath, Half-Bath, Bedroom-Abv-Gr, Kitchen-Abv-Gr, Tot-Rms-Abv-Grd, Fireplaces, Garage-Yr-Blt, Garage-Cars, Garage-Area, WoodDeck-SF, Open-Porch-SF, Enclosed-Porch, 3Ssn-Porch, Screen-Porch, Pool-Area, Misc-Val, Mo-Sold, Yr-Sold and Sale-Price. *Interval* and *ratio* are the two types of numerical variables we encounter in most of the data analytical applications. Statistical aspects of the numerical univariate analysis include: count, minimum, maximum, mean, median, mode, quantile, range, variance, standard deviation and skewness. Data visualization techniques, such as histogram, box plot and scatter plot are used to analyze the numerical variables. We have shown *sale price*, *overall quality*, *garage live area* and *year built*, in the Figures (4) and (5) as a few sample plots from the numerical analysis. Corresponding code snippet is shown in Figure (3).

```
# python code - analyze numeric variables
numerical_features = [f for f in train.columns
if train.dtypes[f] != object]

nd = pd.melt(train, value_vars = numerical_features)
plt.figure(figsize = (5,3))
plot = sns.FacetGrid (nd, col=variable, col_wrap=4,
sharex=False, sharey = False)
plot = plot.map(sns.distplot, value)
```

Figure 3: Code - Numerical Analysis

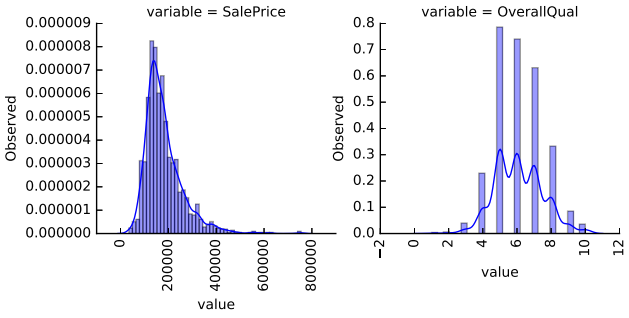


Figure 4: Graph - Sale Price and Overall Quality

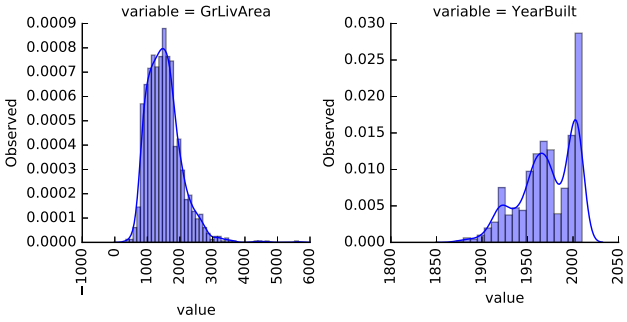


Figure 5: Graph - Ground Live Area and Year Built

4.3 Analyze Categorical Variables

There are 43 categorical variables in the combined dataset. List of categorical variables are: MS-Zoning, Street, Alley, Lot-Shape, Land-Contour, Utilities, Lot-Config, Land-Slope, Neighborhood, Condition1, Condition2, Bldg-Type, House-Style, Roof-Style, Roof-Matl, Exterior-1st, Exterior-2nd, Mas-Vnr-Type, Exter-Qual, Exter-Cond, Foundation, Bsmt-Qual, Bsmt-Cond, Bsmt-Exposure, Bsmt-Fin-Type1, Bsmt-Fin-Type2, Heating, Heating-QC, Central-Air, Electrical, Kitchen-Qual, Functional, Fireplace-Qu, Garage-Type, Garage-Finish, Garage-Qual, Garage-Cond, Paved-Drive, Pool-QC, Fence, Misc-Feature, Sale-Type and Sale-Condition. We have analyzed all categorical variables and found the ways to fill the missing values. We has also evaluated proper approaches to convert them into numerical factors. Bar and pie charts are used to visualize categorical variables. Later on in the feature engineering section, we will go through more details on numerical factors. Categorical variable factors and the corresponding code snippet for *neighborhood* and *sale type* are shown in Figure (6) and Figures (7).

```
# python code - analyze numeric variables
cat_features = [f for f in train.columns
if train.dtypes[f] == object]
print(cat_features)

plt.figure(figsize = (5,3))

p = pd.melt(train, id_vars=SalePrice,
value_vars=cat_features)

g = sns.FacetGrid (p, col=variable, col_wrap=4,
sharex=False, sharey=False, size=5)

g = g.map(barplot, value,SalePrice)
```

Figure 6: Code - Categorical Analysis

4.4 Analyze Correlations

Numpy package offers correlations functionality to analyze the variables that are positively or negatively correlated with the *sale price* and also analyze any interdependencies among the variables.

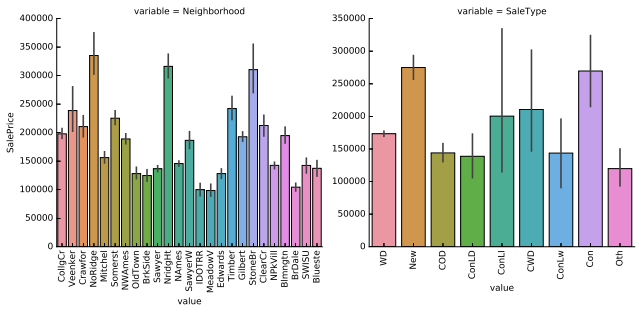


Figure 7: Graph - Neighborhood and Sale Type

Figure (8) and (9) shows the code snippet and the correlations plot. From that we can list the top 10 features those are strongly correlated with the target variable - *sale price*. We can visualize a few pair-wise correlation graphs with sale price for further detailed analysis. Figures (10) and (11) show how *overall quality*, *ground live area*, *garage cars* and *garage area* are positively correlated with *sale price*.

```
# python code
corr = alldata[numerical_features].corr()
mask = np.zeros_like(corr)
mask[np.triu_indices_from(mask)] = True
plt.figure(figsize = (15,8))
sns_plot = sns.heatmap(corr, cmap=YlGnBu,
linewidths=.5,  mask=mask, vmax=.3)
```

Figure 8: Code - Correlations

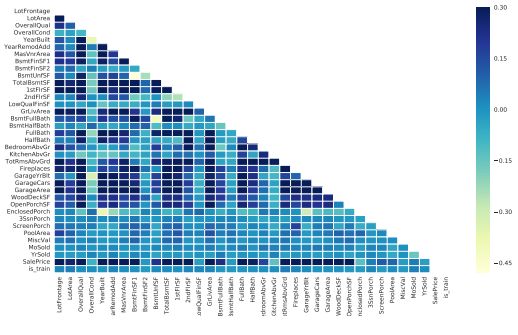


Figure 9: Graph - Correlations with Sale Price

(1) OverallQual: Overall material and finish quality
(2) GrLivArea: Above ground living area square feet
(3) GarageCars: Size of garage in car capacity
(4) GarageArea: Size of garage in square feet
(5) TotalBsmtSF: Total square feet of basement area
(6) 1stFlrSF: First Floor square feet
(7) FullBath: Full bathrooms above grade
(8) TotRmsAbvGrd: Total rooms above ground

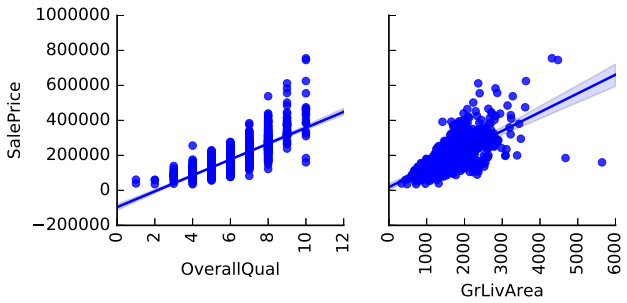


Figure 10: Graph - Overall Quality and Ground Live Area

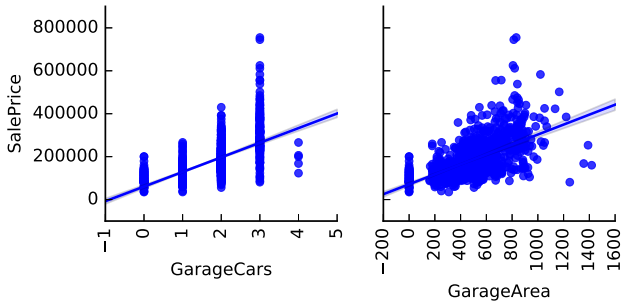


Figure 11: Graph - Garage Cars and Garage Area

(9) YearBuilt: Original construction date
(10) GarageYrBlt: Garage built year

4.5 Skewed Data Analysis

From the numerical analysis, we have identified that there are a few numerical variables need further analysis to identify the skewed data. We did not find any key variables those have skewed more than 75%. However, we wanted to replace the *sale price* with corresponding logarithmic value for the predictive models and later convert it back to the exponential value before saving the predictions. Figure (12) shows the *sale price*, before and after applying the logarithmic value.

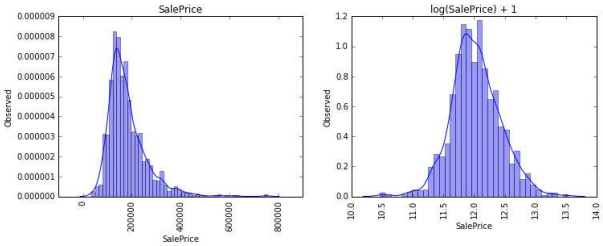


Figure 12: Graph - Sale Price skewness

4.6 Outlier Analysis

Continuing with exploratory analysis, we have analyzed the outliers using *Cooks distance*. Cooks distance is a measure calculated

from a regression model to find out the influence exerted by each observation (row) on the predictions. As a practice, those observations that have a Cooks distance greater than 4 times the mean value may be classified as an outlier. Outlier detection can be done using univariate and multivariate analysis. In univariate model, the outliers are those observations that are present outside of 1.5 * IQR, where IQR (*Inter Quartile Range*) is the difference between 75th and 25th quartiles. Analyzing outliers in any observations based on single variable may lead to incorrect inferences. Cooks distance generalizes the outlier analysis using multivariate approach [7]. Figure (13) is the code implementing Cooks distance to find the outliers from training dataset and Figure (14) shows the scatter plot with outliers being marked as bubbles. The bigger the bubbles, the bigger outlier deviations from the mean value. We have further analyzed two key variables - *ground live area* and *garage area* that are in high correlation with the *sale price.* From the scatter plot shown in Figure (15), we can see that *garage live area* has 4 outliers with values greater than 4,000 sq ft. We can also visualize 4 outliers in *garage area* scatter plot with values greater than 1,200 sq ft. as shown in Figure (16). We have removed the 8 outlier rows related to these two variables from the training dataset, the corresponding code snippet shown in Figure (17).

```
# python code - outlier analysis
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols(formula = SalePrice ~
GrLivArea + GarageArea, data=train)
fitted = model.fit()
plot = sm.graphics.influence_plot(fitted,
criterion=cooks)
```

Figure 13: Code - Outlier Analysis

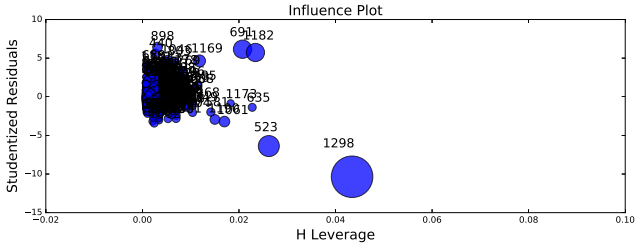


Figure 14: Graph - Outliers using Cooks distance

4.7 Feature Engineering

Feature engineering is a technique to analyze all the variables those influence target variable for better predictions. Part of feature engineering, we may need to create new features to make the data to be more expressive. One of the key intents, in analyzing categorical variables, is to convert them into numerical factors as most of the machine learning algorithms expect all the variables to be numeric for them to work more effectively. Feature engineering

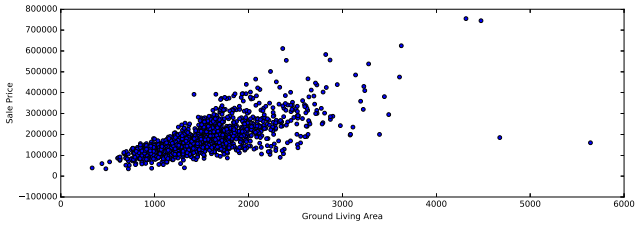


Figure 15: Graph - Garage Live Area Outliers

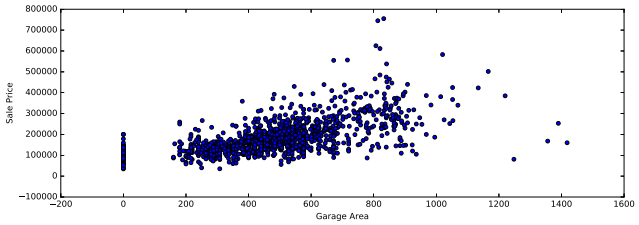


Figure 16: Graph - Garage Area Outliers

```
# python code - remove outlier rows
# fix all extreme outliers based on outlier analysis
# 8 rows will be deleted
train = train[train.GrLivArea <= 4000]
train = train[train.GarageArea <= 1200]
```

Figure 17: Code - Delete Outliers

is a difficult task; majority of the effort is manual and requires lots of domain knowledge.

4.7.1 Numerical Encoding. Some of the categorical variables are ordinal. we can use T-shirt sizes: small, medium and large as an example to explain an ordinal variable. When we convert this category variable into numeric encoding, we need to retain the fact that there is an implicit order within the values. Supposing, we give ordinal encoding as - small = 1, medium = 2 and large = 3; we will satisfy the implicit order or weightage and that helps in modeling the system by elevating the importance of this implicit ordering in the values of the ordinal variable.There are a few other encoding techniques, such as one-hot, binary, polynomial and helmert to factorize categorical variables. We will use ordinal and one-hot encoding techniques for this dataset. Following are a few categorical variables converted to numerical:

- *Lot shape* is encoded as: 1 - regular, 2 - Irregular-I, 3 - Irregular-II, 4 - Irregular-III
- *Alley* is encoded as: 1 - none, 2 - gravel, 3 - paved
- All quality variables such as *garage quality* are encoded as: 0 - none, 1 - poor, 2 - fair 3 - typical 4 - good, 5 - excellent
- *Building type* is encoded as: 1 - single-family, 2 - two-family, 3 - duplex, 4 - townhouse end unit, 5 - townhouse inside unit
- *Overall quality* is encoded as: 1 to 3 - bad, 4 to 6 - average, 7 to 10 - good

4.7.2 One-hot Encoding. One-hot encoding converts the category variable into many binary vectors, one new numeric variable for each value in the category. Assume that we have a categorical variable called signal-light with three possible values: green, yellow and red. We will need to convert these values into numeric - green = 1, yellow = 2 and red = 3. When we apply one-hot encoding on this variable, basically we are creating three new categorical variables - signal-light-green, signal-light-yellow and signal-light-red along with the original variable - signal-light, each is pretty much a binary vector having 1s for all the corresponding values; otherwise 0s. With hot-encoding, we are basically increasing dimensions in the model. After extensive feature engineering applied on the housing dataset, we have added *228* new features (variables). Figure (18) shows the python methods to factorize categorical variables using one-hot encoding techniques.

```
# python code - factorize and one-hot
def get_one_hot(df, col_name, fill_val):
if fill_val is not None:
df[col_name].fillna(fill_val, inplace=True)

dummies = pd.get_dummies(df[col_name], prefix=_ + col_name)
df = df.join(dummies)
df = df.drop([col_name], axis=1)
return df
#end def
```

Figure 18: Code - factorize and one-hot encoding

4.7.3 New Features. By adding new features that fill the gaps in domain model, we can guide the model predictions more accurately. We can, easily, create more meaningful new features from existing features, such as:

- What is the total area of the house? - This variable is sum of 18 existing variables that are contributing to the overall size of the house, such as *lot frontage*, *lot area*, *ground live area*, *pool area* and *garage area*.
- Whether house has been ever remodeled? - We can find this out using two variables: *year built* and *year remodel added*.
- House remodeled since? - We can find this out using two variables: *year sold* and *year remodel added*.
- Is it a very new house? - This can be calculated based on *year built*
- What is the age of the house? - This is a calculated value from *year build* (formula: 2010 - *year built*)
- When was it last sold? - This is a calculated value from *year build* (formula: 2010 - *year sold*)
- Which season house was last sold in? - This is a calculated value from *month build*

4.7.4 Handling Null Values.

- LotFrontage: Calculated the median of the LotFrontage grouping by neighborhood and assigned the median value for the homes with null values.
- Street: Filled null values with *Grvl*.
- Alley: Filled null values with *NA*.
- Lot Shape: Filled null values with *Reg*.
- Land Contour: Filled null values with *lvl*.
- Land Slope: Filled null values with *Gtl*
- Neighborhood_Good: Filled null values with 0
- YearRemodAdd: Filled null values with Year Built value.
- GarageYrBlt: Filled null values with 0
- Exterior1st: Filled null values with Mode of this variable
- Exterior2nd: Filled null values with Mode of this variable
- MasVnrArea: Filled null values with 0
- ExterQual: Filled null values with *TA* (numeric factor = 2)
- BsmtQual: Filled null values with *TA* (numeric factor = 2)
- BsmtFinType1: Filled null values with Mode of this variable
- BsmtFinType2: Filled null values with Mode of this variable
- PoolQC: There are entries with PoolArea > 0 and PoolQC as NA, so filled the values with average condition - *TA*

5 ALGORITHMS AND METHODOLOGY

Broadly speaking, there are three types of machine learning algorithms: supervised, unsupervised and reinforcement learning. *Supervised learning algorithms*: decision trees, linear regression, support vector machines (SVMs), Naive Bayes, neural networks, etc. are popular for classification and regression problems by analyzing labeled training data. K-means clustering algorithms are good for *unsupervised* datasets to categorize based on the identified patterns in unlabeled data. While there are so many factors - nature of the domain, sample size of the dataset and number of attributes defining characteristics of the data - decide which machine learning algorithm works better, Deep Learning neural network algorithms are, getting greater traction, addressing complex analytics tasks including high-dimensionality and automatic creation of new features from existing complex hierarchical features, very well. *Reinforcement Learning* algorithms focus on how to maximize the learning based on rewards and punishments.

Linear regression predicts the target variable using best possible straight line fit to the set predictor variables. The best fit is usually the one that minimizes the root mean squared error (RMSE) between the actual and predicted data points. Logistic regression solves classification problems to predict discrete values based on given training dataset. Simple decision tree algorithms also target solving classification problems. Naive Bayes is also popular for classification type of problems where it assumes independence among the variables. However, with complex problem space such as the housing prices dataset, we have lots of variables relating to the target variable in a non-linear fashion. Trivial supervised learning algorithms will not be effective to provide accurate *sale price* predictions. To overcome this challenge, we have applied various advanced supervised learning algorithms, such as Support Vector Machine (SVM), Random Forest, Lasso, Ridge, XGBoost and Neural Network, to predict the test data housing prices. Following are some of the aspects that are common to all the algorithms:

5.1 Underfitting and Overfitting

Underfitting happens when the model is trivial and does not fit the data properly. As a result it is unable to learn the model properly

hence gives incorrect predictions. Underfitting suffers from low *variance* but high *bias* from the predicted model. Variance measures the variation in learning from different training sets. Variance does not properly filter outliers that are part of the model. Bias prevents generalization beyond the training dataset. Overfitting occurs when the predicted model learns the training dataset including the noise and results negatively impacting the performance and accuracy of the model. Overfitting happens more likely with non-linear and non-parametric algorithms those offer more flexibility. Overfitting, as expected, exhibits low *bias* and high *variance*. Balancing between bias and variance is a challenge and model may have to compromise one over the other.

5.2 Cross Validation

Before applying the trained model onto the testing dataset, we need to validate it. Cross-validation is a technique to validate the trained model by partitioning the original training dataset into two parts - training and cross validation datasets. The cross validation dataset is basically to evaluate the trained model before applying on the actual test dataset. Usually 70% of the original training dataset is kept for training the model and 30% of it for cross validation. This type of cross validation is called *holdout method. K-fold cross validation* is more improved and effective cross validation method, where the dataset is divided into *k* subsets, and the *holdout* method is repeated k times. In each iteration, one of the k subsets is selected as a test dataset and the remaining *k-1* subsets will be part of the training dataset. In the end, the average error across all *k* attempts is computed.

5.3 Model Evaluation

We can use evaluation metrics to check accuracy of the trained model. Accuracy, precision, recall and f-score are typical metrics used to evaluate the classification models. Regression models use mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination and relative scored error (RSE) as metrics to verify the accuracy and performance of the model. MAE evaluates how close the predictions are to the actual target variable, hence a lower score is better. RMSE metric summarizes the error in predicted model. By squaring the error, the over and under predictions are controlled. RSE normalizes the total squared error by diving the total squared error of the actual target variable values. Coefficient of determination (R^2) represents the prediction as value between 0 and 1; 0 - means the model is random and 1 - means the model is a good fit.

5.4 Support Vector Machine (SVM) Algorithm

Support Vector Machine (SVM) algorithms can be used to solve classification and regression problems. SVM creates larger margins between categories of data so that they are linearly separable. SVM regression relies on kernel functions for modeling the data. SVM handles non-linearly separable data, mainly for regression problems, using kernel functions, such as polynomial, radial basis function (RBF) and sigmoid, to project the data onto a hyperplane. Figure (19) shows the python implementation for *sale price* predictions of the housing test dataset.

```
# python code - SVM algorithm
from sklearn.svm import SVR

_svm_algo = SVR(kernel = rbf, C=1e3, gamma=1e-8)
_svm_algo.fit(train, target_vector)

y_train = target_vector
y_train_pred = _svm_algo.predict(train)

#root mean squared error (RMSE)
rmse_train = np.sqrt(rmse(y_train,y_train_pred))

y_test_pred = _svm_algo.predict(test)
```

Figure 19: Code - SVM Algorithm

We have used *sklearn.svm* package to implement the SVM algorithm in Python. The SVM kernel used, for the *sale price* prediction, is *radial basis function*. Cost parameter, with a value of *1e3*, is used to increase the margin for better linear separability. Gamma controls the trade-off between error due to bias and variance in the trained model. We have used gamma value as *1e-8*. Once the SVM algorithm is instantiated, we fit the model by passing the training dataset and the *sale prices* vector of the training dataset as *Y* - target variable. After training the model is done, we checked the *root mean squared error* (RMSE) of the trained model predictions with the actuals to make sure the desired accuracy is being met. In this case, RMSE is calculated as 0.2069. Finally we predict the *sale price* of test dataset and make sure the sale prices are meaningful.

5.5 Random Forest Algorithm

Random Forest is an advanced machine learning algorithm for predictive analytics. Random Forest ensembles multiple decision trees to create an additive learning model from the sequence of base models created by each decision tree that worked on a subsample dataset. Random Forest models are suitable to handle tabular datasets with hundreds of numeric and categorical features. Along with missing values, non-linear relations between features and the target, will be handled well by random forest algorithms. By tuning the hyper-parameters of the random forest algorithm, it can perform well with decent accuracy in the predictions without overfitting the model. Random forest uses *voting* concept to predict the target variable and selects the highly voted predicted values as the final selected predictions. Unlike similar regression models, it does not offer feature coefficient information but it provides *feature ranking* functionality very nicely. Figure (20) shows the random forest algorithm details for the *sale price* predictions implemented using *sklearn* package and the Figure (21) shows the top 10 important features selected by random forest to model the predictions.

We have used *sklearn.ensemble* and *sklearn.metrics* packages to implement random forest algorithm. The hyper-parameters used in this algorithm are: n_estimators = 100, oob_score = True and random_state = 123456. Parameter: *n_estimators* is for the number of trees in the random forest. Parameter: *oob_score* is a boolean to indicate whether to use out-of-bag samples to estimate the generalization accuracy. Parameter: *random_state* is used as seed for

```
# python code - random forest algorithm
from sklearn.ensemble import RandomForestRegressor

_algo = RandomForestRegressor(n_estimators=100,
oob_score=True, random_state=123456)
model = _algo.fit(train, target_vector)

feat_imp = pd.Series(_algo.feature_importances_,
train.columns).sort_values(ascending=False)
feat_imp[:10].plot(kind=bar,
title=Feature Ranmkingt)
y_train = target_vector
y_train_pred = _algo.predict(train)

#root mean squared error (RMSE)
rmse_train = np.sqrt(rmse(y_train,y_train_pred))
y_test_pred = _algo.predict(test)
```

Figure 20: Code - Random Forest Algorithm

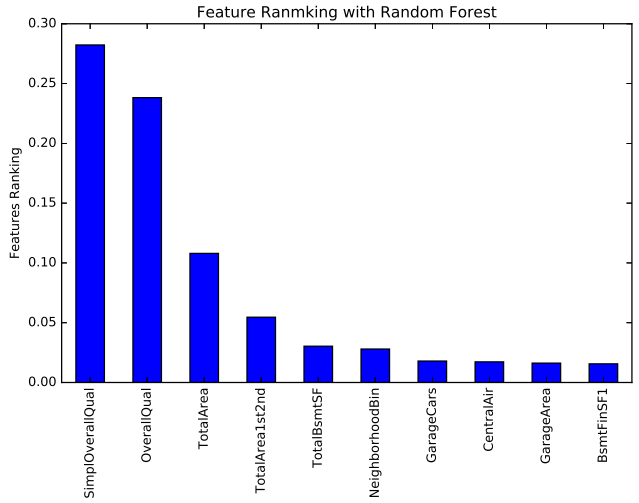


Figure 21: Graph - Random Forest Feature Ranking

the random number generator. Training dataset and the *Sale Price* vector are used as input to fit the model and verified the predicted output of the training dataset. The RMSE is calculated as 0.0519. Finished the implementation by predicting the *sale price* of the test dataset.

5.6 Lasso Algorithm

There are four techniques to model the predictions using linear regression - (1) simple linear regression, (2) Ordinary Least Squares (OLS), (3) Gradient Descent and (4) Regularization. First two techniques are basically using statistical analysis to calculate the coefficients, such as means, standard deviation, covariance and correlations, where as gradient descent uses sum of the squared errors to scale down the coefficients to minimize the error. Regularization not only minimizes the squared error in the training dataset but also reduces complexity of the overall model. Two well known algorithms to create the regularized regression models are - Lasso and Ridge regressions. While lasso performs *L1 regularization*, ridge applies *L2 regularization* techniques in modeling the predictions. L1 regularization adds penalty to the variables equivalent to *absolute value of the magnitude* of the coefficients, where as L2 adds the penalty equivalent to *square of the magnitude* of the variable coefficients.

Lasso is a regression model that uses shrinkage to bring data points towards the center, similar to the mean value of all the data points. Lasso stands for Least Absolute Shrinkage and Selection Operator. It is a regularized linear model with penalty term *lambda* to minimize the error. Parameter penalization controls overfitting the input data by shrinking variable coefficients to 0. Essentially this makes the variables no effect in the model, hence reduces the dimensions. Figure (22) shows the lasso algorithm implementation for *sale price* predictions in python.

```
# python code - lasso algorithm
from sklearn.linear_model import Lasso

_best_alpha = 0.0001
_lasso_algo = Lasso(alpha = _best_alpha,
max_iter = 50000)
model = _lasso_algo.fit(train, target_vector)

y_train = target_vector
y_train_pred = _algo.predict(train)

#root mean squared error (RMSE)
rmse_train = np.sqrt(rmse(y_train,y_train_pred))
y_test_pred = _lasso_algo.predict(test)
```

Figure 22: Code - Lasso Algorithm

We have used *sklearn.linear_model* package to implement lasso algorithm. *sklearn.metrics* is used for RMSE calculations. The two key hyper-parameters used in this algorithm are: alpha = 0.0001 and max_iter = 50000. The parameter: *alpha* is used as a constant term that multiplies the L1 term. L1 is explained in the following Ridge Algorithm section. We have given alpha by finding the best value through cross validation. Training data and the *sale price* are sent to the *fit* method to fit the model. RMSE is calculated as 0.1015 to evaluate the accuracy of the trained model against the *sale price* of the training dataset. Finally the algorithm predicted the *sale price* of the test dataset.

5.7 Ridge Algorithm

Ridge algorithm is very similar to lasso algorithm with the same goal. Ridge uses *alpha* parameter to control the balance between minimizing the *residual sum of squares* (RSS) and minimizing sum of squares of square of coefficients. As the alpha value increases, the complexity of the model decreases. However, significant value of alpha might cause the model to go underfitting. Figure (23) shows the python implementation of the ridge algorithm for the *sale price*

predictions. Figures (24) and (25) show the top 10 positively and top 10 negatively influencing variables with *sale price*.

```
# python code - ridge algorithm
from sklearn.linear_model import Ridge
#found this best alpha value through cross-validation
_best_alpha = 0.00099
_ridge_algo = Ridge(alpha = _best_alpha,
normalize = True)
_ridge_algo.fit(train, target_vector)

y_train = target_vector
y_train_pred = _ridge_algo.predict(train)

#root mean squared error (RMSE)
rmse_train = np.sqrt(rmse(y_train,y_train_pred))

y_test_pred = _ridge_algo.predict(test)
```

Figure 23: Code - Ridge Algorithm

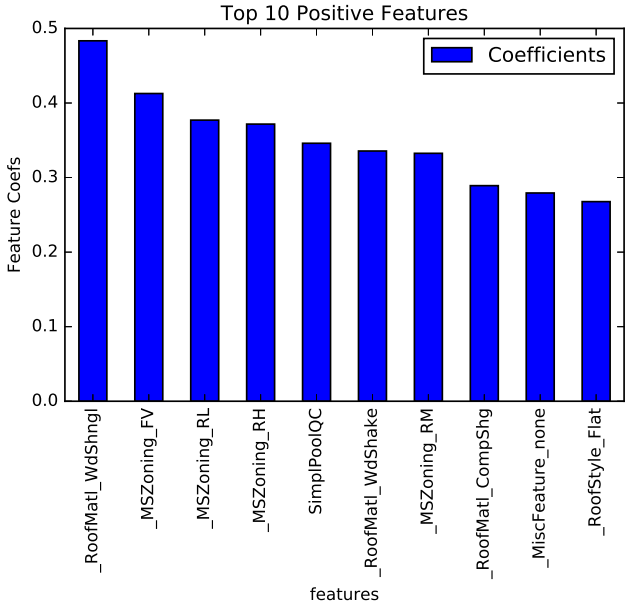


Figure 24: Graph - Ridge Top 10 Positive Features

We have used *sklearn.linear_model* package to implement ridge algorithm. The two key hyper-parameters used in this algorithm are: alpha = 0.00099 and normalize = True. The parameter: *alpha* is used denote the *regularization* strength. We have given alpha by finding the best value through cross validation. Parameter: *normalize* is used when the value is True to normalize the regressors *X* before regression by subtracting the mean and dividing by the L2 norm. Training data and the *sale price* are sent to the *fit* method to fit the model. RMSE is calculated as 0.09888 to evaluate the accuracy

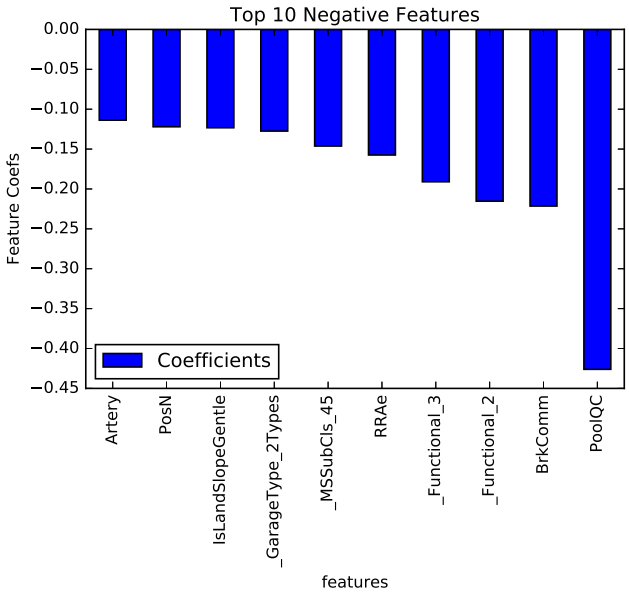


Figure 25: Graph - Ridge Top 10 Negative Features

of the trained model against the *sale price* of the training dataset. We have also extracted the top 10 positive and negative features influencing the target variable - *sale price*. Finally the algorithm predicted the *sale price* of the test dataset.

5.8 XGBoost Algorithm

XGBoost (eXtreme Gradient Boosting) is one of the Gradient Boosted Machine algorithms. It ensembles (combines) optimized model by taking trained models from all the preceding iterations. XGBoost regularizes the variables (parameters) using L1 and L2 regularizations to reduce the overfit and can work well with variables having missing values. It is empowered with built-in cross validation to reduce the boosting iterations; hence offers better performance along with parallel processing on multi-core CPU and also can also work with very large datasets in distributed environments. Execution speed and the performance of model creation make XGBoost a very good choice. There are many variations of XGBoost, such as gradient boosting machine, stochastic gradient boosting and additive regression tree, and all of them use gradient descent methods to minimize the loss function. By tuning the XGBoost hyper parameters, we can achieve well optimized model that can make more accurate predictions. XGBoost uses *F-Score* to measure the importance of variables. Our implementation of *sale price* predictions using XGBoost is shown in Figure (26). Following list explains the hyper-parameters of XGBoost algorithm.

- Maximum Iterations - Number of trees in the final model. More the trees, more accuracy.
- Maximum Depth - Depth of each individual tree to control overfitting.
- Step Size - Shrinkage, works similar to learning rate; smaller value takes more iterations.

[9]

- Column Subsample - Subset of the columns to use in each iteration.

```
# python code - XGBoost algorithm
import xgboost as xgb

_xgb_algo = xgb.XGBRegressor(
colsample_bytree=0.8,
colsample_bylevel = 0.8,
gamma=0.01,learning_rate=0.05,
max_depth=5, min_child_weight=1.5,
n_estimators=6000, reg_alpha=0.5,
reg_lambda=0.5,subsample=0.7,
seed=42, silent=1)

_xgb_algo.fit(train, target_vector)

y_train = target_vector
y_train_pred = _xgb_algo.predict(train)

#root mean squared error (RMSE)
rmse_train = np.sqrt(rmse(y_train,y_train_pred))
y_test_pred = _xgb_algo.predict(test)
```

Figure 26: Code - XGBoost Algorithm

We have used *xgb* package to implement the XGBoost algorithm. Various hyper-parameters are used to tune the algorithm and a few of them are explained in the above list. The training dataset and the *sale price* vector are used to fit the model. Top 10 features selected by XGBoost algorithm are: TotalArea, LotArea, GarageArea, All_Liv_SF, BsmtUnfSF, 1stFlrSF, BsmtFinSF1, LotFronage, TotalBsmtSF and Age, including the hot-encoded variables and new features created through feature engineering. We have captured the *feature ranking* as a graph and evaluated accuracy of the predictions by calculating RMSE on training dataset. Finally, we have predicted *sale prices* of the test dataset. Figure (27) shows the top 10 features selected by the XGBoost.

5.9 Neural Network Algorithm

Neural Network is, a *directed graph*, organized by layers and layers are created by number of interconnected neurons (or nodes). Every neuron in a layer is connected with all the neurons from previous layer; there will be no interaction of neurons within a layer. The performance of a Neural Network is measured using *cost or error function* and the dependent input *weight* variables. *Forward-propagation* and *back-propagation* are two techniques, neural network uses repeatedly until all the input variables are adjusted or calibrated to predict accurate output. During, forward-propagation, information moves in forward direction and passes through all the layers by applying certain weights to the input parameters. *Back-propagation* method minimizes the error in the *weights* by applying an algorithm called *gradient descent* at each iteration step.

Deep Learning is an advanced neural network, with multiple hidden layers (thousands or even more deep), that can work well with supervised (labeled) and unsupervised (unlabeled) datasets.

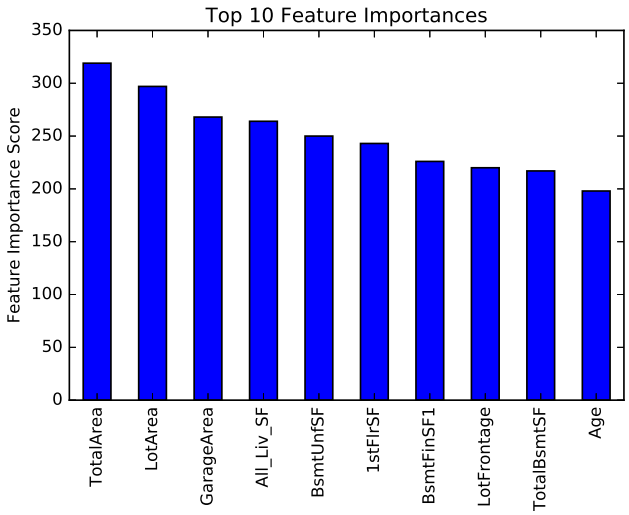


Figure 27: Graph - XGBoost Feature Importance

Applications, such as speech, image and behavior patterns, having complex relationships in large-set of attributes, are best suited for Deep Learning Neural Networks. Deep Learning vectorizes the input and converts it into output vector space by decomposing complex geometric and polynomial equations into a series of simple transformations. These transformations go through neuron activation functions at each layer parameterized by input weights. Deep Learning makes neurons learn new features themselves, in an unsupervised manner, from existing features distributed in several hidden layers. *Stacked Autoencoder* (AE) is, a Deep Belief Network algorithm, to create advanced predictive models for large datasets having thousands or even millions of dimensions, automatically, with complex hierarchical attributes in non-linear fashion for simpler computing. Though AE is sophisticated, it is very difficult to understand the algorithm logic and so unable to reuse the learnings from the modeling to other systems. We have used *TensorFlow* python library to predict the *sale price* of housing dataset using simple feed-forward neural network. TensorFlow uses *tensors*, special multi-dimensional arrays to store the datasets for easier linear algebra and vector calculus operations.

We have implemented Neural Network algorithm by creating a TensorFlow based work-flow. We have created various objects, such as input variables, loss computation, optimizer and the predictions for TensorFlow to create the model. TensorFlow tunes the hyper-parameters using number of cross-validation iterations before finally predicting the *sale prices* of test dataset. Deep Learning, by design, allows parallel programming, as each module - with all the dependencies among neurons - can run independently and parallelly from other modules within the network. Using Graphics Process Unit (GPU), module networks can achieve parallel programming without needing much of Central Processing Unit (CPU) allocation of a computer. Though GPU is intended for graphical processing, it works efficiently to run thousands of small mathematical functions, such as matrix multiplications, in parallel. Cloud computing and edge analytics offer flexible scale out distributed processing

options using virtualization and containerization. Sophisticated algorithms and distributed computing make Deep Learning scale and perform well to process huge datasets.

5.10 Model Ensembling

We can create a robust predictive model with better accuracy by merging two or more machine learning algorithms. This technique is called *model ensembling*. Ensembled algorithms may be similar in functionality or may entirely be different from each other. Individual algorithms may not perform great but by ensembling them, the overall system can offer much better performance and accuracy. Variations in the predicting logic in each of these individual algorithms will bring unbiasedness into the unified model. *Bagging*, *boosting* and *stacking* are popular ensembling techniques. Many of the advanced machine learning algorithms use ensembled approaches to achieve accurate classifications or predictions. Random Forest uses bagging, XGBoost uses boosting and Neural Network applies stacking ensembling techniques. To optimize the predictions, we have created an ensembled model by averaging *Sale Price* of the top 3 performing ensembled algorithms - XGBoost, Lasso and Neural Network. As predicted, ensembled model has predicted better with less RMSE (*root mean squared error*), compared to all the individual algorithms. Following list displays each algorithm and the corresponding *root mean squared error* (RMSE).

- SVM - RMSE = 0.2069
- Random Forest - RMSE = 0.0519
- XGBoost - RMSE = 0.0432
- Lasso - RMSE = 0.1015
- Ridge - RMSE = 0.0988
- Neural Network - RMSE = 0.20

6 DEVELOPMENT ENVIRONMENT

6.1 OS and Programming Language

We have used *Ubuntu 16.4* Operating System that runs in Windows 10 through Oracle Virtual Box 5.2. Python 2.7 has been used as the programming language for this project. Data visualizations are done using *seaborn* and *matplotlib* packages. Most of the algorithms implemented in this project are using *sklearn* package. For the neural network algorithm, we have used *tensorflow* package as it offers simple programming interface to the complex processing needed by the algorithm. Our code is placed in gitbub repository at *git@github.com:bigdata-i523/hid306.git*.

6.2 Project Folder Structure

Project is organized in three folders - code, data and images. Code folder has all the python code files. Data folder contains the *house pricing* sample datasets that we used for the exploratory analysis and *sale price* predictions. We also stored all the *sale price* prediction output files from various algorithms in the data folder. Images folder contains all the data visualization files that we have created during the analysis and in processing the algorithms. We wanted to create interactive and sharable code files that contain not only the python code but also corresponding explanation along with data visualizations. Jupyter Notebook application is ideal for such facilitation with python code components. Using Jupyter Notebook, it would be easy to share live code with the reviewers. Such environment allows to explore the code-base easily along with the interactive code execution and visualize all the corresponding exploratory analysis results with the graphs.

6.3 Project Files

We have a total of 11 Jupyter Notebook driven python code files. First 4 files are focused on doing the exploratory data analysis and the next 6 files are meant for six supervised machine learning algorithms - SVM, Random Forest, Ridge, Lasso, Neural Network and XGBoost. Last code file is dedicated for ensembling the top 3 algorithms with best predictions of housing sale prices in the test dataset. We have named them in a sequence as there is an implicit order in the execution of these files. We wanted to do the data analysis first before running the predictive algorithms.

6.4 List of Code Files

Following is the list of code files:

- Exploratory Analysis Numerical - To load datasets and analyze all numerical variables
- Exploratory Analysis Categorical - To analyze categorical variables in the dataset
- Outlier And Skewed Data Analysis - Handles outlier and skewed data analysis
- Feature Engineering - All the feature engineering is done in this file
- SVM Algorithm - Implementation of SVM algorithm
- Random Forest Algorithm - Implementation of Random Forest algorithm
- Ridge Algorithm - Implementation of Ridge algorithm
- Lasso Algorithm - Implementation of Lasso algorithm
- Neural Network Algorithm - Implementation of Neural Network algorithm
- XGBoost Algorithm - Implementation of XGBoost algorithm
- Ensembled Model - Implementation of Ensembled algorithm

6.5 List of Data Files

Following is the list of data files:

- Housing Dataset with Sale Price - Sample training dataset with housing attributes along with the sale price
- Housing Dataset without Sale Price - Sample testing dataset similar to training dataset without the sale price
- SVMs Algorithm Predictions - Predicted Housing Sale Prices from SVM algorithm
- Random Forest Algorithm Predictions - Predicted Housing Sale Prices from Random Forest algorithm
- Ridge Algorithm Predictions - Predicted Housing Sale Prices from Ridge algorithm
- XGBoost Algorithm Predictions - Predicted Housing Sale Prices from XGBoost algorithm
- Lasso Algorithm Predictions - Predicted Housing Sale Prices from Lasso algorithm

- Neural Network Predictions - Predicted Housing Sale Prices from Neural Network algorithm
- Ensembled Model Predictions - Predicted Housing Sale Prices from Ensembled algorithm

7 CONCLUSION

Though the datasets we have used are old, the complexity of the dataset is challenging enough for us to find the optimized algorithms to get the accurate predictions. Generally, ensemble models performs better compared to individual algorithms. However, there are a few factors that influence accuracy and performance of the algorithms, such as handcrafted feature engineering, proper cost function with regularized input to address non-linearities in the training datasets and tuning hyper-parameters of the algorithms. While Deep Learning Neural Networks are good for image processing, K-Nearest Neighbor algorithms can handle unsupervised datasets with less complexity. Domain knowledge and algorithm selection play vital role in getting accurate predictions. XGBoost, Random Forest, Lasso and Neural Networks are advanced machine learning algorithms dominating in the Big Data analytics for classification and regression related tasks. With ensembling and iterative learning techniques, they can scale well and offer better predictions for huge datasets having large number of features.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski and the Teaching Assistants for their support and great suggestions. Authors would also want to thank Kaggle Website for the sample datasets and the contributed developers for their valuable information, ideas and contributions.

REFERENCES

[1] AiO. 2017. House Prices: Advanced Regression Techniques. (Feb. 2017). https://www.kaggle.com/notaapple/detailed-exploratory-data-analysis-using-r
[2] Tanner Carbonati. 2017. Detailed Data Analysis & Ensemble Modeling. (Aug. 2017). https://www.kaggle.com/tannercarbonati/detailed-data-analysis-ensemble-modeling/notebook
[3] Yeshwant Chillakuru, Michael Arango, Jack Crum, and Paul Brewster. 2017. Using Neighborhood Level Data to Predict the Residential Sale Price of Properties in Ames, Iowa. (May 2017). https://rpubs.com/jackcrum/281471
[4] Aarshay Jain. 2016. Complete Guide to Parameter Tuning in XGBoost. (March 2016). https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/
[5] Charles McLellan. 2015. The internet of things and big data: Unlocking the power. (March 2015). http://www.zdnet.com/article/the-internet-of-things-and-big-data-unlocking-the-power/
[6] James O'brien. 2014. 5 Ways Big Data Is Changing Real Estate. (July 2014). http://mashable.com/2014/07/09/big-data-real-estate/#_dujSE2o.gq2
[7] Selva Prabhakaran. 2017. Outlier Treatment. (Dec. 2017). http://r-statistics.co/Outlier-Treatment-With-R.html
[8] Siddharth Raina. 2017. Regularized Regression - Housing Pricing. (Jan. 2017). https://www.kaggle.com/sidraina89/regularized-regression-housing-pricing
[9] Kevin Rands. 2017. 8 companies using big data to disrupt real estate. (Aug. 2017). https://www.cio.com/article/3211601/data-science/8-companies-using-big-data-to-disrupt-real-estate.html
[10] Athena Snow. 2017. Why Big Data is a Game Changer for Agents. (May 2017). https://www.coldwellbanker.com/blog/cbx-app-game-changer-for-agents/
[11] Kevin Wong. 2016. Predicting Ames House Prices. (Dec. 2016). http://kevinfw.com/post/predicting-ames-house-prices/
[12] Cnarlie Young. 2017. Big data takes over real estate: The best tech for attracting buyers and satisfying sellers. (May 2017). https://www.inman.com/2017/05/12/big-data-takes-real-estate-best-tech-attracting-buyers-satisfying-sellers/
[13] Ricky Yue and Jurgen De Jager. 2016. Advanced Regression Modeling on House Prices. (Sept. 2016). https://nycdatascience.com/blog/student-works/advanced-regression-modeling-house-prices/

A SAMPLE DATASET FILE DETAILS

The training and testing sample datasets contain the same variables explaining the housing real estate aspects. Training dataset contains the sale price information whereas the testing dataset does not the sale price as that is the target variable we need to predict using supervised machine learning algorithm. Following are the list of variables describing the housing real estate domain. Good understanding of the domain is needed for better exploratory data analysis and to apply the matching machine learning algorithms to the problem space.

- Id: Row Id
- SalePrice: Sale price of the house in dollars. This is the target variable to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition

- CentralAir: Central air conditioning Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: Dollar Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

A.1 Factorization of categorical variables

Following are the factorized categorical variable details:

A.1.1 Street (Nominal). : Type of road access to property

- Grvl - Gravel
- Pave - Paved

A.1.2 Alley (Nominal). : Type of alley access to property

- Grvl- Gravel
- Pave - Paved
- NA - No alley access

A.1.3 Lot Shape (Ordinal). : General shape of property

- Reg - Regular
- IR1 - Slightly irregular
- IR2 - Moderately Irregular
- IR3 - Irregular

A.1.4 Land Contour (Nominal). : Flatness of the property

- Lvl - Near Flat /Level
- Bnk - Banked - Quick and significant rise from street grade to building
- HLS - Hillside - Significant slope from side to side
- Low - Depression

A.1.5 Land Slope (Ordinal). : Slope of property

- Gtl - Gentle slope
- Mod - Moderate Slope
- Sev - Severe Slope

A.1.6 Utilities (Ordinal). : Type of utilities available

- AllPub - All public Utilities (E,G,W, and S)
- NoSewr - Electricity, Gas, and Water (Septic Tank)
- NoSeWa - Electricity and Gas Only
- ELO - Electricity only

A.1.7 Lot Config (Nominal). : Lot configuration

- Inside - Inside lot
- Corner - Corner lot
- CulDSac - Cul-de-sac
- FR2 - Frontage on 2 sides of property
- FR3 - Frontage on 3 sides of property

A.1.8 Neighborhood (Nominal). : Physical locations within Ames city limits (map available)

- Blmngtn - Bloomington Heights
- Blueste - Bluestem
- BrDale - Briardale
- BrkSide - Brookside
- ClearCr - Clear Creek
- CollgCr - College Creek
- Crawfor - Crawford
- Edwards - Edwards
- Gilbert - Gilbert
- Greens - Greens
- GrnHill - Green Hills
- IDOTRR - Iowa DOT and Rail Road
- Landmrk - Landmark
- MeadowV - Meadow Village
- Mitchel - Mitchell
- Names - North Ames
- NoRidge - Northridge
- NPkVill - Northpark Villa
- NridgHt - Northridge Heights
- NWAmes - Northwest Ames
- OldTown - Old Town
- SWISU - South and West of Iowa State University
- Sawyer - Sawyer
- SawyerW - Sawyer West
- Somerst - Somerset
- StoneBr - Stone Brook
- Timber - Timberland
- Veenker - Veenker

A.1.9 Condition 1 (Nominal). : Proximity to various conditions

- Artery - Adjacent to arterial street
- Feedr - Adjacent to feeder street
- Norm - Normal
- RRNn - Within 200 feet of North-South Railroad
- RRAn - Adjacent to North-South Railroad
- PosN - Near positive off-site feature–park, greenbelt, etc.
- PosA - Adjacent to positive off-site feature
- RRNe - Within 200 feet of East-West Railroad
- RRAe - Adjacent to East-West Railroad

A.1.10 Condition 2 (Nominal). : Proximity to various conditions (if more than one is present)

- Artery - Adjacent to arterial street
- Feedr - Adjacent to feeder street
- Norm - Normal
- RRNn - Within 200 feet of North-South Railroad
- RRAn - Adjacent to North-South Railroad
- PosN - Near positive off-site feature–park, greenbelt, etc.
- PosA - Adjacent to positive off-site feature
- RRNe - Within 200 feet of East-West Railroad
- RRAe - Adjacent to East-West Railroad

A.1.11 Bldg Type (Nominal). : Type of dwelling

- 1Fam - Single-family Detached
- 2FmCon - Two-family Conversion; originally built as one-family dwelling
- Duplx - Duplex
- TwnhsE - Townhouse End Unit
- TwnhsI - Townhouse Inside Unit

A.1.12 Variable: MS Zoning. MS Zoning (Nominal): Identifies the general zoning classification of the sale.

- A - Agriculture
- C - Commercial
- FV - Floating Village Residential
- I - Industrial
- RH - Residential High Density
- RL - Residential Low Density
- RP - Residential Low Density Park
- RM - Residential Medium Density

A.1.13 House Style (Nominal). : Style of dwelling

- 1Story - One story
- 1.5Fin - One and one-half story: 2nd level finished
- 1.5Unf - One and one-half story: 2nd level unfinished
- 2Story - Two story
- 2.5Fin - Two and one-half story: 2nd level finished
- 2.5Unf - Two and one-half story: 2nd level unfinished
- SFoyer - Split Foyer
- SLvl - Split Level

A.1.14 Overall Qual (Ordinal). : Rates the overall material and finish of the house

- 10 - Very Excellent
- 9 - Excellent
- 8 - Very Good
- 7 - Good
- 6 - Above Average
- 5 - Average
- 4 - Below Average
- 3 - Fair
- 2 - Poor
- 1 - Very Poor

A.1.15 Overall Cond (Ordinal). : Rates the overall condition of the house

- 10 - Very Excellent
- 9 - Excellent
- 8 - Very Good
- 7 - Good
- 6 - Above Average
- 5 - Average
- 4 - Below Average
- 3 - Fair
- 2 - Poor
- 1 - Very Poor

A.1.16 Roof Style (Nominal). : Type of roof

- Flat - Flat
- Gable - Gable
- Gambrel - Gabrel (Barn)
- Hip - Hip
- Mansard - Mansard
- Shed - Shed

A.1.17 Roof Matl (Nominal). : Roof material

- ClyTile - Clay or Tile
- CompShg - Standard (Composite) Shingle
- Membran - Membrane
- Metal - Metal
- Roll - Roll
- Tar and Grv - Gravel and Tar
- WdShake - Wood Shakes
- WdShngl - Wood Shingles

A.1.18 Exterior 1 and 2 (Nominal). : Exterior covering on house

- AsbShng - Asbestos Shingles
- AsphShn - Asphalt Shingles
- BrkComm - Brick Common
- BrkFace - Brick Face
- CBlock - Cinder Block
- CemntBd - Cement Board
- HdBoard - Hard Board
- ImStucc - Imitation Stucco
- MetalSd - Metal Siding
- Other - Other
- Plywood - Plywood
- PreCast - PreCast
- Stone - Stone
- Stucco - Stucco
- VinylSd - Vinyl Siding
- Wd Sdng - Wood Siding
- WdShing - Wood Shingles

A.1.19 Mas Vnr Type (Nominal). : Masonry veneer type

- BrkCmn - Brick Common
- BrkFace - Brick Face
- CBlock - Cinder Block
- None - None
- Stone - Stone

A.1.20 Bsmt Cond, Exter Qual and Exter Cond (Ordinal). : Evaluates the quality of the material on the exterior

- Ex - Excellent
- Gd - Good
- TA - Average/Typical
- Fa - Fair
- Po - Poor

A.1.21 Foundation (Nominal). : Type of foundation

- BrkTil - Brick and Tile
- CBlock - Cinder Block
- PConc - Poured Concrete
- Slab - Slab
- Stone - Stone
- Wood - Wood

A.1.22 Bsmt Qual (Ordinal). : Evaluates the height of the basement

- Ex - Excellent (100+ inches)
- Gd - Good (90-99 inches)
- TA - Typical (80-89 inches)
- Fa - Fair (70-79 inches)
- Po - Poor (¡70 inches)
- NA - No Basement

A.1.23 Bsmt Exposure (Ordinal). : Refers to walkout or garden level walls

- Gd - Good Exposure
- Av - Average Exposure (split levels or foyers typically score average or above)
- Mn - Minimum Exposure
- No - No Exposure
- NA - No Basement

A.1.24 BsmtFin Type 1 (Ordinal). : Rating of basement finished area

- GLQ - Good Living Quarters
- ALQ - Average Living Quarters
- BLQ - Below Average Living Quarters
- Rec - Average Rec Room
- LwQ - Low Quality
- Unf - Unfinished
- NA - No Basement

A.1.25 BsmtFinType 2 (Ordinal). : Rating of basement finished area (if multiple types)

- GLQ - Good Living Quarters
- ALQ - Average Living Quarters
- BLQ - Below Average Living Quarters
- Rec - Average Rec Room
- LwQ - Low Quality
- Unf - Unfinished
- NA - No Basement

A.1.26 Heating (Nominal). : Type of heating

- Floor - Floor Furnace
- GasA - Gas forced warm air furnace
- GasW - Gas hot water or steam heat
- Grav - Gravity furnace
- OthW - Hot water or steam heat other than gas
- Wall - Wall furnace

A.1.27 Electrical (Ordinal). : Electrical system

- SBrkr - Standard Circuit Breakers and Romex
- FuseA - Fuse Box over 60 AMP and all Romex wiring (Average)
- FuseF - 60 AMP Fuse Box and mostly Romex wiring (Fair)
- FuseP - 60 AMP Fuse Box and mostly knob and tube wiring (poor)
- Mix - Mixed

A.1.28 HeatingQC (Ordinal). : Heating quality and condition

- Ex - Excellent
- Gd - Good
- TA - Average/Typical
- Fa - Fair
- Po - Poor

A.1.29 Central Air (Nominal). : Central air conditioning

- N - No
- Y - Yes

A.1.30 KitchenQual (Ordinal). : Kitchen quality

- Ex - Excellent
- Gd - Good
- TA - Typical/Average
- Fa - Fair
- Po - Poor

Big Data Applications in Real Estate Analysis

Elena Kirzhner
Indiana University Bloomington
3209 E 10th St
Bloomington, Indiana 47408
ekirzhne@iu.edu

ABSTRACT

Big Data analysis reveals and comforts buyers with knowledge and facts about the neighborhood, its people and trends. Reducing risk of buying and predicting changes in home value for potential buyers.

KEYWORDS

i523, hid320, Big Data Applications and Analytics, Real Estate

1 INTRODUCTION

When one mentions American dream, home ownership is first aspect that comes to mind. Another part of the American dream is financial success and wealth building. Buying your dream home to raise the family is obvious part of the real estate. For most Americans buying a home is the largest purchase they will ever make. Coupled with the fact that most conventional mortgages span 30 years research and analysis required to make educated choice should not be taken lightly as it will have implications on lifestyle for practically 40 percent of your lifetime. Successful investment in your home, potential rental property or land can lead to financially windfall. Failure to make right choices in real-estate purchases may have disastrous consequences. Financial ruin is obvious part of the equation. Majority of divorces in the united states are caused by financial duress in the households. Resulting in stress negatively affecting one's health.

The latest trend in real estate is application of Big Data. Big Data manipulation is booming and transforming the industry. We are seeing a huge move in usage of Big data and analytics. Companies build property matching online software based on customers behavior and their needs. The opportunities of Big Data are truly endless. It creates the power to change our thinking in decision making and develops efficient business approach by extracting variety of collected data points and reducing risks for consumers.

Big Data is already changing real estate industry by optimizing consumers search, offers recommendations on real estate websites to potential buyers and sellers. Utilizing Big Data in real estate could match customers with their desired home. It might include how many bedrooms they need, what neighborhood fits best, affordability, schools, crime rates, potential business property for rent, location and communities.

When using Big Data and analytics, it is possible to review patterns to understand whether the property is a good investment and a great match to potential customer. It is also possible to analyze what buyers are selecting more often and based on that data create a model.

When selecting a specific house for sale, Big Data integration within online websites made it possible to analyze local surroundings, sale patterns and neighborhood personality of each area. It created a knowledge comfort by having facts of the neighborhood, its people; and therefore reducing the risk of buying or investing in the wrong property.

2 BIG DATA IN REAL ESTATE BUSINESS

Risk mitigation is essential part of the way Big Data is transforming real estate. Open data across the internet and variety of Big Data tools added strong force for analysis in decision making of choosing right property or home. It equipped customers with the valuable information by extracting the data and cross analyzing it.

Big real estate agencies such as Realtor [13], Zillow [22] and Trulia [16], are pioneering those tools and provide estimated forecast of the property value from 1 to 10 years. Additionally, they provide information about the neighborhood trends, estimate mortgage payment, cost of ownership, history of the property and current value. The calculation is based on variety of public data records, market information, user data points [21] by using Big Data analysis formula developed in-house.

2.1 Real Estate Industry Evolution

Automated valuation methods have been used for a very long time. For decades banks utilized "Automated Valuation Model" to estimate home values. At one point banks wanted to exclusively rely on this model more than home values provided by professional appraisers. That practice led to problems with by omitting important nuances about condition resulting in overvaluation and undervaluation of properties. Big Data analysis and property estimates generated by online real estate giants are the next step in the evolution of real estate industry. This evolution diminishes importance and need for a real-estate agents as it is able to gather a lot of tribal information known only to experts in the area. That means this change can impact job market for over six hundred thousand active agents in the US.

2.2 Real Estate and Artificial Intelligence

Real estate businesses worry that unlocking the vast amount of data about properties could transform the business to be powered by artificial intelligence.

However, based on the GeekWire article [7] big data and artificial intelligence will not replace real estate agents. Robots are just big help and enrichment to the business. It created much better and safer decision making models. Artificial intelligence will help to deliver information about real estate transactions and trends to consumers. It says that in future Amazon voice or Siri could provide

useful information about popular housing trends and market value. Additionally, it can reveal the data on how many people were interested in the property and bids.

So far it is not a robot that is thinking and proactively making decision, it is just a voice based system that extracts the information from Big Data.

For the last twenty years, industry worries about loosing jobs in that area. However, the industry stayed the same. People still want an advice before making an important decision. Even thought, there so much more information and streamlined sales, individuals that want relationships, empathy and connecting with people are still there.

Obviously, there is some fear in real estate that robots can rock the world for real estate business. However, Big Data empowers agents with information and data, it is making them better providers with higher service.

2.3 Online Real Estate Agencies

Online real estate agencies calculate market value by using proprietary formulas. They are not providing expert estimation but a starting point in estimated property monetary worth. It is calculated from public data and surveys, by utilizing special features, market conditions and location. Additionally, they encourage consumers and homeowners to expand online data by doing other investigations such as comparing market prices for around areas, working with a real estate agent, getting an appraisal from an expert and visiting the house [21].

For example Zillow, developed a Zestimate prediction[21], which is Zillow's estimate of a home that currently on sale, one to ten years from now. The provided information based on current house and market condition. Other real estate agencies with online presence competing with Zillow, like Trulia and Realtor for example have developed similar proprietary formulas to assist customers.

Also, the companies provide rent estimates that would help evaluate potential monthly rental price by developed in-house algorithmic formula. Variations in rental prices can also happen because of different factors, additional investments, or length of lease.

Big Data information affects the forecasting. As an example, the amount of rental listings in a specific area affects how much we know about approximate prices in that area for condos, apartments and houses. Based on number of properties for rent, the prediction becomes more accurate. Homeowners can also update and provide information online about their needs or property, which helps even more for predicted accuracy.

The formula they use to estimate rent prices is comparing similar homes and apartments in the given area. Comparing bedrooms, square footage and other details. Then prices are being compared, and pattern to rental prices is shown.

Big data analysis provides unbiased information. Although, majority of real estate agents are esteemed professionals looking out for clientfis best interests they are still. There are still those that would like to manipulate clientfis opinion to benefit themselves. For example, if a particular house have been on the market too long and the agent might lose the listing there is a possibility that some shortcoming of the property will be omitted by the agent in order to complete the sale. Same can be said about agents trying to achieve some sales goals or quotas. However, if the potential buyer conducts the research using Big Data all information will be available. Put simply, data does not lie.

3 REAL ESTATE ANALYSIS

Big Data is widely used by agents and real estate agencies to understand and improve how to target potential buyers. But the great thing about Big Data is that customers benefit from it as well. They can use free public resources with tons of data and information maps with different data analyzing tool options.

Latest tools allow to utilize Python to cross mix and match different values and data sets to analyze complex data. Prior to having these tools available such analysis would be an impossible task for individual users and required immense human and computing effort to complete. It is possible to visualize it by rendering correlations and trends. It reveals stunning insights in to chosen property for rent, business or home.

There is so much information that it is important to understand which data is relevant to consumers and improves decision making. It is useful to analyze the data-sets when considering investing [3]. The analysis can provide variety information and make the educated decision on the investment.

3.1 Big Data Tools

Analysis of these featured data points could be done with Python tool sets and libraries.

Python is a great programming language with variety of options. It is object oriented, semantically structured and great for scripting programs as well as connecting other programmable components. Python is considerably easy to learn and because of its high productivity and also became one of the favorite tools for programmers and data scientists. It contains libraries that are script importable and usable for a lot of use cases, such as image modification, scientific data analysis and server automation. Python world has been around for thirty years and a lot of code was written with multiple contributors. Variety of options built up on how to visualize the data [19].

The most common type of visualization is a simple bar chart and line graph [14]. It is popular and commonly used type of visualization to make comparison between values and variety of categories. It can be vertically or horizontally oriented by adjusting x and y axes, depending on what kind of information or categories the chart requires to present. Parameters need to be identified, such as axes, similarities, title and decided on what exactly the visualization supposed to show.

To make a simple bar chart, a number some of the most popular tools and libraries that have been invented for plotting the data could be utilized. These include the most used and common tools such as: Pandas, Seaborn, Bokeh, Pygal and Ploty.

Additionally, just like any other programming language issues, errors or questions with the libraries can be found on stack overflow page by Google search.

3.2 Data Analysis

For the purpose of this project, bar chart and graphs visualization methods with pandas modules in Python have been rendered and

explained. The simple form of this plot looks acceptable and easy to read.

The techniques were done within Jupiter notebook [9].

Jupiter notebook is great for running data sets analysis and for calculation projects. Jupiter notebook documents are readable files having the analysis description and the results in figures and tables as well as exportable files which can be executed to perform data analysis. It allows to render images and move values back and forth between different modules and coding languages.

The data-sets collected from clrsearch.com [5], data.gov [17], zillow.com [22] and uploaded to the class's Google Drive to demonstrate the trends and patterns between each output.

The data includes both geographic and social data-sets evaluated by ratings in rows and titles in columns to keep it simple. The data set for both cities is being used for all examples that are demonstrated below. The point of the visualization is to understand the data in visual platform and make an informative decision based on rendered data.

A simple example of two properties in Tarzana, California versus Calabasas, were compared and exported for read.

Tarzana City is a wealthy neighborhood in the San Fernando Valley region of the city of Los Angeles, California. Tarzana was purchased in 1919 and developed on the site of local elites and named by Edgar Rice Burroughs, author of the popular Tarzan books. He established Tarzana and later sold it to local farmers [20].

Calabasas City located in the hills west of Malibu, in the San Fernando Valley region of the city of Los Angeles, California. The area established in 1991 and the name was derived from Spanish word "calabaza", meaning pumpkin.The legend has it that in 1824, a Mexican rancher spilled a wagon of pumpkin seeds and it spouted alongside the road. Therefore, the area was named Calabasas, the pumpkin land [20].

From a quick glance both areas are very similar and are located within 10 miles of each other. Both Tarzana and Calabasas are influential and desirable neighborhoods with lots of high priced homes. How does one differentiate between the two in order to find the right investment?

Big Data is the answer. Specifically in states like California. In California Big Data application benefits greatly from availability of public records such as sale price as apposed to certain "non-disclosure" states. There sale prices for homes are not disclosed in public records.

The analysis combines several main components, including property characteristics in the area, crime rate, quality of life, pollution, race and ethnicity, population growth, family household, house value, business field, employment, schools and future home value.

3.3 Property Characteristics

Big Data analytics can help in connecting needs of a buyer and providing neighborhood demographics. The quality of population in the neighborhood will influence who buys the house and who lives there. It is important to identify what is important to you and make sure those items are covered in the research. For example, if you are a student you will probably look for a densely populated location around universities, closer to food locations and communities. Things like public transportation,nightlife, and bars will be very important to you and will be prioritized over other things. If you are married with kids, your best choice would be location with good schools and low crime. Parks, playgrounds and traffic and noise pollution around the house will be paramount. Most parents would love to find a nice quite cul-de-sac house. Young working professional would prefer to be right in the middle of things on a busy boulevard.

Latest Big Data collections made it all possible for real estate website to provide that information to potential buyers. Websites such as United States Zip-codes [2] collect information from public records and make it available in exportable format as well as for reviewing and analyzing local neighborhoods by states and zip codes input.

3.4 Crime Rate Indexes

Crime Big Data is available now and helps to see patterns and avoid areas with unfavorable statistics. The Los Angeles Police Department [1] already uses the data to show which areas in Los Angeles are hot-spots of crime.

Crime rates are being calculated by comparing the national levels of the average 100 [6]. For example, if score is 150 it means that it is 1.5 higher risk of crime that national average level. The data is coming from police department reports and public records. Additionally, the Federal Bureau of Investigation also provides factual information for ranking [18]. Furthermore, the research on crime can be extracted from United States Department of Justice via Uniform Crime Reporting Program [11].

In this example [9], running the crime data sets of Tarzana and Calabasas showed that Calabasas is in much better shape and safer place to live as compared to Tarzana, which is around the average of national rate. The total crime risk in Tarzana slightly higher than national average, it is 108, meanwhile Calabasas is almost 3 times lower, it is 24. The murder risk is 118 compared to 24 in Calabasas. Rape risk is 70 in Tarzana and 35 in Calabasas. Robbery risk in Tarzana is almost twice higher than in Calabasas, it is 125 versus 77. Assault risk three times higher in Tarzana, it is 127 versus 48 in Calabasas. Burglary risk twice higher in Tarzana as well, it is 55 versus 27. Larceny risk in Tarzana is overwhelmingly high, it is 73 versus 9 in Calabasas. Motor vehicle theft risk in Tarzana is 118 versus 39 in Calabasas. Based on these findings, it is defiantly safer to live in Calabasas [fig 1].

Based on these finding, it is defiantly safer to live in Calabasas as shown in Figure 1 [9].

3.5 Education Levels

Next run was done on educational level of residents. Big Data includes data of resident's education level and makes it possible to collect data about an individual resident and provides insightful information about social level interaction. The data extracted and combined from variety of sources including international school districts.

The education rating filtered by zip codes represents the percentage of people in the area who have attended colleges and received degrees. It does not represent performance and specific schools.

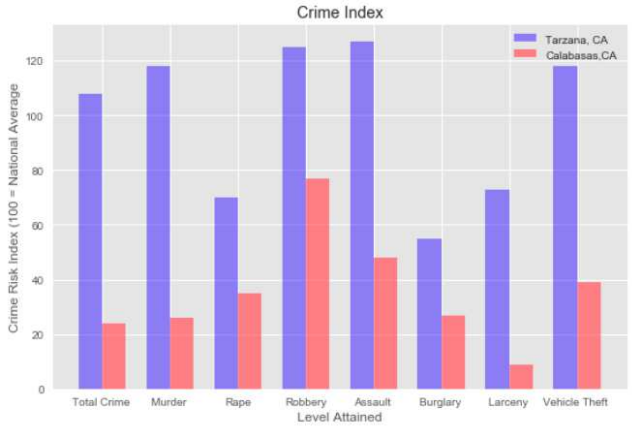


Figure 1: Crime rate in Tarzana, CA compared to Calabasas, CA (100 = National Average) [9].

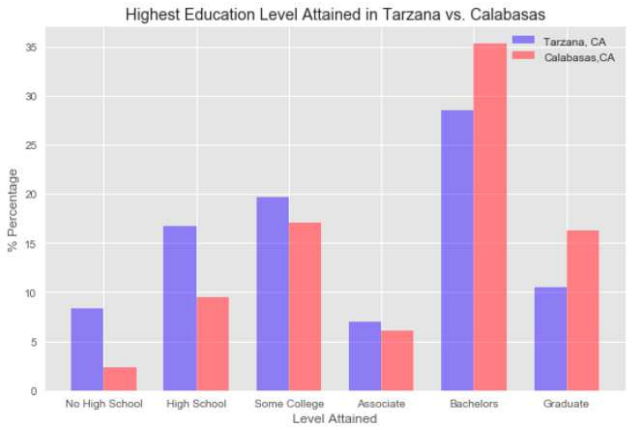


Figure 2: Educational percentage of people in Tarzana, CA compared to Calabasas, CA (Population Age 25+) [9].

The rendered data showed [9] that residents in Calabasas are higher educated by 7 percent with Bachelor degrees and 6 percent higher with graduate degree, as shown in Figure 2 [9].

Based on the Economic Policy Institute study [4], there is a clear correlation between higher educated workforce and economic success within state and ability to grow. Additionally, higher educated people are good for state budgets, since workers with higher income contribute more through taxes.

3.6 Life Quality Standards

The next important consideration in buying a property, searching for a house and making a decision is quality of life standards in that area. Big Data and and latest methods of data collection can lead to improvements in quality of life for residential areas. It can find neighborhoods that are safer, cleaner, more entertaining and a better place to live specifically tailored to potential buyer.

The data-set of life quality obtained from variety of sources, including public Google searches, social media and local study

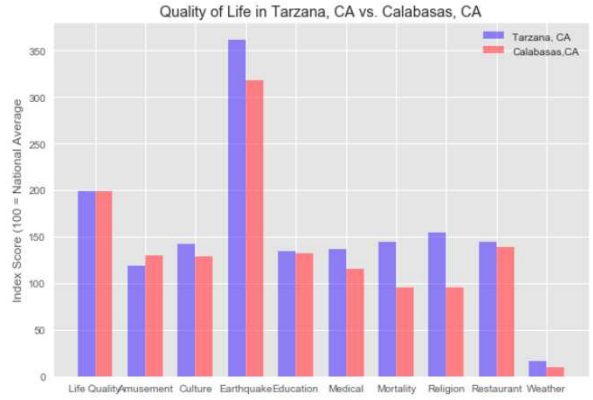


Figure 3: Life quality of people in Tarzana, CA compared to Calabasas, CA [9].

groups. The quality of life is being measured by how residents are being effected by crimes, weather, education, entertainment, religion, medical support and food supply. The positive decision variables calculated by amusement, education, culture, media, religion, weather and restaurants. The negative decision is based on the level of crime, natural disasters and mortality. The national level is being compared to 100 [5].

Rendered data showed that amusement index equal to 110 in Tarzana and 130 in Calabasas. What that means is that in Calabasas there are more community events and entertainment. Culture is 142 in Tarzana versus 129 in Calabasas. Culture refers to artistic development. Earthquake index 362 in Tarzna compared to 318 in Calabasas. this is a very intresting point considering that both neighborhoods are very close to each other. But since Tarzana's Earthquake index is higher associated insurance will likely be higher as well. Raising cost of ownership. Medical index is 137 in Tarzana and 116 in Calabasas.If you are working the the medical field this might be an important topic for you as it will help you find employment closer to home. Reduce your commute time, minimizing wasted time spent in California's infamous gridlock traffic. Mortality is much higher in Tarzana, it is 144 versus 95 in Calabasas. Religion is better in Tarzana, it is 154 compared to 96 in Calabasas. Religion refers to houses of worship and religious establishments. Restaurant index about the same, it is 144 in Tarazana and 139 in Calabasas. Weather is better in Tarzana, it is 16 versus 10 in Calabasas. That is another interesting observation considering that both neighborhoods are minutes away from each other.

Based on the data, overall quality of life is equal between two cities, as shown in Figure 3 [9].

3.7 Air Pollution

Big data can control and reveal pollution levels of particular area [8]. It is one of the main causes of health problems in the population and preventive cause death.

Over 80 percent of residents living in urban areas are vulnerable to poisoning from pollution. Cancer is one of the leading cause of deaths for both men and women; and exposure to pollution at early may have life-long negative consequences.

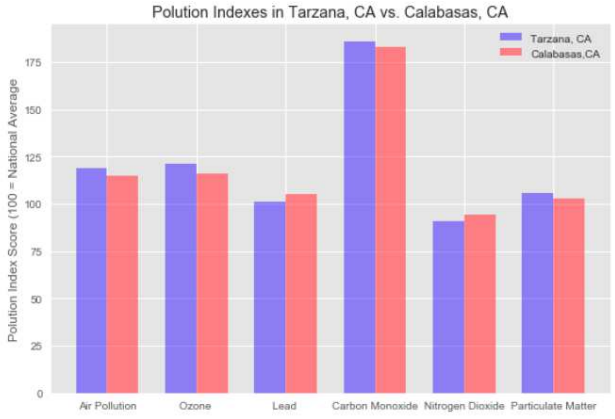


Figure 4: Air Pollution Indexes in Tarzana, CA compared to Calabasas, CA [9].

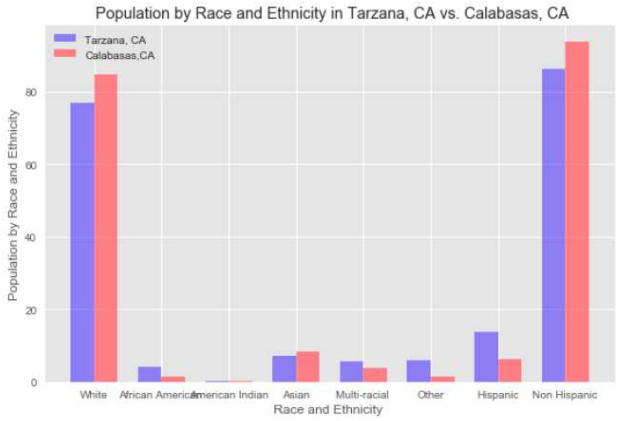


Figure 5: 2012 Population by Race and Ethnicity in Tarzana, CA compared to Calabasas, CA [9].

Monitored areas show that air quality levels exceed the safety levels [12]. Additionally, the World Health Organization warns that most populated states are most affected.

Government is aware of this problem, therefore collecting and monitoring the data regarding air quality has increased. The data is being shared between universities and air quality maps for further development. The data is openly shared and prepared for Big Data analysis.

Even thought Big Data will not reduce the pollution by itself, it provides tools to visualize the problem which is especially helpful when choosing a place to live.

The exported data-sets showed [9] that carbon monoxide is extremely high in both cities. It is 186 in Tarzana and 183 in Calabasas. The national level is being compared to 100 [5]. Based on the data, overall air pollution index is about the same in both areas, as shown in Figure 4 [9].

3.8 Race and Ethnicity

Big Data can reveal a lot of information about population by using zip codes. It shows profiles of people who live there. Understanding ethnicity and identity of the community influence will help with decision.

The standard of maintaining, collecting and presenting federal data on race and ethnicity [10] were revised and improved on collecting quality about two decades ago. In accordance to best analysis practices, federal agencies conducting researches to better understand ethnic and race diversity.

The language to describe the ethnicity and race keeps changing to resonate with the category of residence and adding new meaning to not make it discriminatory. The general rule became that race and ethnicity should not be interpreted as being a science.

Based on the rendered graph, most population in Tarzana and Calabasas consist of white and non-Hispanic residents, as shown in Figure 5 [9].

That information provides insight about communities and relatedness to the buyer.

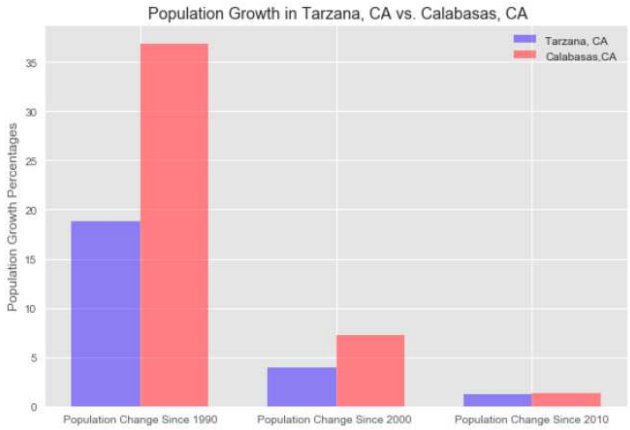


Figure 6: Population change since 1990 in Tarzana, CA compared to Calabasas, CA [9].

3.9 Population Growth

Leveraging Big Data in population growth might be helpful for economic growth prediction and future development.

For the recent centuries, population growth jumped dramatically [15]. How fast the population is growing can influence area homes and businesses development. Allowing for more business opportunities.

Educated people can contribute to the development with increased skills and knowledge. However, it is also important to look not only on the total population size, but also population growth rate.

Based on the data visualization, population size in Tarzana is higher by 3,000 residents than in Calabasas, as shown in Figure 6 [9].

Both in Calabasas and Tarzana, the rate was rapidly increasing from 1900-2000, and there was not much progress since then. Population density in Tarzana is 4,048 versus 856 in Tarzana. City area

size in square miles is 7.44 and 31.67 in Calabasas. This information provides insight that Calabasas has much more opportunities for future growth and development. New housing and real estate development is achilies heel in California. State struggles to provide all existing residents with affordable housing. Compiled with population growth and migration of new residents the problem becomes even harder resolve. By having additional development space Calabasas growth potential is much higher compared to Tarzana.

3.10 Family Household

Big Data and internet of things are making its existence common place in each household. Only 15 years ago home computers were the only smart device in the house. Now even vacuums and thermostats are connected. Our homes are goldmines of data. Getting family household data summary instantly tells about the type of people in these areas and obtained knowledge can be used to help with buying decision.

Household definition refers to type of family and people living in a household structure. Household data is useful when consumer wants to know about the type of people living in that area and relativity.

Based on the combined data-set results [9], full family household is 64 percent in Tarzana and 76 percent in Calabasas. 48 percent are married in Tarzana, and 62 percent in Calabasas. Therefore for married families with kids it makes more sense to live in Calabasas.

3.11 Property Value

Big Data is being used to analyze property values. Real estate agencies, such as Zillow [22], estimate values based on Big Data collection tools and using their algorithm [21].They combine information from variety of sources and provide insightful information to buyers, sellers or brokers.

Based on the data analysis [9], it shows that Calabasas prices are higher than in Tarzana by 23 percent. That insight shows that more financially able residence live in Calabasas.

To confirm that, the income data was calculated. Based on the rendered data as shown in Figure 7 [9], it proves that residence in Calabasas are more influential with higher income than in Tarzana.

The total income in Calabasas is higher by approximately 20 percent.

3.12 Employment and Occupation

The employment breakdown that derived from data, published by the Bureau of Labor Statistics showed that business field compared with employment field could help with predicting job opportunities.

Based on compared data sets, Health-care is leading employment field in Tarzana and Managment in Calabasas, as shown in Figure 8 and Figure 9 [9].

3.13 Public Schools

Big Data in public schools are being used to fix education institutions and improve student scores and results. whereas in the past school performance was judged simply on average API scores of the students now student attributes data is further analyzed. This allows to identify subgroups of under-performing students. For example income levels of households are tracked to make sure that students from low-income families have the same opportunities to have better scores and grades as families from high-income families. It also provides tracking and comparison with schools in different districts. This helps school boards to allocate additional resources to schools that lack them. It also helps parents and home buyers identify schools and neighborhoods where their child could flourish academically.

The mined data could be used for decision making in property investment as well. Prospective buyers with kids are not only looking for good education and safe schools for their own kis, but also from stand-point of property value since homes located in good school districs are more desirable. The detailed information that can be found online made it easy to be properly informed. Compared data between two cities, showed that elementary schools have 38 percent higher rating in Calabasas, middle schools are 25 percent higher and high schools are the same. Schools in Calabasas are better based on these rating scores, as shown in Figure 10 [9].

3.14 Available Houses for Rent and Sale

Another shift in demographic preferences that has been observed is related to home ownership vs renting. Millenniums are changing their spending habits when compared to previous generations. Food, health and entertainment take priorities over burdens expenses associated with home ownership. If that trend continues return on investment generated by buying rental properties will rise.

The best way to know if a house is a good investment is to check the rental properties near the area.

There is also a 1 percent rule of thumb to keep in mind. The rule is that a purchased home should be rented for 1 percent of the cost.

Based on the rental data, medium price in Tarzana 4,210 dollars per month, and Calabasas 4,085 dollars per month. It actually reveals that Tarzana rental properties are more expensive than Calabasas, even thought the home prices in Calabasas are higher, as shown in Figure 11 [9].

Additionally, square footage was calculated. To get the price per square footage, the price of the area was divided by its square footage. The results showed that in Tarzana rent is slightly higher than in Calabasas, as shown in Figure 12 [9].

Therefore, it makes more sense to buy renting properties in Tarzana.

The lowest price of property in Tarzana is 700,000 us dollars, and in Calabasas it is 975,000 us dollars [9].

Based on the 1 percent rule, it does not make sense to buy and rent out in Tarzana or Calabasas.

3.15 Future Value

California housing is booming and crashing. Massive home equity destruction happened few years ago and reversed back.

When data-sets are analyzed, they can reveal insightful information and guide consumer decision making.

Based on the sales data was taken and generated, suggests that in spite of prices drops the value of houses goes up, as shown in Figure 13 and Figure 14 [9].

Calculated housing investment for the last 20 years had a growth rate of 5.46 percent [9]. By knowing a starting and ending value, it is

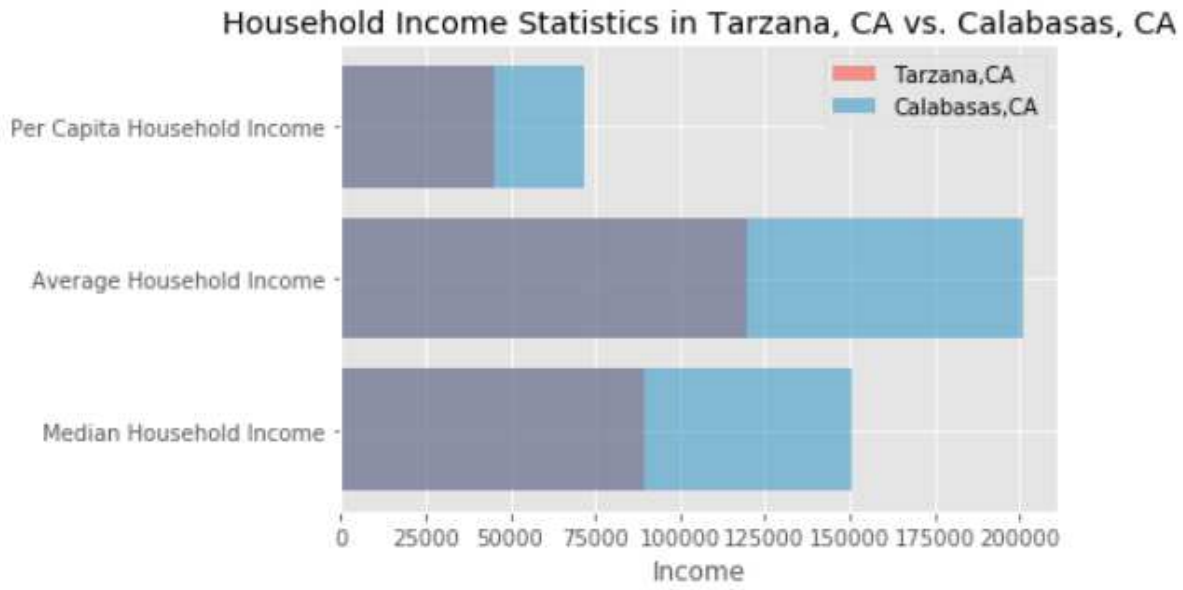


Figure 7: Income in Tarzana, CA compared to Calabasas, CA [9].

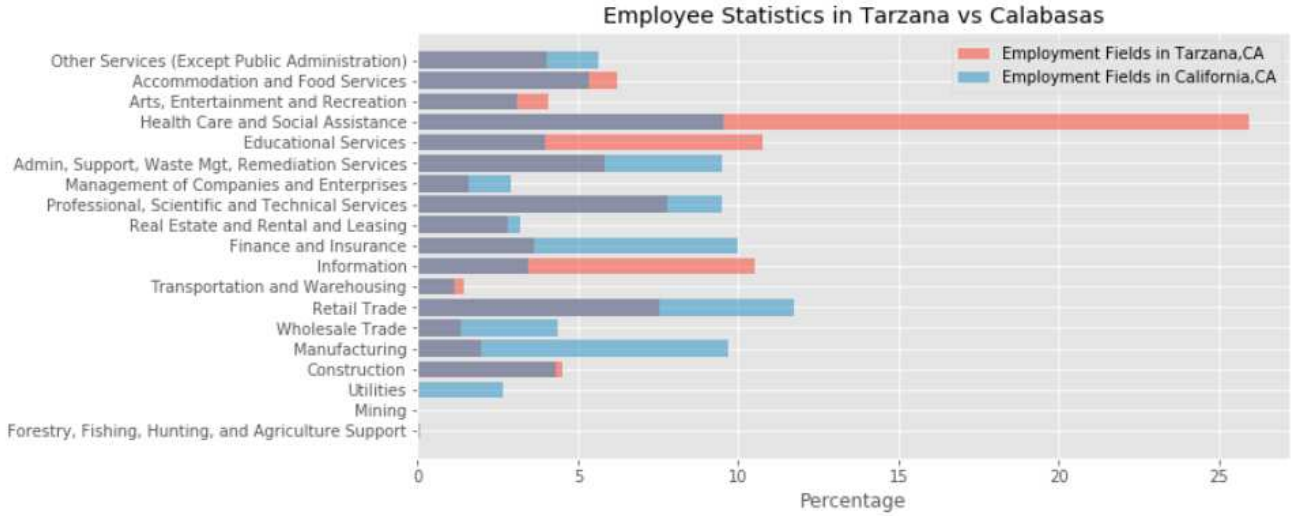


Figure 8: Employment field in Tarzana, CA compared to Calabasas, CA [9].

possible to calculate the future value of an investment. Referencing the previous calculations [9], it predicts that house value will grow by 63 percent in the next 20 years.

4 CONCLUSION

Big Data potential to transform decision making in real-estate is immense. Home ownership is part of the American dream and Big Data will play a huge role in that process. It will allow potential buyers to have a better understanding of historic data and how it correlates to investment potential.

Big data will provide powerful insight to augment decision making process. Yet, it will not eliminate all risks associated with investment in real-estate. All risks must be evaluated and analyzed before buying and big data will provide plenty of tools for that.

Based on this analysis, it was determined that Tarzana and Calabasas properties are overpriced. Currently, renting is low compared to buying a property.

It is impossible to find properties in California that generate rents at around 1 percent of total property cost. You can not justify

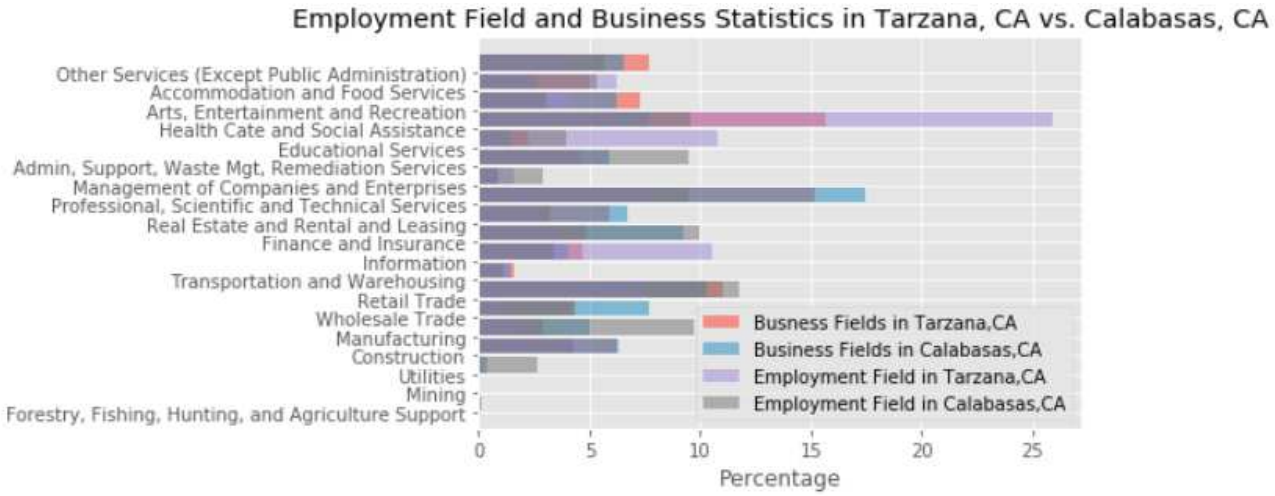


Figure 9: Business fields in Tarzana, CA compared to Calabasas, CA [9].

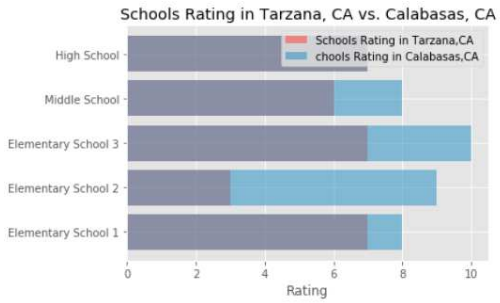


Figure 10: Public schools in Tarzana, CA compared to Calabasas, CA [9].

Figure 11: Houses for rent in Tarzana, CA compared to Calabasas, CA (3bd+ House For Rent (1,500-2,500 Sqft)) [9].

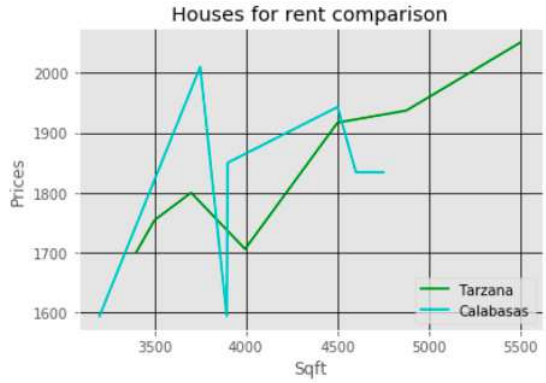


Figure 12: Price per sqft for rent in Tarzana, CA compared to Calabasas, CA [9].

Figure 13: Prices Growth Index in California [9].

the prices and it is only for the privilege of living in San Fernando Valley region of the city of Los Angeles, California.

However, if you do still want to invest, Calabasas is a better choice for investing in a family home property and Tarzana for a rental property.

8

Figure 14: Prices Growth Index in California [9].

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski, Juliette Zerick and Miao Jiang for their help, support and suggestions to write this paper.

REFERENCES

[1] 0 2017. *Federal Register The Daily Journal of the United States Government.* 0. https://www.federalregister.gov
[2] 0 2017. *United States Zip Codes.org.* 0. https://www.unitedstateszipcodes.org/91356/
[3] Andrew Beattie . 2017. *Top 10 Features of a Profitable Rental Property.* 0. https://www.investopedia.com/articles/mortgages-real-estate/08/buy-rental-property.asp
[4] Noah Berger. 2013. *A Well-Educated Workforce Is Key to State Prosperity.* 0. http://www.epi.org/publication/states-education-productivity-growth-foundations/
[5] CLRsearch.com. 2012. *Tarzana, CA 91356 Population Growth and Population Statistics.* 0. https://www.clrsearch.com/Tarzana-Demographics/CA/91356/Population-Growth-and-Population-Statistics
[6] CLRsearch.org. 2012. *Community Demographic Information FAQ.* 0. https://www.clrsearch.com/demographics/Demographic_Information.jsp
[7] John Cook. 2017. *Robots in real estate?* 0. https://www.geekwire.com/2017/robots-real-estate-theres-nothing-see-zillow-co-founder-says-agent-jobs-safe/
[8] Arantxa Herranz. 2017. *Big data will control pollution in your city.* 0. http://blog.ferrovial.com/en/2017/04/big-data-pollution-control-in-cities/
[9] Elena Kirzhner. 2017. *Big Data Applications in Real Estate.* 0. https://github.com/bigdata-i523/hid320/blob/master/project/project.md
[10] Management and Budget Office. 2016. *Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity.* 0. https://www.federalregister.gov/documents/2016/09/30/2016-23672/standards-for-maintaining-collecting-and-presenting-federal-data-on-race-and-ethnicity
[11] FBIfis Crime Statistics Management. 2017. *Uniform Crime Reporting Statistics: Their Proper Use.* 0. https://ucr.fbi.gov/ucr-statistics-their-proper-use
[12] World Health Organization. 2016. *Air pollution levels rising in many of the worldfis poorest cities.* 0. http://www.who.int/mediacentre/news/releases/2016/air-pollution-rising/en/
[13] Realtor.com. 2017. *Realtor.com - resource for home buyers and sellers.* 0. https://www.realtor.com
[14] Naomi B Robbins. 2012. *Creating more effective graphs.* Wiley, 0.
[15] Max Roser and Esteban Ortiz-Ospina. 2017. *World Population Growth.* 0. https://ourworldindata.org/world-population-growth/
[16] Trulia.com. 2017. *Trulia is a mobile and online real estate resource.* 0. https://www.trulia.com
[17] U.S. General Services Administration, Technology Transformation Service. 2017. *Real Estate Sale History.* 0. https://www.data.gov
[18] Mark van Rijmenam. 2017. *The Los Angeles Police Department Is Predicting and Fighting Crime With Big Data.* 0. https://datafloq.com/read/los-angeles-police-department-predicts-fights-crim/279
[19] Guido Van Rossum and Fred L Drake. 2011. *The python language reference manual.* Network Theory Ltd., 0.
[20] Wikipedia. 2017. *Tarzana, Los Angeles.* 0. https://en.wikipedia.org/wiki/Tarzana,_Los_Angeles
[21] Zillow.com. 2017. *The Zestimate home valuation.* 0. https://www.zillow.com/zestimate/#what
[22] Zillow.com. 2017. *Zillow is the leading real estate and rental marketplace.* 0. https://www.zillow.com

Big Data Analytics in Identifying Factors Affecting Bitcoin

Ashok Kuppuraj
Indiana University
Bloomington, Indiana 43017-6221
akuppura@iu.edu

ABSTRACT

Pricing of Blockchain based cryptocurrencies are like a black box, as per theory the pricing compared to U.S dollar is based on a number of transactions however lot other factors like Dollar price, social media, Online threats supersede the transaction count. Big data and Analytics helps to identify the metrics impacting this variation and identify the correlation between them.

KEYWORDS

i523, hid324, Big data, Predictive analytics, Random Forest, correlation, Blockchain, Bitcoin, Ethereum

1 INTRODUCTION

The start of the 21st century witnessed the evolution of various disruptive technologies, right from Big data, IoT, VR to Blockchain. When it comes to the blockchain, the sole winner is Bitcoin, with the growth rate of over 1327 percent [5], Bitcoin is disrupting the way banking system works. As the Bitcoin grows the acceptance and adoption grow along with that. Similar to any other currency in the world, Bitcoin's price deviates widely towards the positive side which created the opportunity for investment in it. Even though the same is not widely accepted everywhere, there is a grace to own Bitcoin citing its growth rate. Though the transaction counts haven't grown up, the retention of the coin has grown up making it a Digital Gold [17].

2 BITCOIN

Bitcoin is a progressed cryptographic cash and shared ledger that is completely decentralized, which implies it relies upon peer-to-peer trades with no bureaucratic oversight. Trades and liquidity inside the framework are somewhat based on cryptography. The concept was first introduced in 2009 [8] and is at this moment a prospering open-source gathering and portion sort out. In perspective of the uniqueness of Bitcoin's tradition and its creating choice, the Bitcoin is grabbing stacks of thought from associations, clients, and monetary experts alike. Specifically, for this technology to thrive, we need to recreate budgetary organizations and things that starting at now exist in our traditional, fiat cash world, make them available and specially fitted to Bitcoin, and other rising computerized types of cash. In technical terms, Bitcoin's is a shared ledger or a database running by a set of clusters, as the clustering is involved, a competition is set for the individual machines to acquire and update the ledger. The competition is in terms of hashing problem. The hashing needs multiple GPU's to perform validations and update the ledger. This competition eliminates the slower machines to be part of the network and improve the infrastructure's capacity, only by winning the competition a machine can be awarded some Bitcoin as an incentive. Since one machine cannot process the competition problem, a set of peers come together to form a Mining pool and share their capacity and the incentives. We can gather useful mining statistics information from these mining pools.

3 PRICE PREDICTION

The Bitcoin market's cash-related basic is, clearly, a securities trade. To support money related to reward, the stock market prediction has turned out to be known ground which can be reused with the presence of high-repeat,low-dormancy trading hardware joined with solid machine learning figurings. Henceforth, it looks good that this desire is imitated in the domain of Bitcoin, as the framework expands more conspicuous liquidity and more people develop an excitement for placing profitably in the structure. To do accordingly, it is essential to utilize machine learning and Big data advancements to foresee the cost of Bitcoin [10].

3.1 Data Source

As Bitcoin is a decentralized and a transparent system, all the source of data can be gathered from the peer-to-peer networks. This peer-to-peer network is called as Bitcoin-mining pool [2].The rate of block creation is adjusted every 2016 blocks to aim for a constant two week adjustment period (equivalent to 6 per hour.) The number of Bitcoins generated per block is set to decrease geometrically, with a 50 reduction every 210,000 blocks, or approximately four years. The result is that the number of bitcoins in existence is not expected to exceed 21 million [6]. The true source of data for Bitcoin analysis would be from Bitcoin mining pool. Coinbase is one of the main members of bitcoin pool from which we can gather mining statistics. In the process of identifying the features impacting Bitcoin's price fluctuations, not only the transaction volume impacts, even the popularity and people's trend towards it impact the price of the coins. Hence, data from Google is also gathered. As a currency's price also been altered by its exchange, supply, and demand, Ethereum's price data and transactional data is also acquired from Ethereum's exchange point. With all these data sources, we analyze the features impacting the Bitcoin's market price.

3.2 Feature Selection

Feature selection is one of the vital steps in any meaningful analysis of an expected outcome. A set of features have been selected to analyze its interdependence with Bitcoin's evaluation. The features are selected based on three wide areas, the first is Bitcoin mining data, second is social data and the last one is exchange data. The internal activities in the Bitcoin's infrastructure definitely reflect the changes or the fluctuations in the Bitcoin's network, Bitcoin's mining data is gathered from Coinbase. This is extracted from the

web service API provided by Quandl.com [9]. By making a REST call, CSV files containing the historical data is downloaded and processed. The second is the social data, which is extracted as a static data from Google trends [7], the main reason behind this data is when the popularity grows people tend to know or show interest in being part of the growth. With the impressive growth of more than 1000 percent in a year, this is considered as an important data. The last one is the exchange data, as a currencies price is directly proportional to the supply and demand, the supply of the currency can be impacted by the exchange to other currencies or commodity [18]. Ethereum is known to show a similar pattern in terms of growth and deviations [3]. Hence, there's price in US dollars and transaction volume is considered one of the features.

4 BIG DATA IN FEATURE ANALYSIS AND ALGORITHM'S EXECUTION

Feature extraction, transformation, and prediction can be synonymous with a conventional ETL methodology. Though few of the extraction is handled manually and the volume is comparably low, it is assumed that the data volume will be increased by modifying the extraction to real-time systems. When the extraction systems are changed, our code must be able to handle streaming data which can be related to "variety and volume " of the data. The next step is validating the data for anomalies, data miss and cleanse the data of issues which is synonymous with data cleansing. The later one is data processing, which includes data processing with multiple iterations and permutation consuming a lot of memory and other resources. These processing needs lead us in adopting Big data technologies in the entire lifecycle of the implementation. Apache Spark framework is identified as the end-to-end processing environment which is pre-loaded with redundancy, fault tolerance, in-memory processing, parallel processing, streaming, and Machine learning modules.

4.1 Execution with Apache Spark

"Apache Spark is a fast and general-purpose cluster computing system. It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs". It also provides extensive support to Machine learning libraries(MLlib) and to streaming through Spark Streaming. The in-memory processing is implemented with the help of Resilient distributed dataset (RDD) [12].

Spark's architectures given in the figure 1 provide a glimpse of how different system in Spark is interfaced. The first level of interfacing to Spark is with high-level languages like Scala, Java, Python and R. Users implement their functionalities in these high -level languages. The primary executing components in Spark are Driver and Executor modules. The driver is the entry point for any implementation, the written programs will be executed in the main function of the Driver module, later converted to set of Directed Acyclic graph by the Spark APIs. DAGs are then executed in executors in the data nodes based on the data placement policy of the infrastructure. Four modules built on spark for serving the user's needs are SparkSQL, Spark Streaming, MLlib, and GraphX. Spark SQL and Machine Learning libraries(MLlib) are consumed in our implementation and the future improvement would be on Spark