National College of Ireland

Statistical Analysis on Second hand vehicle sales using Big Data Technologies and Linear Regression

Programming for Data Analytics
Msc Data Analytics

Yash Balaji Iyengar
Student ID: x18124739

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Yash Balaji Iyengar
Student ID:	x18124739
Programme:	Msc Data Analytics
Year:	2018
Module:	Programming in Data Analytics
Supervisor:	Dr. Muhammad Iqbal
Submission Due Date:	16th August 2019
Project Title:	Statistical Analysis on Second hand vehiclesales using Big Data Technologies and LinearRegression
Word Count:	3141
Page Count:	11

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	16th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Statistical Analysis on Second hand vehicle sales using Big Data Technologies and Linear Regression

Yash Balaji Iyengar
x18124739

Abstract

In the modern age, an automobile vehicle has become more of a necessity than a luxury. Due to the increase in living costs and demands of the modern world variety of vehicles are launched every ear around the world. But not all can afford to buy a brand new car. So in developing countries, people buy vehicles on lease or contract. This gives rise to a market for used vehicles. Unlike the mainstream automobile market, the prices of the vehicles fluctuate in the second-hand sales market. In order to create a stable price range for different vehicles in the market, a methodology has been proposed. This paper uses Craig's list dataset on used cars and performs analysis on it in a Big Data Environment. Regression analysis performed on the dataset and a model has been proposed which can predict the price of the vehicle in the used-car market.

1 Introduction

The automobile industry in any country is one of the major factors to improve the country's economy. Used car sales prediction has gained high importance in recent years. Vehicle production has increased exponentially over the years Gegic et al. (2019). With the growing standard of living of people all over the world, people buy a new car or a second-hand one depending on their budget. In economically emerging nations people lease cars on loans or on a contract basis. At the end of the contract, the car is returned to the dealer. Thus it is very common to see the used car sales market surpass the new automobile sales. Now the car price changes while it is sold in the secondary market. In the car sales market, the cost of the new car is decided by the manufacturer and it remains standard. But in the second-hand car market the cost of a car is relative, varying and depends on a lot of factors like the brand or manufacturer of the vehicle, make of the car, age of the car is, cylinder capacity and type of fuel used Pai and Liu (2018). Due to this, there is no standard set for assigning the selling price of the second-hand car.

Sometimes car dealers take advantage of the ignorance of the customer and inflate the prices of vehicles. So understanding the needs of the market and devising a way to maintain a standard price range for different cars in the second-hand market is an interesting challenge. In other scenarios, the customer is not willing to spend much on the vehicle or many times the customers choice of purchasing a vehicle is influenced by the features of the car and how comfortable the car feels. Due to the scenario discussed above both parties that is the customer and the car dealer trying to figure out what the

other one needs and how do both get the best deal. The car sellers try to advertise their best deals with the help of social media platforms and thus try to increase their trade. In this paper, we will try to analyse a used car sales data set which was made public on Craig's list. The dataset was collected from Kaggle [1] and it consists of 1.7 million records of car sales data from the year 1900 to 2019. The data set is pre-processed and cleaned in R. Then the data is loaded into MySQL. Using Sqoop the data is moved from MySQL to Hive and Hadoop Distributed File System. The data is also moved to Pig. Data analysis is performed in the Pig, Hive and the MapReduce environment. The hdfs stored data is then processed and regression analysis is performed on the data to predict the price of the used cars.

2 Related Work

There is an increase in the number of car sales recently. Residents of the developing nations use the option of leasing the cars instead of buying as it is more economical due to which the sales of used cars have amplified. Keeping this in mind, the car sellers are setting unrealistic prices as there is high demand. Hence there is a requirement to build a model that sets the price for a car by determining its key attributes. This research proposes a supervised learning approach i.e. Random forest to predict the cost of used cars Pal et al. (2019). A random Forest with 500 decision trees was formed in order to train the data. Random Forest is mainly utilized as a Classification model but, in this research, it was employed in the form of a regression model and the related problem was transformed in a corresponding regression problem. A Grid Search Algorithm was deployed to determine the appropriate count of trees and best efficiency deduced was 500 decision trees. The attributes selected is equal to the attributes in the input data. After executing the model, the results determined were 95.82% training accuracy and the testing accuracy deduced was 83.63%. Random Forest provided with better results in comparison with other models which were used in prior researches in the data.

Due to the expanding social media, stating the views on commodities has become effortless. Data available on social media can be used as a potential input in order to predict the sales of vehicles. Stock market estimates to have an impact on sales. This research proposes a methodology to use multivariate regression with social media data and stock market estimates and time-series models are used to forecast monthly total vehicle sales Pai and Liu (2018). The least-square support vector regression (LSSVR) model was employed to handle multivariate regression data. The data used in this research are sentiment score of tweets, stock market values and hybrid data to predict the car sales in the USA. The results specify that predicting vehicle sales using hybrid multivariate regression data combined with deseasonalizing methodologies can deduce accurate prediction outputs as compared to other models. The employment of hybrid data comprising sentimental analysis of social media and stock market estimates can improvise to predict the accuracy. The deseasonalizing methodologies in condition variables and decision variables aid in expanding the forecasting efficiency.

Presently, there is an astonishing amount of alternatives and selections accessible to car buyers. Manufacturers and vendors proceed to examine various ways and use remarkable advertising approaches to satisfy potential consumers to buy their cars. This

[1] `https://www.kaggle.com/austinreese/craigslist-carstrucks-data#craigslistVehicles.csv`

research proposes to examine and create an intelligent system for forecasting the marketing usefulness of cars on the basis of utilizing selected car aspects and supervised neural network prediction model Khashman and Sadikoglu (2019). The neural network prediction model which is based on propagation neural network was employed and its benchmarks enhanced deducing a rapid training time of 0.017s and a proper combined prediction rate (CPR) of 85.07%. Hence ensuring a practical resolution to this feasibility forecasting undertaking.

Gegic et al. (2019) has proposed research on car price prediction which is one of the hot topics in the research field. It states that to build a highly accurate and reliable model of prediction is a complex task where the number of distinct attributes needs to be taken into consideration. As this model is based on car prices in Bosnia and Herzegovina which according to the author is a potentially growing market and the automobile makers must target more strategically. There are three such models applied on the data namely Random Forest, ANN and Support Vector Machin. The data was scraped from the website called autopijaca.ba and the tool used was PHP. The model was given 90% training data and 10% testing data. And data evaluation is done using 10-fold cross-validation. The author is trying to compare these models against each other and find the best performing model based on accuracy. Where the SVM outperformed the ANN in terms of the best model. Finally, for the prediction model integration was done in JAVA.

This research study Zhang et al. (2017) proposes how big data can improve car sales. Making use of data mining techniques and java program to analyse the application of the big data. Data mining is used to help the manufacturer in increasing the production and also to reduce the inventory and also to cut on the wastage of resources. The data was collected of chine market which it had data ranging from 2005 to 2015. The author considered three features from this which are as follows 1. Analysis of a large volume of data rather than small data analysis, 2. Rather than having accuracy look for complication in data, 3. Exploring any unwanted anomalies in data. The approach followed in this research is the KDD methodology as data is very large and it follows unsupervised learning. Through this research, there was many objectives answer like which model is most popular in which category of customer, how many cars to be produced in the coming months and which models give most profit.

Use of big data in the automobile industry is ever increasing. More and more car manufacturer is making use of big data analytics to improve aspects like production, sales and brand value. One such example is AUDI which in paper Dremel et al. (2017) has descried as the digital transformation. The researcher suggests that big data gives insights into new business models, digital services and innovation. There are three models of the use of big data been proposed in this paper. And the how automobile makers are connecting these digital opportunities of business with data services. It also states that there should be a collective force between I.T. and Business dept. With understanding from this research, there should be a cross-departmental and cross-discipline task to further thrive the business.

3 Methodology

The above flow diagram 1 gives a pictorial representation of the entire workflow of the project. For a detailed explanation, this section is divided into the following sections.

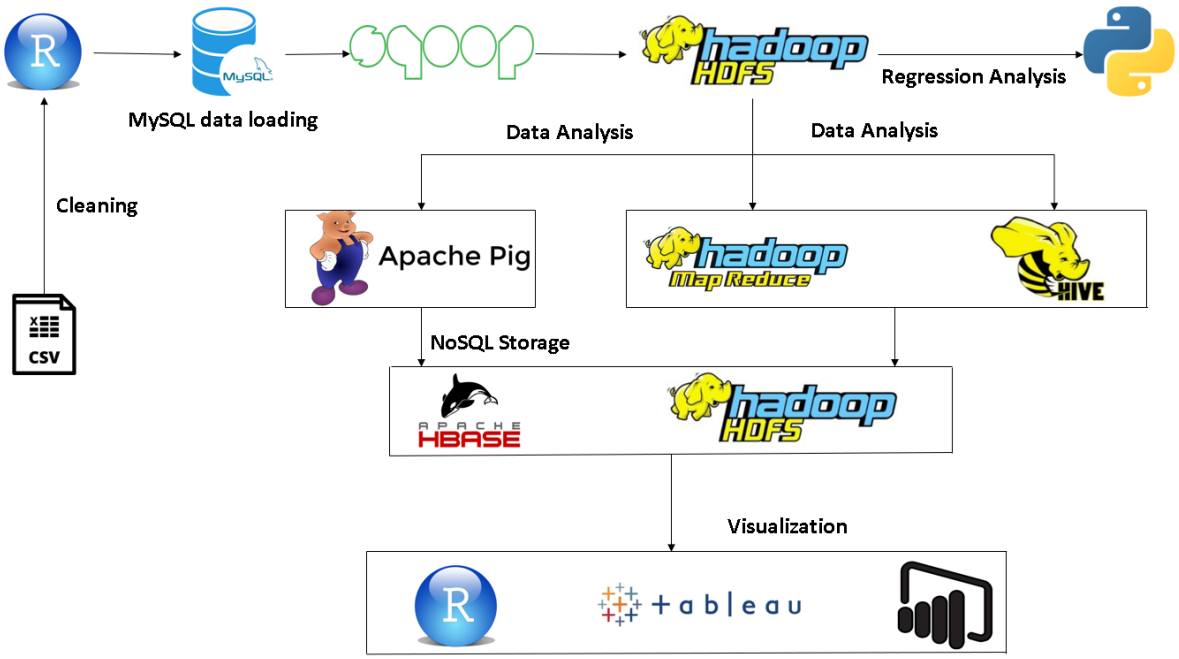

Figure 1: Second-hand Car Sales Analysis and Price Prediction Work Flow

3.1 Environment Setup and Data Pre-Processing

First, an instance was created on NCI OpenStack Cloud Environment. The Big Data Environment was set up from scratch by installing the required software in the following order. Hadoop was installed, then HBase was installed and run in distributed mode. Then MySQL was installed. A Test Database was created and a dummy table was created for testing purposes. Then Sqoop was installed. Then the functionality of Sqoop was tested by transferring the dummy table from MySQL to HDFS and HBase. Then Pig and Hive were installed in the virtual environment. The data set is obtained from Kaggle [2]. The data set consists of 26 features and 1.7 million rows. The data is downloaded and loaded in the R - studio environment. The data set is then checked for null values using the Amelia library and miss_mapp function. This function visualises the data and gives a distribution of the missing values in the dataset. Unwanted columns were dropped and the null values were omitted using the na.omit function available in base R. Factor levels of the categorical variables were checked and altered to avoid data anomalies and duplication of data. The final clean data set consisted of 331473 rows and 20 columns.

3.2 Data processing in Big Data Environment

In order to improve the processing and loading performance, we load the data into the MySQL database. The dfs and yarn services are switched on. First, the dataset is loaded into the local memory in the house profile. A new database is created named cars and a schema is created describing the data types of each column in the data. Then the data is loaded into MySQL. This data is then loaded to the Hadoop Distributed File system Environment. Since the size of the data is huge we use Sqoop for transferring the data.

[2] `https://www.kaggle.com/austinreese/craigslist-carstrucks-data#craigslistVehicles.csv`

Sqoop is a software developed by Apache to facilitate the transfer of huge chunks of data between relational databases and hdfs [3]. Hadoop uses the java database connectivity and MySQL to connect with different databases, so the data is transferred to HDFS from MSQL. Similarly, the data is transferred to Hive using Sqoop. The data stored on Hadoop is then processed using multiple frameworks for exploring and analysing the data. Mapreduce is used for our two BI Queries. Hive is used to calculate three BI Queries and Pig is used for two BI Queries. We will discuss them in detail in the next section.

3.3 Software Used

3.3.1 Hadoop Mapreduce:

It is a processing framework which works on HDFS [4]. We have used it to calculate the used cars price distribution over the years and also to find out the vehicle of which size is preferred by the customers. The entire Mapreduce functioning can be divided into two parts and the code can be explained in three parts. The mapper assigns a key-value pair to each element in a column and passes it to the reducer. The reducer performs an aggregate function on the input key-value pair received from the mapper depending on the application.

3.3.2 Hive:

Hive is a MySQL like database which is used to store large datasets and alter them [5]. We use Hive for processing our data to find answers to our three BI Queries they are which state has most listings of used cars, Which manufacturer is more trusted in the second-hand vehicle market and the third query is the vehicle with what cylinder capacity is sold most. The select, group by and the count syntax have been used to calculate the above queries. The output of the BI queries has been stored in HDFS

3.3.3 Pig:

It is a framework that is used for exploring patterns in datasets with a scripting syntax. We have used Pig to find out the most popular vehicle type sold in the used car market and customer's tendency to buy which colour vehicle. The group by, for each and the count functions, have been used to answer the BI Queries. The output of the queries has been stored in HBase database.

3.3.4 HBase:

HBase a NoSQL database which is used to store semi-structured as well as structured data. It is integrated with Hadoop and it is used to store vast and widespread data. It has column families and is a schemaless database. In this project, we have used HBase to store the output of the BI queries obtained from Pig framework.

[3] `https://sqoop.apache.org/`

[4] `https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html`

[5] `https://hive.apache.org/`

3.3.5 R - Programming Language:

R is a statistical language used for statistical analysis. We have used it for data pre-processing and in the final stage for visualization of the BI queries.

3.3.6 Python Programming Language:

It is a programming language used extensively in the data science domain. In this project, it is used for performing regression analysis and building a model that predicts the price of used cars. The results of the model will be discussed in the results section.

3.4 Tableu and PowerBI:

Both are data visualization tools used to plot data and extract important meaningful insights from the data. We have visualised our data with the help of this software.

4 Results

The results of the analyses performed on the dataset will be explained in this section.

4.1 BI Query 1: Trend of Used Car Sales over the years

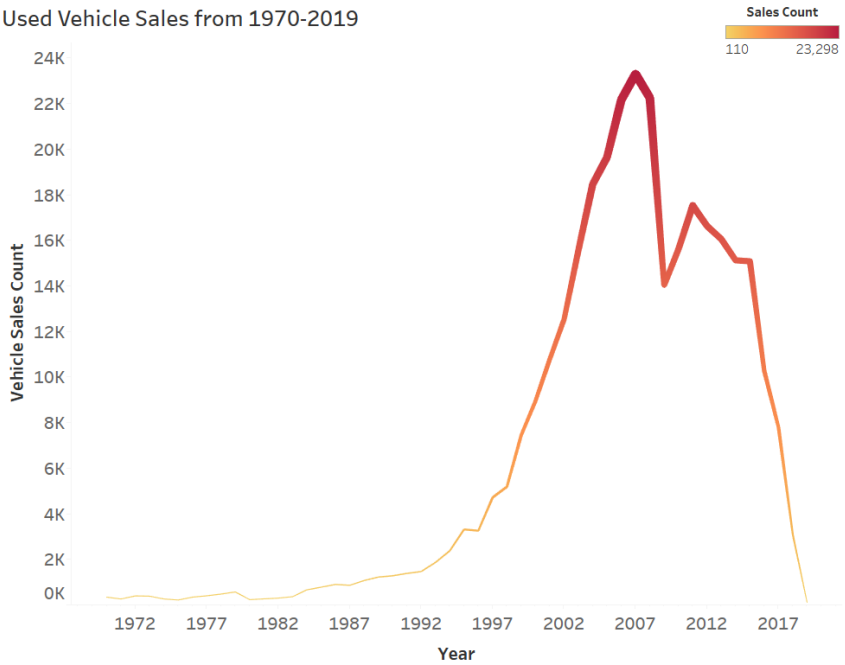

Figure 2: Results for BI Query 1

In Figure 2 Count of car Sales has been plotted against years in a line graph. This output was obtained from Hadoop MapReduce. From the trend, we can see that there is a gradual increase in the second car sales market from the year 1985 and it can be observed that there is a steep descend in the sales in the year 2009. This might have happened because of the recession and the economic loss that year.

4.2 BI Query 2: Can an inference be made based on the vehicle size distribution ?

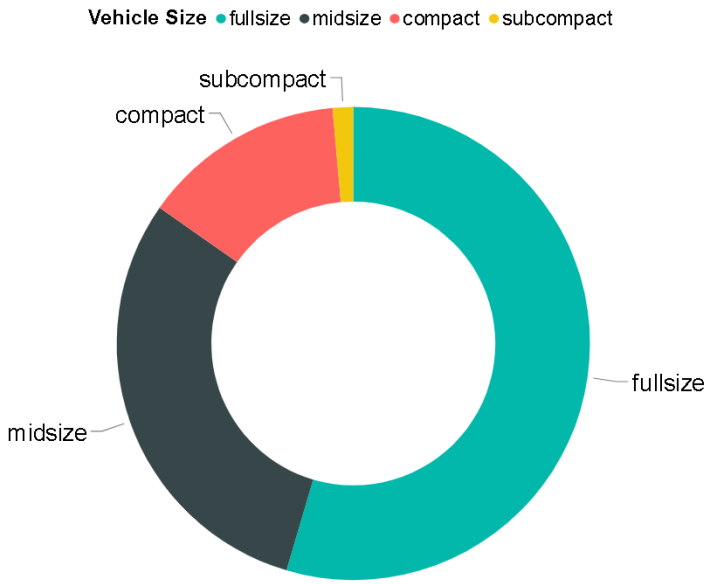

Figure 3: Results for BI Query 2

In Figure 3 the vehicle size variable has been plotted in a doughnut chart. The output file of this query is obtained by performing analysis on Hadoop MapReduce. The chart shows the tendency of the majority of customers to buy full-size vehicles. From this, we can infer that main customers buy heavy vehicles and must be truck drivers or lorry drivers or delivery contractors. So based on this data new marketing strategies can be designed to increase sales.

4.3 BI Query 3: Statewise distribution of car sales

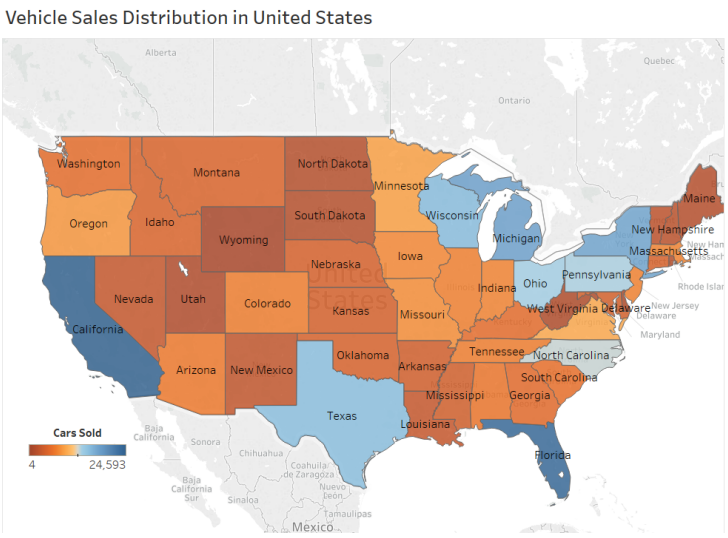

Figure 4: Results for BI Query 3

Above Figure 4 shows the state-wise distribution of used cars sales. It can be observed that Florida and California have established themselves as the top leading second-hand vehicle market followed by Texas, Michigan and Wisconsin. This gives the customers a better idea of where to find reliable car dealers and helps the vehicle sales company new customer base to target.

4.4 BI Query 4: Which Manufacturer is the most trusted in the second hand vehicle sales market ?

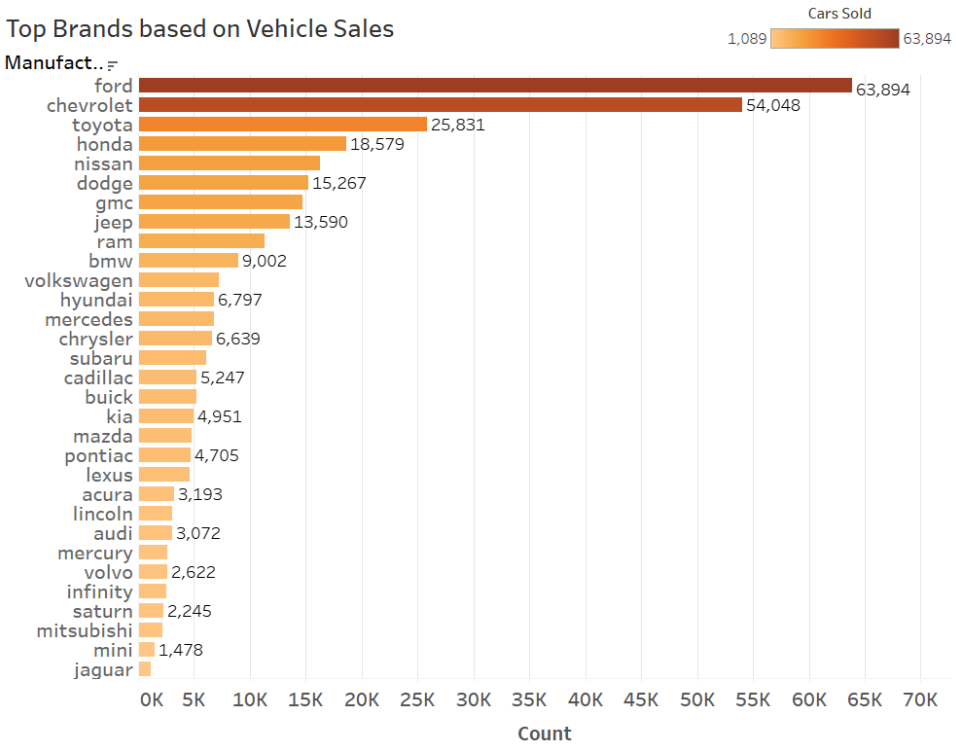

Figure 5: Results for BI Query 4

Figure 5 shows the most trusted brand in the used cars market. This query is processed in the hive framework. Ford and Chevrolet lead the market by a high margin followed by Toyota, Honda, Nissan and others. So the vehicle dealers can set a price of the vehicle based on this statistic and the customer can benefit by narrowing down the range of vehicles to search from to purchase.

4.5 BI Query 5: Vehicle Sales distribution based on number of cylinders in a vehicle

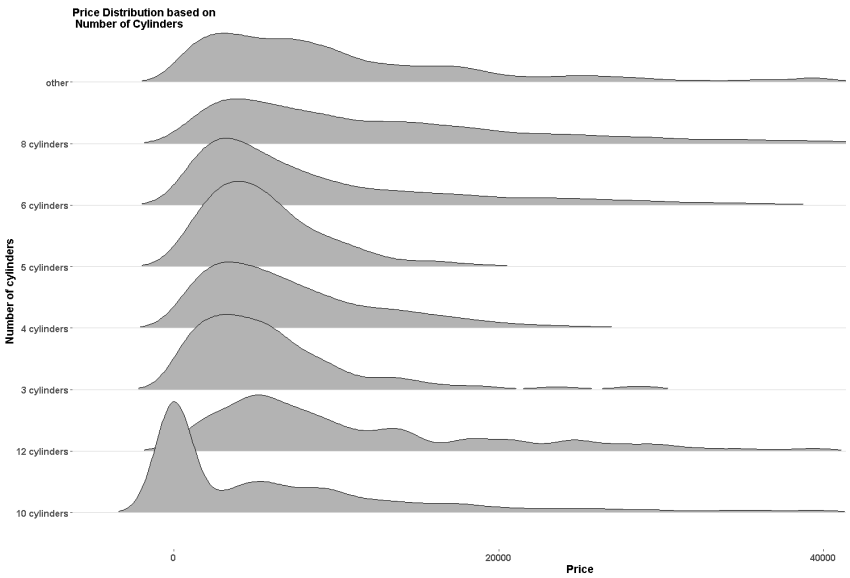

Figure 6: Results for BI Query 5

The data for this query has been processed in Hive and we can observe that customers prefer a vehicle with 10 cylinder engine. This visualization has been plotted in R programming language.

4.6 BI Query 6: Which type of vehicle are most sold in the market?

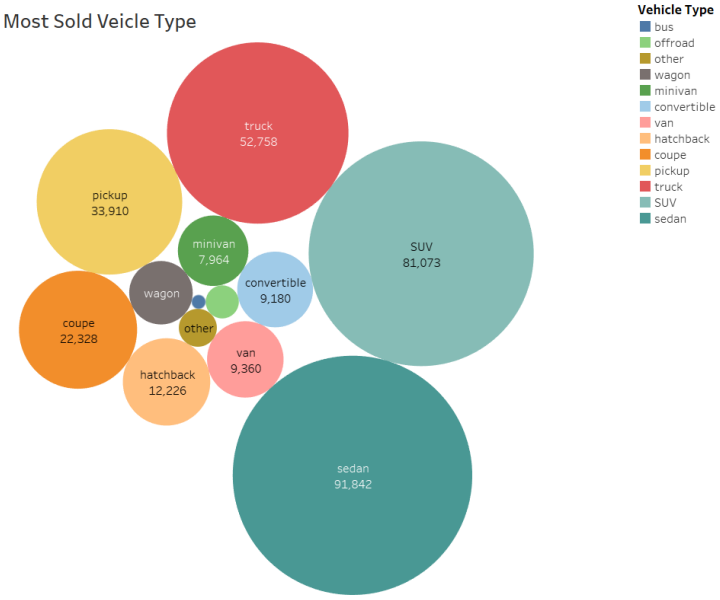

Figure 7: Results for BI Query 6

This query has been processed in the Pig Framework and the output has been stored in HBase as discussed in subsection 3.3.3. We can see that Sedan, Truck and SUV are most purchased. If we observe closely we can co-relate it with the 4.2. Both these queries bring more credibility to our assumption that the main customers are the truck drivers and delivery contractors in the second-hand vehicle sales industry.

4.7 Linear Regression in Python

The processed data in HDFS is then used for Regression Analysis to build a model for predicting the prices of the used cars. First, the clean data is loaded into the Jupyter environment. Then Exploratory Data Analysis is performed and outlier values are checked. The price and odometer column are filtered to rid the data from anomalies and outlier values. Only numeric features are selected which are the year, odometer and price is the dependent variable. Then the data is portioned into training and testing with partition split ratio as 75 to 25. Linear Regression model is run on the training data to obtain a root means square error value of 7549.25. The model is then checked on test data to get an RMSE value if 7488.68. This means, the model correctly predicts the price of the cars 74.88% of the times.

5 Conclusion

In this study, the Hadoop and its Big Data Family features have been used to demonstrate the ease of working on large datasets in a virtual environment. In this paper, an analysis has been performed on the used cars dataset which helps us understand the various factors and its importance on the dependent variable that is price. A proper workflow has been devised as mentioned in the figure1 and all the tasks have been performed according to the design flow.

References

Dremel, C., Herterich, M., Wulf, J., Waizmann, J.-C. and Brenner, W. (2017). Digital Transformation Often Requires Big Data Analytics Capabilities1 2 How AUDI AG Established Big Data Analytics in Its Digital Transformation, *MIS Quarterly Executive* **16**(81): 81–100.
URL: *http://www.misqe.org/ojs2/index.php/misqe/article/viewFile/765/459*

Gegic, E., Isakovic, B., Keco, D., Masetic, Z. and Kevric, J. (2019). Car price prediction using machine learning techniques, *TEM Journal* **8**(1): 113–118.

Khashman, A. and Sadikoglu, G. (2019). *Data Coding and Neural Network Arbitration for Feasibility Prediction of Car Marketing*, Vol. 2, Springer International Publishing.
URL: *http://dx.doi.org/10.1007/978-3-030-04164-9_34*

Pai, P. F. and Liu, C. H. (2018). Predicting vehicle sales by sentiment analysis of twitter data and stock market values, *IEEE Access* **6**: 57655–57662.

Pal, N., Arora, P. and Kohli, P. (2019). How Much Is My Car Worth ? A Methodology for Predicting Used Cars ' Prices Using Random Forest, Vol. 1, Springer International

Publishing, pp. 413–422.
URL: *http://dx.doi.org/10.1007/978-3-030-03402-3_28*

Zhang, Q., Zhan, H. and Yu, J. (2017). Car Sales Analysis Based on the Application of Big Data, *Procedia Computer Science* **107**(Icict): 436–441.
URL: *http://dx.doi.org/10.1016/j.procs.2017.03.137*