Rohan Mahajan
SU ID - 270982114

Applied Data Science Portfolio Essay

Portfolio Link:

https://github.com/rohanmahajan0396/Applied-Data-Science-Portfolio

Introduction:

Data Science is the ongoing trending field in the tech industry. Not only can it be done from scratch regardless of any user’s background, but it’s usage has been incredible. The impact of data science has been incredible and the top companies of the tech industry are demanding aspiring data scientists for their firms.

Companies like Google, Facebook and Amazon generate billions of data everyday based on user usage and the amount of manipulation done using this data can lead to incredible insights, trends and predictions.

After giving a presentation on data storage in Hadoop, I went deeper into how this data could be used and what could one possibly do with so much of data. I then discovered the field of data science and started learning it from scratch. Being a Computer Science major during my undergraduate years, I was familiar with how programming languages and software use and work with datasets. That is when I decided to go out of my comfort zone and move to the United States to pursue a Masters in Applied Data Science at the School of Information Studies.

As an analytical mind, my eagerness to learn and make decisions based on data is where my strengths lie. Transforming and utilizing data into decisions is something I always aspired to do.

Over the course of almost 2 years in the iSchool, I have learned to process, manipulate, optimize and interpret data in an efficient way.

I have outlined my journey in the iSchool and my progression towards diving deep into the data science filed in my portfolio along with a project I have done every semester.

Summary and Learning outcomes:

The Applied Data Science program focuses on applying data science concepts to solve real-life scientific and business problems. It is an interdisciplinary course provided by the School of Information Studies and the Whitman School of Management.

The learning outcomes of the program focus on the following:

1. A basic overview on data science with it’s major practice areas.
2. Collecting and organizing data from real-world scenarios followed by identifying trends and patterns in datasets through visualizations, statistical analysis and data mining.

3. Using alternative strategies based on datasets and implementing new business decisions derived from the analysis.
4. Synthesizing the ethical dimensions of data science.

Below I have given a summary on how I went on with achieving this learning outcomes and how my projects aligned with them throughout the semesters.

1. Overview of Data Science Concepts, Statistical Analysis and Data Visualization – Fall 2018:

In my first semester of my program, I got an insight as to how basic regression models work. I worked on projects that helped me understand how variables are correlated with each other and how basic statistical analysis is done. I also learned to interpret datasets using data visualizations and find patterns and trends.

MBC 638 – Restaurant Feedback Analysis | Data Analysis and Decision Making

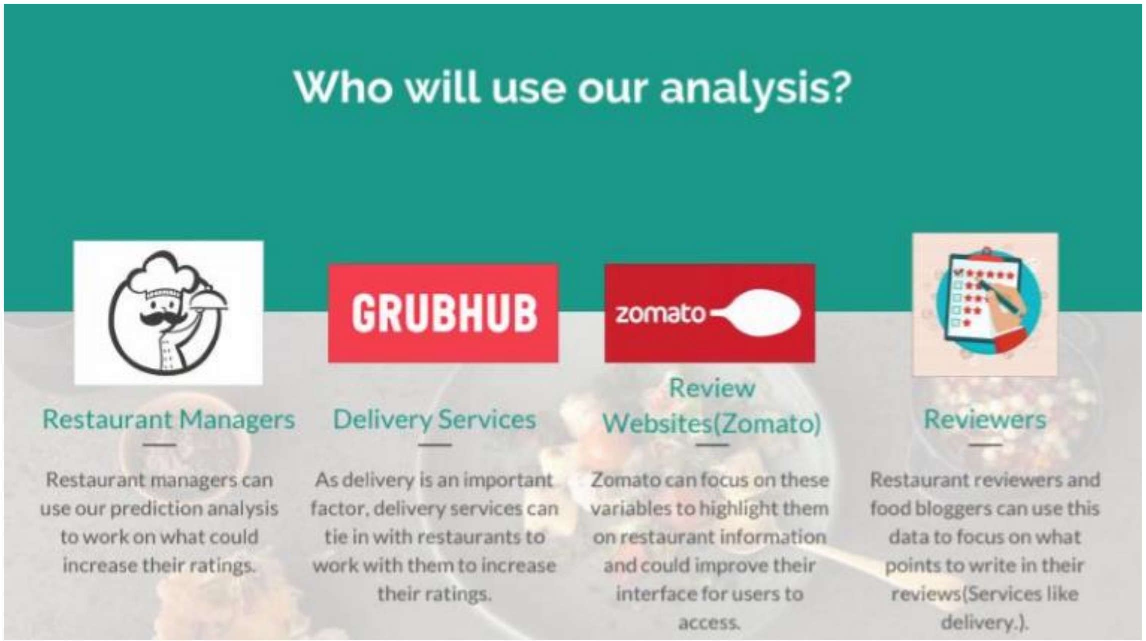
I first dived into data analytics in my Data Analysis and Decision Making course which I took during my first semester in the Fall 2018 semester. This course focused more into using Regression and Statistics in excel to conduct analysis and insights on datasets. This course was taught by Professor Anna Chernobai at the Whitman School of Management and was a core course in my program.

Our main focus was to interpret our variables for the project and to provide insights to a business problem. My group members and I chose a Zomato Restaurant Review dataset where we conducted a single regression, multiple and stepwise regression analysis on the aggregated Restaurant Ratings to determine what factors affect the ratings of the restaurant.
Since our focus was to determine what factors would affect the restaurant rating, interpreting our variables was important in this case. Therefore, we only proceeded with using regression as that is the best proven way to interpret our variables.

We manipulated our variables to have the best possible explained value (R-Squared value) for our variables on our target variable.

Our best determined variables/factors and actionable insights were the following:
1. Adding an online table booking service.
2. Adding an option for delivery.
3. Adding an option for booking deliveries online.

This project helped me and my group members, who at the time were also just starting their Masters program understand how regression and the variables associated to it work.
The Use-Cases for our recommendations are given below:



Who will use our analysis?

Restaurant Managers

Restaurant managers can use our prediction analysis to work on what could increase their ratings.

Delivery Services

As delivery is an important factor, delivery services can tie in with restaurants to work with them to increase their ratings.

Review Websites(Zomato)

Zomato can focus on these variables to highlight them on restaurant information and could improve their interface for users to access.

Reviewers

Restaurant reviewers and food bloggers can use this data to focus on what points to write in their reviews(Services like delivery.).

We worked with a clean dataset and this however just scratched the surface for what was yet to come.

2. Identifying different strategies for datasets – Spring 2019:

While I worked on clean datasets in my first semester, I went on to work with datasets given on Kaggle for my future projects. These datasets were relatively unclean and processing needed to be done before any models or algorithms could be used.

I also moved on from the traditional regression approach and focused more on classification and unsupervised learning. This laid the foundation for my summer internship and the work I did for my Fall 2019 semester.

Working with variables itself was also a big issue as I learned how to use Principal Component Analysis. However, interpreting PCA is relatively difficult as we merge our variables in the process.
I also learnt that having a higher accuracy does not generally solve business problems. Making sure how variables are interpreted to provide findings is also important which is why we focused more on our AUC Score for our project.

IST 718 – Loan Default Analysis | Big Data Analytics

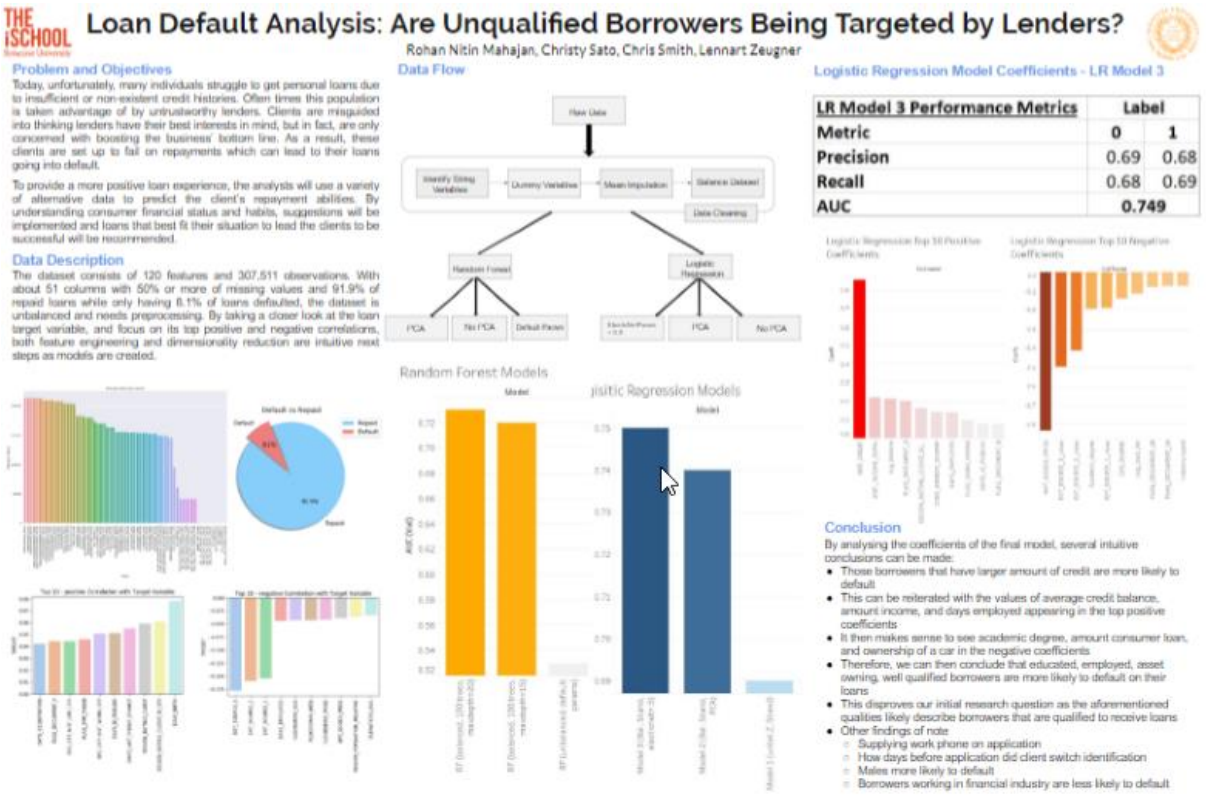
I learned further on data preprocessing and manipulation and using classification algorithms in my Big Data Analytics Project. My group members and I worked on a Home Credit Risk Default dataset where we predicted whether a user would default a loan or not with a variety of alternative data. Our hypothesized problem statement was to answer whether unqualified borrowers are being targeted by lenders.

In this project, knowing what variables to use and describing our dataset was also a big issue. As most of our variables had missing columns, our data needed to be preprocessed before going through any further analysis.

Moreover, our major issue was that our target variable was severely imbalanced. Almost 92% of our data had Repaid loans and 8% of our data had Defaulted loans. For this, we had to first normalize our data in terms of our target variable. This is a common issue in real-world situations as well and as our problem was a classification problem, we would have achieved a high accuracy but a low recall value.

This project taught me that while using models on our dataset is important, data preprocessing, exploratory analysis and the medium that we use to work our models on is as important.
As we had over 120 features (after preprocessing) and around 300,000 rows, we used Principal Component Analysis(PCA) to reduce our variable size for our models.
Furthermore, we also used a non-PCA approach by using the 20 most correlated features to our target variable. Our non-PCA approach gave us a better understanding of our variables and also gave us a better AUC score. The reason behind this is since we reduce our variables using PCA, we also reduce the overall interpretability of them.

While using PCA is highly efficient, interpreting our variables in this problem is also important. Therefore, we went on with our Logistic Regression model, which also performed better than any of our other models for presenting our findings.



Loan Default Analysis: Are Unqualified Borrowers Being Targeted by Lenders?

Rohan Nitin Mahajan, Christy Sato, Chris Smith, Lennart Zeugner

Problem and Objectives

Today, unfortunately, many individuals struggle to get personal loans due to insufficient or non-existent credit histories. Often times this population is taken advantage of by untrustworthy lenders. Clients are misguided into thinking lenders have their best interests in mind, but in fact, are only concerned with boosting the business' bottom line. As a result, these clients are set up to fail on repayments which can lead to their loans going into default.

To provide a more positive loan experience, the analysis will use a variety of alternative data to predict the client's repayment abilities. By understanding consumer financial status and habits, suggestions will be implemented and loans that best fit their situation to lead the clients to be successful will be recommended.

Data Description

The dataset consists of 120 features and 307,511 observations. With about 51 columns with 50% or more of missing values and 91.9% of repaid loans while only having 8.1% of loans defaulted, the dataset is unbalanced and needs preprocessing. By taking a closer look at the loan target variable, and focus on its top positive and negative correlations, both feature engineering and dimensionality reduction are intuitive next steps as models are created.

Data Flow

Logistic Regression Model Coefficients - LR Model 3

LR Model 3 Performance Metrics	Label	
Metric	0	1
Precision	0.69	0.68
Recall	0.68	0.69
AUC	0.749	

Conclusion

By analysing the coefficients of the final model, several intuitive conclusions can be made:

- Those borrowers that have larger amount of credit are more likely to default
- This can be reiterated with the values of average credit balance, amount income, and days employed appearing in the top positive coefficients
- It then makes sense to see academic degree, amount consumer loan, and ownership of a car in the negative coefficients
- Therefore, we can then conclude that educated, employed, asset owning, well qualified borrowers are more likely to default on their loans
- This disproves our initial research question as the aforementioned qualities likely describe borrowers that are qualified to receive loans
- Other findings of note
  - Supplying work phone on application
  - How days before application did client switch identification
  - Males more likely to default
  - Borrowers working in financial industry are less likely to default

We came to the conclusion that surprisingly educated, employed, well qualified, asset owning borrowers were actually defaulting on their loans as compared to less qualified borrowers.

This course was taken by professor Daniel Acuna where we dived deeper into machine learning concepts that included using variables into Principal Component Analysis, using pipelines, focusing on model cross validation and also scratched the surface of using tensorflow modules.

3. Identifying unique patterns by data mining and visualizations and deriving business insights based on analysis – Fall 2019

In my third semester, I went on a more diverse approach on data science. I took 4 courses, Accounting Analytics, Financial Analytics, Natural Language Processing and Data Warehouse all of which were of different fields.

In Accounting Analytics, I learnt how to work with Financial Statement Analysis as well as brushing up on Key Performance Indicators for businesses and clients.

Financial Analytics helped me combine unsupervised learning algorithms with financial concepts. Factors like risk premium and risk free rate helped us determine what factors are used for stock investments.

Rather than focusing on traditional datasets like I did in my previous semesters, I got the opportunity to work with text data, language models and regular expressions in my Natural Language Processing course, which helped me work with different layers of data along with manipulating text data to combine it with classification models.

While I got to work with a huge diverse set of data this semester, I also learned how to load and stage this data to use it for Business Intelligence solutions in my Data Warehouse class.
I did two significant projects this semesters. These include:

IST 664 - Cyberbullying Detection on Social Media | Natural Language Processing

In my Natural Language Processing course, I worked on detecting cyberbullying content on a social media website. As mental health is a huge topic in today's world, my group project members and I decided to work in something related to that field.

We used data from formspring.me, which is a question-answer website open to all users online. We used this text data and used a sentiment algorithm to generalize what was considered as a bullying statement and what was not.

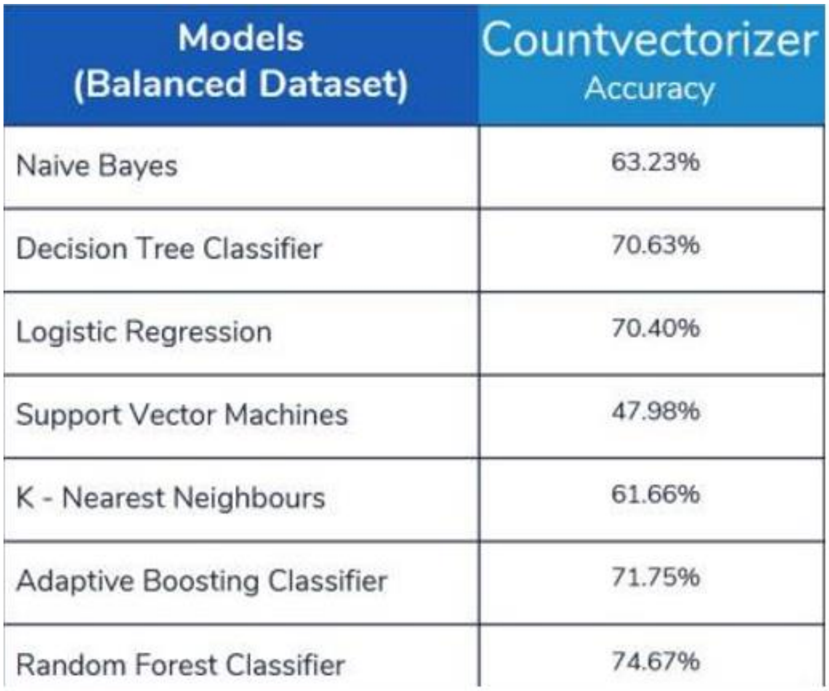
As this data was previously labeled, it was relatively easy to run the classification on it. For this project, just like my Loan Default Analysis project, we went through the same problem which was our labeled data was imbalanced. We had more data that had non-bullying comments as compared to bullying comments. To solve this, we had to sample and equalize both labels.
We then proceeded with cleaning our text data using regular expressions and removal of stopwords. This is important as it increases the importance of words inside the text body and removes redundant unnecessary text.

To use our sentiment algorithm and our classification models, we first need to convert our text data into sentence vectors. For this, we used the TF-IDF and Countvectorizer vectorization methods. We used two methods to make a comparison for later models to evaluate which method was the best method.

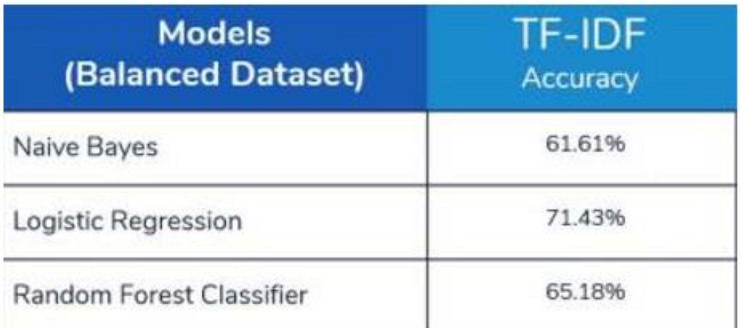
Based on our sentiment algorithm and classification models, we compared it with the previously labeled data and got a decent accuracy for our models to detect Cyberbullying comments.

While working with text data was a primary goal in this project, another primary goal was model evaluation. We used a wide range of models along with our two vectorization methods for making our predictions.

These included basic Logistic Regression, Naïve Bayes, Support Vector Machines and Random Forests. Our model accuracies on our normalized dataset used for the analysis is given below:



Models (Balanced Dataset)	Countvectorizer Accuracy
Naive Bayes	63.23%
Decision Tree Classifier	70.63%
Logistic Regression	70.40%
Support Vector Machines	47.98%
K - Nearest Neighbours	61.66%
Adaptive Boosting Classifier	71.75%
Random Forest Classifier	74.67%



Models (Balanced Dataset)	TF-IDF Accuracy
Naive Bayes	61.61%
Logistic Regression	71.43%
Random Forest Classifier	65.18%

In this case, we found out that our Random Forest Model with the Countvectorizer approach was the best method out of all of the methods used.

This helped us generate basic insights to unsupervised models as well but at the same time, which model is the best in general is very subjective and depends on the problem statement and dataset used.

IST 722 – Business Solutions for Fudgemart Inc. | Data Warehouse

For my Data Warehouse class, we were provided with 2 databases:

1. Fudgemart – A Walmart-like database which had products with their reviews.
2. Fudgeflix – A Netflix-like database which had movies and TV shows with their reviews.

Both databases were comprised under Fudgemart Inc. – A company that was our client for our group project.

My group project members and I were tasked with providing 2 business solutions for Fudgemart Inc. to maximize profits and provide an insight towards the data stored.
Users bought products from Fudgemart and rented/bought movies from Fudgeflix and reviewed both accordingly. Our task was to load, stage and evaluate this data and give a Business Intelligence solution.

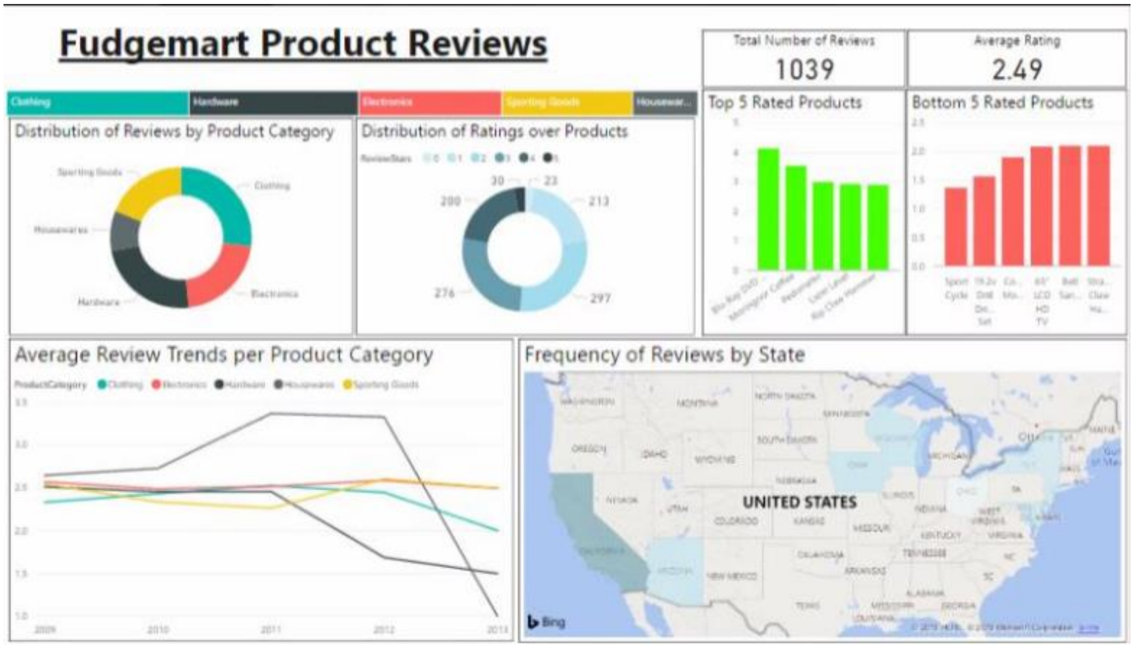
We first modelled our data by conducting High Level and Detailed Level Dimensional Modelling. We then generated SQL queries to load and query our given databases. As Fudgemart and Fudgeflix were two separate databases with different entities, we had to make sure that all of our given datatypes were compatible with each other, as our business processes included manipulation of both datasets.

We focused on Fudgemart's product reviews and Fudgemart Inc.'s total sales and made a trend analysis.

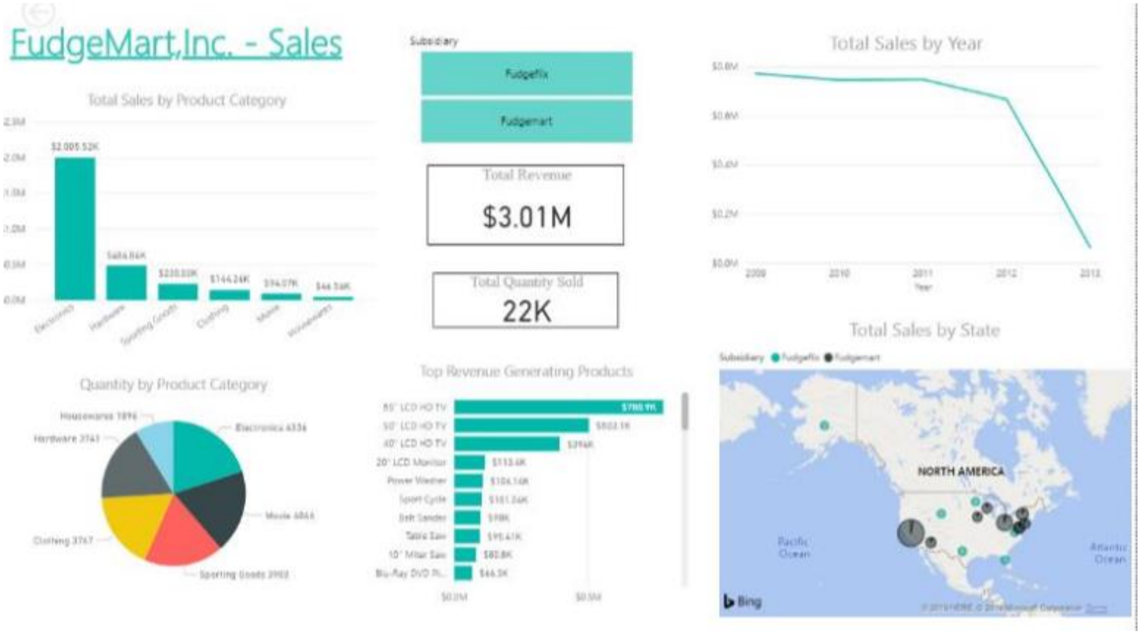
We also highlighted the popular items/categories that were sold and the ones that were relatively lower. This could help Fudgemart Inc. work on increasing sales and reviews for products that do not do well in the market while capitalize over the ongoing popular products for increasing overall profits.

We interpreted our findings and analysis through dashboards using PowerBI.

1. Fudgemart Product Reviews:



2. Fudgemart Inc. Total Sales:



4. Web Scraping and Current Work – Spring 2020

While I have worked on diverse fields in the past 2 years, I have gone back to brushing up on my basic Python concepts this semester in my Scripting for Data Analytics course. This course is currently being taken by Professor Ben Nichols and I am currently working on a social media web scraping project for this course. I am planning on scraping data from reddit.com and creating my own dataset from this raw unstructured data to use topic modeling and trend analysis on news data.

My plan is to have this project as a culmination of everything I have learned in the past along with some new concepts that I am currently learning as I am currently diving deeper into data storage.

I am also currently learning Cloud Computing concepts in my Cloud Management course and am working on a Case Study to provide a Cloud Solution to Ensemble Video for my course.

Conclusion:

While I have learned and am still learning newer concepts as I dive deep into the field of data science, this portfolio is the culmination of all of my work done so far. I plan on learning more on using epochs for models and maybe learn more about Neural Networks. Understanding data science and analytics techniques along with machine learning concepts is important and in my opinion, an aspiring data scientist must be proficient from preprocessing to visualizing to fitting models on datasets. The field and community of data science is growing at such a fast face and I am excited to work in the industry in this field.