INTERNSHIP PROJECT REPORT

(Project Term January-June, 2018)

Car and BOSCH OEM Sales Forecast for Indian Market (Predictive Analysis, Machine learning)

Submitted by

Ankit Parichha **Registration Number: 11410859**

Course Code: CSE441

Under the Guidance of

Mr. Ponvannan Ponnuramu, (RBEI/BSW5)
Senior Architect

LOVELY PROFESSIONAL UNIVERSITY

CERTIFICATE FROM THE COMPANY

CERTIFICATE OF ORIGINALITY

This is to certify that the project report entitled **"Car and BOSCH OEM Sales Forecast for Indian Market"**, submitted to Lovely Professional University, Phagwara in partial fulfillment of the conditions for the award of B.Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara is an original work carried out by **Mr. Ankit Parichha**, under guidance of **Mr. Ponvannan Ponnuramu**. The matter embodied in this project is a genuine work done by Ankit Parichha to best of my knowledge and belief and has not been submitted before, neither to this University not to any other University for the fulfillment of the requirement of any course of study.

Signature of the Student **Signature of the Guide**

Designation

ACKNOWLEDGEMENT

The satiation and euphoria that accompany the successful completion of project would be incomplete without the mention of the people who made it possible.

I would like to take the opportunity to thank and express my deep sense of gratitude to my corporate mentor Mr. Ponvannan Ponnuramu and Project manager Mr. Badrinarayanan G R. I am greatly indebted to both of them for providing their valuable guidance at all stages of study, you advice, constructive suggestions, positive and supportive attitude and continuous encouragement, without which it would not have been possible to complete the project.

I owe my whole hearted thanks and appreciation to my entire team for their corporation and assistance during the course of my project.

I hope that I can build upon the experience and knowledge that I have gained and make a valuable contribution towards this industry in coming future.

Ankit Parichha

LPU, Phagwara

Punjab

Name of the Student **Signature**

TABLE OF CONTENTS

1. INTRODUCTION

Car and BOSCH OEM sales prediction is the predictive analysis for forecasting of future car sales for the next nine months and using those sales predict the BOSCH OEM sales for automotive car parts.

1.1 MACHINE LEARNING

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine learning algorithms are often categorized as supervised or unsupervised.

Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the

data and can draw inferences from datasets to describe hidden structures from unlabeled data.

Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources.

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information.

2. PROFILE OF PROBLEM

The project is a console app created in R studio with very easy to use commands and data in an interactive and simple way which could be easily understood by people.

The challenge is to accurately predict future Car and OEM Sales using predictive analytics and machine learning and then to identify the optimal strategy for OEM sales forecast.

Predictive modeling and data visualization used to bring Car Sales Forecast by make, segment and fuel, BOSCH OEM products forecast and identify the optimal strategy for BOSCH OEM sales forecast.

3. EXISTING SYSTEM

3.1 INTRODUCTION

There are numerous number of systems present for prediction of values or sales but this is very specific for BOSCH OEM and before this project there is no project which can implement such type of prediction.

3.1.1 TOOLS USED

For prediction of any regression type of system there are very few algorithms which can work very fine but we need to tune the system in such a manner that it won't affect much of the overall change rather than help us to get more accuracy. Here we have used r language for scripting the project and all the full stack web development for creating the interface. RStudio is used for scripting of the project which uses R language for processing.

3.1.2 RSTUDIO

RStudio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. RStudio was founded by JJ Allaire, creator of the programming language ColdFusion. Hadley Wickham is the Chief Scientist at RStudio.

RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. Prepackaged distributions of RStudio Desktop are available for Windows, macOS, and Linux.

RStudio is available in open source and commercial editions and runs on the desktop (Windows, macOS, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian, Ubuntu, Red Hat Linux, CentOS, openSUSE and SLES).

RStudio is written in the C++ programming language and uses the Qt framework for its graphical user interface.

Work on RStudio started around December 2010, and the first public beta version (v0.92) was officially announced in February 2011. Version 1.0 was released on 1 November 2016. Version 1.1 was released on 9 October 2017.

RStudio develops open source and enterprise-ready professional software for the R statistical computing environment. Our products simplify data analysis with R and provide powerful tools for publishing and sharing.

The RStudio open source and commercial Integrated Development Environment is the premier IDE for the R programming language. RStudio Connect is the place to publish all the work your teams create in R. RStudio Shiny, Shiny Server, Shiny Server Pro and shinyapps.io help you create and publish interactive web applications. The RStudio team also contributes code to many R packages and projects.

3.2 DFD FOR CAR AND OEM SALES FORECAST

We have collected total of 2 datasheets from different places and each of them represent some specific data for some particular use.

The project is completely based on regression type which makes us more complex to achieve the complete result with simple algorithms.

One of the dataset contains list of all the car brands along with their model names. With respect to each car model we have added one dependent variable that is the total sales and independent variables such as GDP rate, Petrol Price, Diesel Petrol, Unemployment rate for every particular month.

The Data flow diagram displays how the historical data that is the sales for all the cars collected for 2 years have been divided into two parts – training and testing. Training and Testing datasets are used to make the machine learn and check whether the prediction is correct or not.

As shown in Figure 1, the data set is used to train the system and using some tuning it should be used for testing and finding the accuracy.

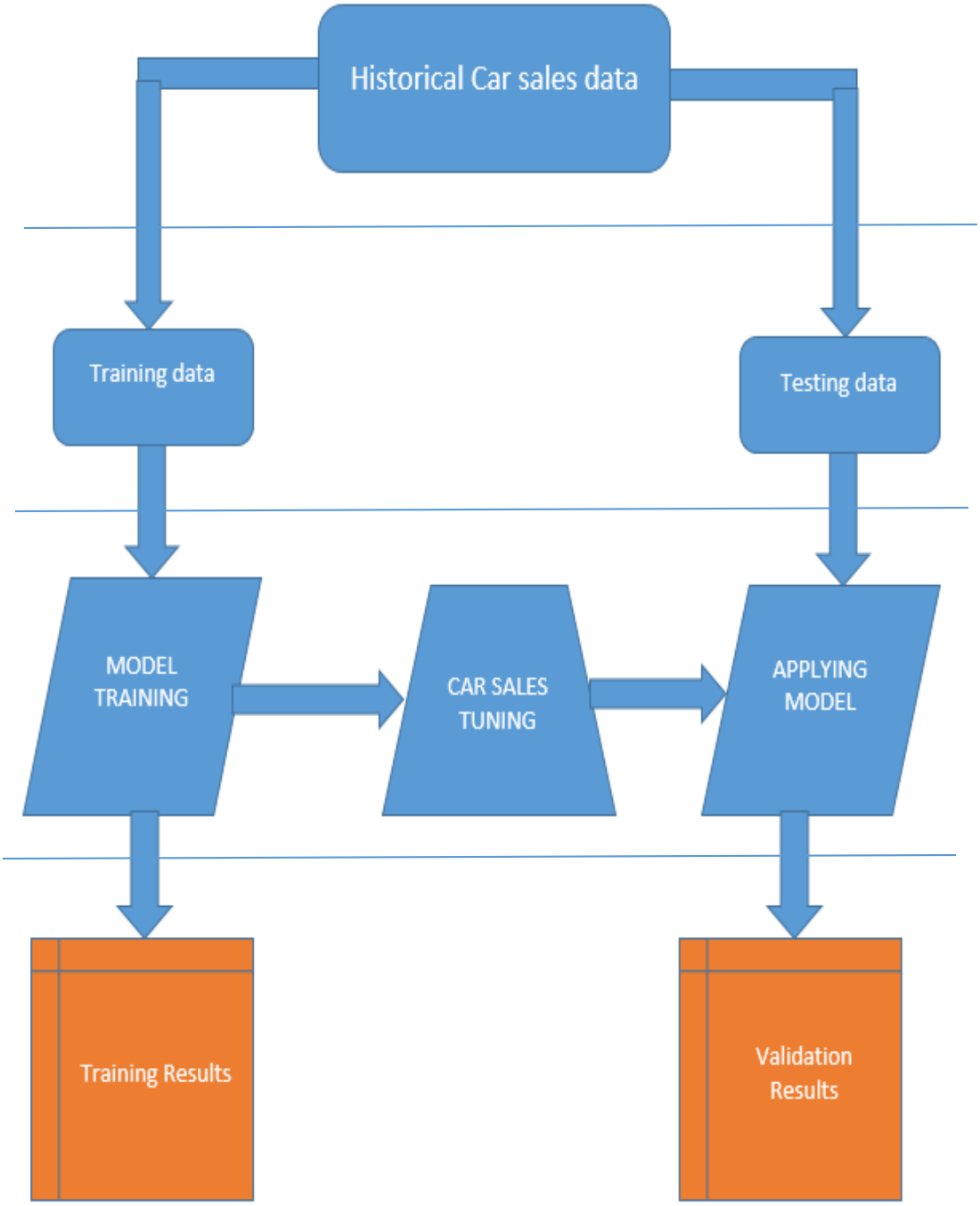


Figure 1: 3.2 DFD FOR CAR AND OEM SALES FORECAST

3.3 WHAT'S NEW IN THE SYSTEM TO BE DEVELOPED

As prediction can never be done with 100% accuracy there will always be some problem with the system and with the accuracy. Sometimes there may be drastic change which may create some problem but we can still train the system and make sure that for the future the problem should be faced with more accuracy. If there are some software which will allows to train the system in this manner then many of the problem can be solved.

Every company wants to do machine learning on a bigger scale and for less cost. Cloud service providers will continue to compete to drive down the costs and increase the capacity of machine learning systems. We've seen Google's cloud services grow from storage to include a suite of machine learning tools across language, speech and images. They have gone so far as to build custom hardware—the tensor processing unit—for helping users train their own machine learning systems quickly. Amazon's AWS and Azure have similar offerings. The end result is a democratization of large scale machine learning tools and infrastructure.

Every research team wants to do machine learning with less data. Acquiring data is expensive and time consuming. One common case: The seemingly simple task of predicting the names of objects in an image can require showing a machine thousands of examples of each object. Researchers are investing in methods to reduce the number of examples needed. A comparison that is often drawn is human child development. How many times does a child have to observe an object before knowing what it is and how it is used?

And because of bigger, faster, cheaper machine learning, which is more accurate with less data, we will see the number of applications and use cases of machine learning continue to rise across all sectors.

4. PROBLEM ANALYSIS

4.1 PRODUCT DEFINITION

Car and OEM sale prediction is an application which predict car and OEM (Original Equipment Manufacturer) sale for upcoming nine months. The application predicts the sale using three different regression models. The application predicts the sale for 104 different car and 15 different OEM's which BOSCH manufactures. With the help of this application, client can easily get to know about the future sale and can work according to that. There are many factors which are dependent on car sale like GDP, petrol price, fuel price and Unemployment rate. With the help of this application the manufacturers get to know about the demands of user like which kind of car is in demand. With the help of this application, manufacturer can also know how many workforce are required to build the cars and its equipment. For providing high computation power azure virtual machine is used. The application can be accessed from anywhere with the help of internet. The data used for prediction is directly from the car manufacturer. Regression algorithm are used to predict Car and OEM sale. Data splitting is done in efficient way for getting higher efficiency. If we will take average sale of upcoming month, then it cannot predict sudden increase and decrease in the sale which may depend on many factors such as GDP, fuel price, unemployment rate. Also the UI of the application is very simple which does require any prior knowledge of using. The UI is very light which can be easily loaded.

4.2 FEASIBILITY ANALYSIS

4.2.1 TECHNICAL FEASIBILITY

All the technical requirements can be easily fulfilled as our application is hosted on azure cloud which gives us enough computing power to handle millions of data. The project is developed in R studio using R script:

- Only the models which works on Regression analysis can be applied.
- Regression algorithms when modified can lead to better result in terms of car and OEM sale prediction.
- User friendly GUI is developed so that naïve user can use this application
- Open Source packages and software are used while development of this application which makes it financially feasible.
- Our approach converts the regressional analysis into classification which makes it possible for finding the accuracy for every model.
- Our application is highly available.

4.2.2 LEGAL FEASIBILITY

Figure 2: 4.2.2 R LANGUAGE

Figure 3: 4.2.2 AZURE FOR DEVELOPMENT

The data used in the application are provided by the car manufacturers which are legal. All the packages and tools which are used are open source except Azure. Azure have special feature of pay for what you use.

All the tool's version that are used are latest. All the package that are used are provided by R studio.

4.3 PROJECT PLAN

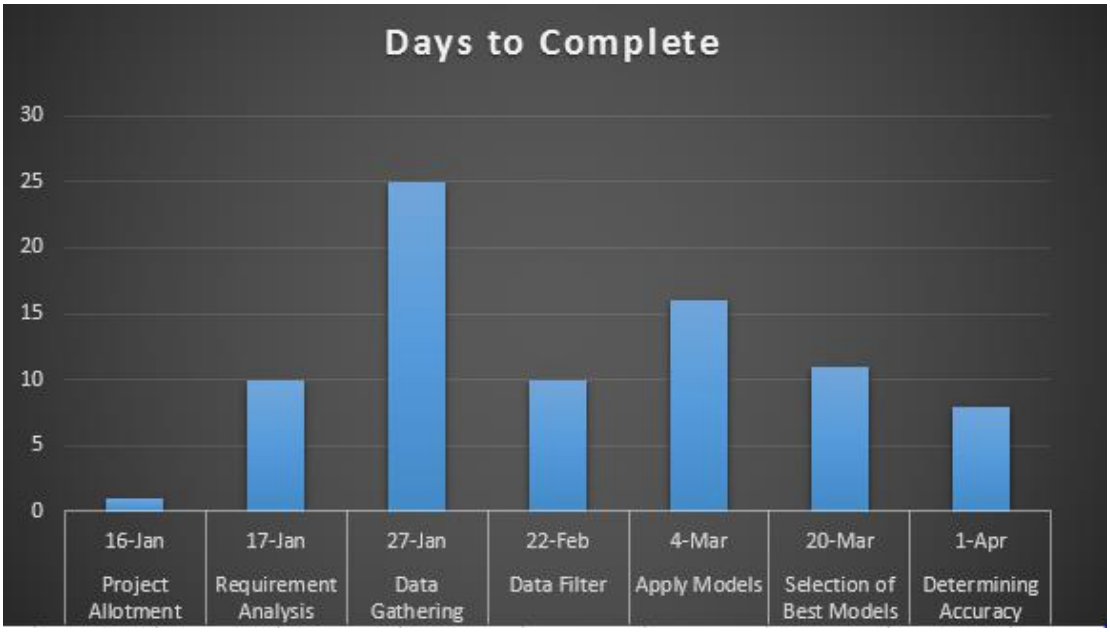


Figure 4: 4.3 PROJECT PLAN FOR CAR AND OEM SALES FORECAST

5. SOFTWARE REQUIREMENT ANALYSIS

5.1 INTRODUCTION

This section will contain all functional and quality requirements of the system. It will give a detailed description of the system and all its features.

5.2 EXTERNAL INTERFACE REQUIREMENTS

5.2.1 USER INTERFACE

Firstly, the user have to select the car manufacturer name, then car name for that particular manufacturer name. After selecting car name, then user have to select the prediction or data visualization options. With data visualization option, user can see the variation of data. With prediction option user can total car sale and OEM sale for upcoming 9 months.

5.2.2 SOFTWARE REQUIREMENTS

The application does not require any high specification. It only requires high speed of internet connectivity. The application can also be accessed with the help of mobile.

5.3 FUCTIONAL REQUIREMENTS

5.3.1 FUNCTIONAL REQUIREMENT 1

ID: FR1

TITLE: Select Car Manufacturer Name

DESC: User have to select particular car manufacturer name from the dropdown list. User can see only those car which are manufactured by those manufacturer which was selected by the user.

5.3.2 FUNCTIONAL REQUIREMENT 2

ID: FR2

TITLE: Select Car Name

DESC: User have to select car name for that particular brand.

5.3.3 FUNCTIONAL REQUIREMENT 3

ID: FR3

TITLE: Get Prediction

DESC: With the help of Prediction user can see next nine month car and OEM sale. The prediction is done using three different models. User can also see the graph of next nine month car sale.

5.3.4 FUNCTIONAL REQUIREMENT 4

ID: FR4

TITLE: Get Accuracy

DESC: User can see the accuracy of the model in percentage. User can also compare the accuracy with help of line graph in which green line represents actual data and blue line represents the predicted data.

5.3.5 FUNCTIONAL REQUIREMENT 5

ID: FR5

TITLE: View Data

DESC: With the help of data visualization user can see the data in scatter plot, line plot and histogram. User can also check correlation between them. User can check also check monthly sale of that car.

5.3.6 FUNCTIONAL REQUIREMENT 6

ID: FR6

TITLE: Reset

DESC: With the help of reset, user can reset all the value which are selected by earlier.

5.3.7 FUNCTIONAL REQUIREMENT 7

ID: FR7

TITLE: Get RMSE

DESC: With the help of RMSE (Root Mean Square Error) user can find the difference of the car between actual and predicted car sale.

5.3.8 FUNCTIONAL REQUIREMENT 8

ID: FR8

TITLE: Continue

DESC: User can see prediction of other car after checking prediction of one car. Data for selection of testing and training data is done randomly.

5.4 PERFORMANCE

High Performance Virtual Machine is used for handling large amount of data. All the data are stored in azure which provide high availability and accessibility. The application is also hosted in azure, which can be accessed from anywhere. Data splitting among training and testing data is also done in efficient way which provide high accuracy.

6. DESIGN

6.1 INTRODUCTION

Development of the application is troublesome as the problem demands Regression Values. Prediction for occurring of an event in future is easy but to predict the numerical value for that event is troublesome. This type of predictions demands changes to be made in the algorithms according to the dataset. Regression Design was followed while development of this applications. The overall design and flow of the data is shown below:-

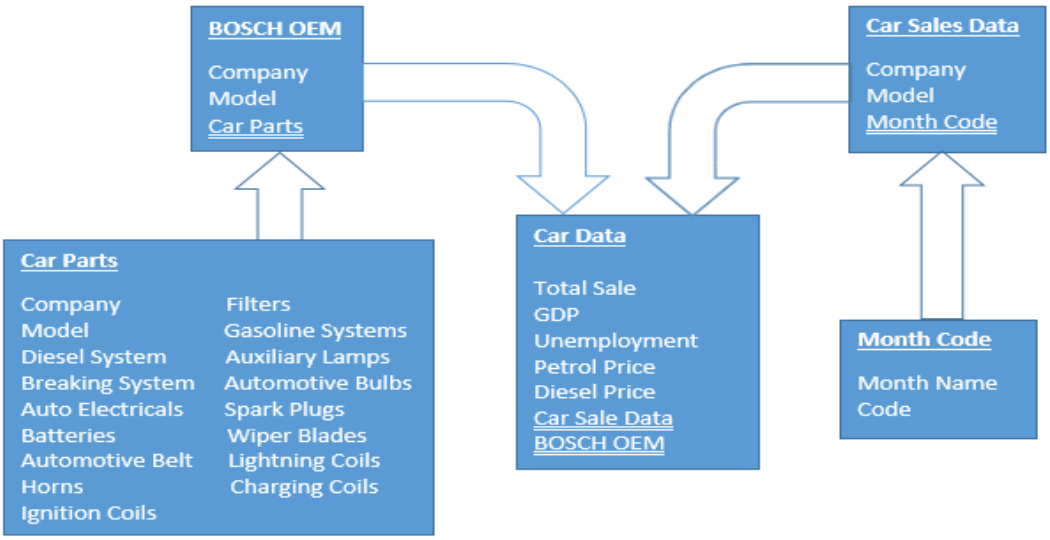
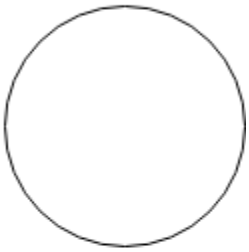


Figure 5: 6.1 DATA FLOW FOR CAR AND OEM SALES FORECAST

6.2 DESIGN NOTATIONS:-

EXTERNAL ENTITY

A producer or consumer of information that resides outside the bounds of the system to be modeled.



PROCESS

A transformation of information (a function) that resides within the bounds of the system to be modeled

DATA OBJECT

A data object; the arrowhead indicates the direction data of data flow.

DATA STORE

A repository of data that is to be stored for use by one or more processes; may be as simple as a buffer or queue or as sophisticated as a relational database.

6.3 DETAILED DESIGN

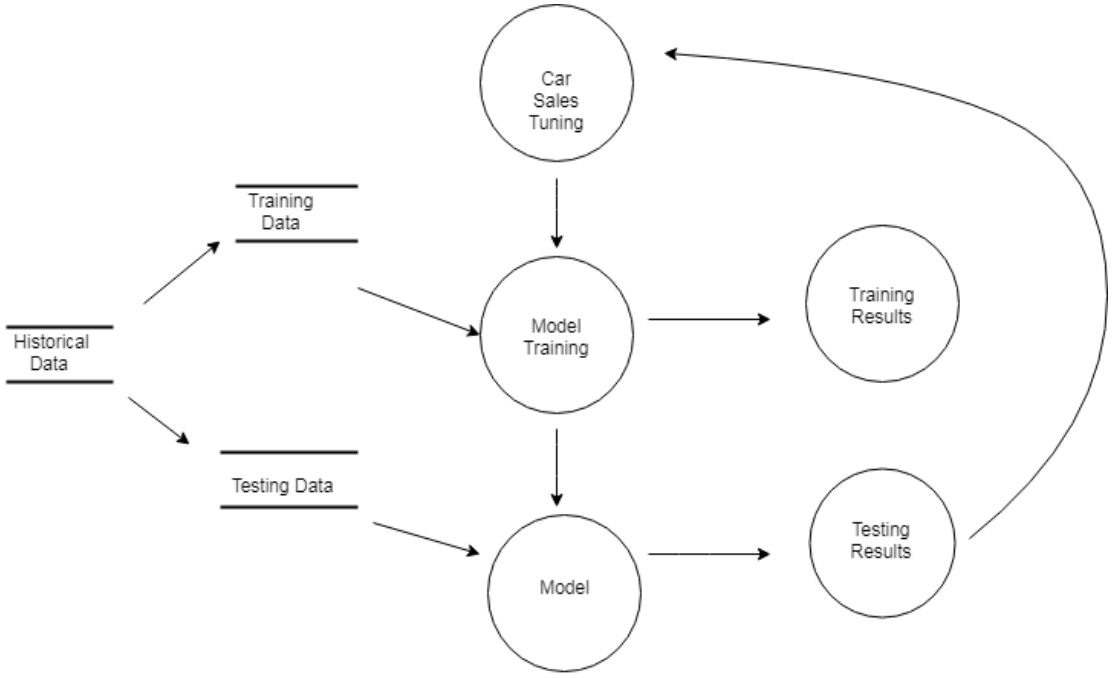


Figure 6: 6.3 DFD FOR CAR AND OEM SALES FORECAST

We have collected total of 2 datasheets from different places and each of them represent some specific data for some particular use.

The project is completely based on regression type which makes us more complex to achieve the complete result with simple algorithms.

One of the dataset contains list of all the car brands along with their model names. With respect to each car model we have added one dependent

variable that is the total sales and independent variables such as GDP rate, Petrol Price, Diesel Petrol, Unemployment rate for every particular month.

The Data flow diagram displays how the historical data that is the sales for all the cars collected for 2 years have been divided into two parts – training and testing. Training and Testing datasets are used to make the machine learn and check whether the prediction is correct or not.

As shown in Figure 6, the data set is used to train the system and using some tuning it should be used for testing and finding the accuracy.

6.4 ARCHITECTURE

Various different architecture were decided to work upon. Since this project is to be implemented on large scale across the organization so first it was developed as a prototype model and small amount of live data was fed in, to check the offline evaluation of live data.

Further the prototype model was modified to be deployed on cloud to work upon live data and online evaluation was performed. It took less than a minute for predictions once the data set was loaded completely into the system either on cloud platform or on offline systems.

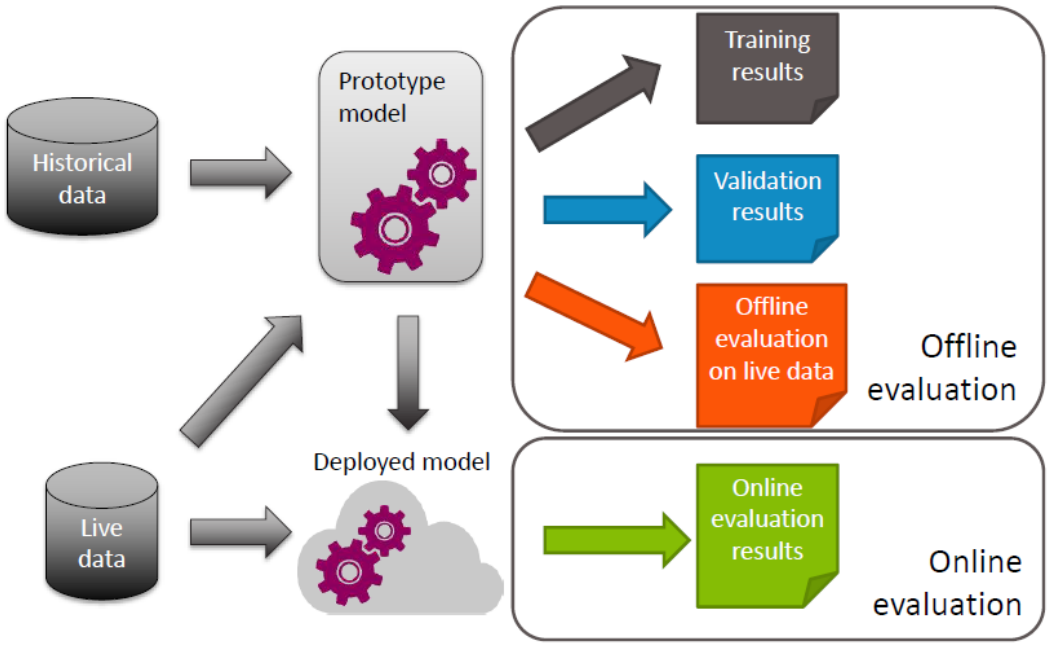


Figure 7: 6.4 ARCHITECTURE FOR CAR AND OEM SALES FORECAST

6.5 PSEUDO CODE

After going through all the classes and demo applications provided by google this was how the application was supposed to work or in exactly this order we are supposed to move in the application cycle to work properly:

i. Installing various library required for implementation of predictive analysis.
ii. Including all the installed packages that were installed
iii. Data Set Inclusion:-
    a. Reading the data set provided (parts , variables, car sales)
    b. Dividing the data set into 70:30 ratio (70 for training set and 30 for testing set)
iv. Conversion of Month from format dd/mm/yyyy to Month Code in form of numerals.
v. Taking the user input for a car company.

vi. Checking if the car company name entered by user exists in our data.
vii. Asking for the particular model from list of all models available for particular car company.
viii. Forecasting the various features (GDP, Unemployment etc.) for next nine months
ix. Applying Various Models (Random Forest, Linear Regression, Logistics, Time Series Analysis, and Decision Tree) for training our model with 70% of the data set.
x. Prediction of sales from forecasted data and model trained for next nine months.
xi. Conversion of Regression Error to Classification Error for accuracy calculation possible for models like Random Forest.
xii. Selection of optimum error rate so that our accuracy calculated binds with the real accuracy.
xiii. Calculation of RMSE (Root Mean Square Error) for differentiating which model fit our predictions best.
xiv. Plot the prediction on Graph.
xv. Plot the Various Attributes against Total Sales for analysis of the dependency.
xvi. Mapping the Car Sales predictions with OEM parts.
xvii. Calculation of exact OEM parts required to be manufactured in next nine months.
xviii. Displaying and visualizing the predictions on the web app.

7. TESTING

7.1 FUNCTIONAL TESTING

Functional Testing is a testing technique that is used to test the features/functionality of the system or Software, should cover all the scenarios including failure paths and boundary cases.

Testing all functions in application independently as well as parallely.

7.1.1 TESTING EACH FUNCTION INDEPENDENTLY

Each functions that do not have any dependency is checked.

1. Testing all data set manually so that no row has N/A values.
2. Testing for reduction of outliers in our data set.
3. Testing for redundancy of data in continuous manner.
4. Testing for missing values in our dataset.
5. Testing for different models being able to work on our dataset
6. Testing of dependency of the features on the total_sales to be predicted.

7.1.2 TESTING THE FUNCTIONS PARALLALLY

1. Allow all data set to be loaded and predictions to be performed on different models.
2. Allow the change in set.seed() value for random testing and training data.
3. Check the graphs for predictions and testing data. Maximizing the cross points for testing and prediction graph.
4. Testing the prediction of past sales with actual sales for last one or two years. Accuracy more than 90% achieved.

7.2 STRUCTURAL TESTING

Simply feeding all data together and running the analysis as well and check for predictions continuously and also changing the set.seed() value for randomness and check for any lag or program crash or any bug.

System should be able to function normally and all functions should work properly.

7.3 LEVELS OF TESTING

7.3.1 TESTING MALUALLY

All the OEM parts predictions were checked for correctness on past data and results were matched with the sales that occurred last year.

set.seed() value was changed at regular interval and accuracy achieved was 90% or more. It was done to check if it is working on different systems or not.

7.3.2 TESTING USING SIMULATOR

The application is build and installed onto a test target and the values of the car sales were simulated and sent using another backend Testing Software named Predict Future Simulator. And again checked for any lag jitter or crashes.

7.3.3 TESTING ON LIVE DATA

After being certain that the Application runs properly and all the bugs were removed, it is finally installed on target and tested there. The dummy data received from outside was modeled and checked again for predictions.
Accuracy more than 90 % was achieved.

8. IMPLEMENTATION

8.1 IMPLEMENTATION OF THE PROJECT

Strategic planning based on reliable forecasts is an essential key ingredient for a successful business management within a market-oriented company. This is especially true for the automobile industry. Reliable forecasts cannot only be based on intuitive economic guesses of the market development. The idea was to develop a program in R which gives the idea of sales of car in near future by utilizing the car sales data of previous two years. Once the prediction on car sales is performed, next goal was to predict the different OEM products sales in near future depending on a particular model sale prediction.

First step was to collect data regarding car sales of previous two years by considering each model of different companies. Since prediction of sales of automobile depends on various inner and exogenous parameter, the major challenge was to select the parameters/attributes that affect the sales dominantly. Also finding the various parameters/attributes that looks relevant but do not affect the sales dominantly was a major challenge. Next challenge was collection of data regarding the various OEM products of the organization used in different cars. Moreover mapping the OEM product sales with the prediction of car sales of a particular model was a difficult task. Choosing of different algorithm for the prediction and comparing and getting best results out of all the models also required special treatment.

The organization has a data set that identifies which of its OEM products are used in which car models. The challenge is to accurately predict future OEM sales requirement using predictive analytics and then to identify the demand of OEM products that may occur in near future on monthly, quarterly and yearly basis depending upon the sales that may be done for cars in near future.

8.2 DATA SET

A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity.

Data available with the organization is the list of car sales in previous two years and a list of OEM products present in various cars.

The mentioned dataset have columns given below:-

1. Make
2. Model
3. Month
4. Total Sales of the model

Further given dataset of car sales in last two years is modified by including various direct and exogenous parameters listed below:

5. GDP (Gross Domestic Product):- Since it is the measure of market value of all final goods and services produced in a period of time, inclusion of this parameter makes sense to have an idea for considering the effect of various financial ups and downs on the sales of car and subsequently on OEM sales of the organization.

6. Unemployment: - Unemployment rate is also included since it directly affects the car sales (country having lesser unemployment rate will observe high car sales)

7. Petrol and Diesel Price: - It looks relevant to include fuel rate in our existing dataset which may affect the car sales.

8.3 OEM DATA SET

A file having data regarding OEM products is also added to project which have information regarding different OEM products used in cars.

OEM products considered in our dataset:-

1. Diesel System
2. Breaking System
3. Auto Electricals
4. Batteries
5. Automotive Belt

6. Horns
7. Filters
8. Gasoline Systems
9. Auxiliary Lamps
10. Automotive Bulbs
11. Spark Plugs
12. Wiper Blades
13. Lightning Coils
14. Charging Coils
15. Ignition Coils

8.4 PROCEDURE

Installing various library needed in our predictive analysis

#install. Packages ("xlsxjars")

#install. Packages (“Random Forest”)

#install. Packages (“caTools”)

Including various library for out predictive analysis:-

Library (rJava)

Library (xlsxjars)

Library (xlsx)

Library (caTools)

Library (rpart)

Library (ggplot2)

Loading various dataset required in the project

my data=read.xlsx(file = "C:/Users/IPN1KOR/CarSalesALL.xlsx", sheetIndex = 1)

moncode=read.xlsx(file = "C:/Users/IPN1KOR/Months.xlsx", sheetIndex = 1)

parts=read.xlsx(file= "C:/Users/IPN1KOR/Parts.xlsx", sheetIndex = 1)

Splitting the given dataset into training and testing data as 70:30 ratio.

```
set.seed(2)
id=sample(2,nrow(data),prob  = c(0.7,0.3),replace=TRUE)
training_data=data[id==1,]
testing_data=data[id==2,]
```

Converting the Months column of the Car Sales data set into Month code by using formula

```
#formula to generate month code--
#12 + (12*(Year-17) + Monthcode)
x=strtoi(year)-17
moncode  = moncode[moncode$MON==mon,]
moncode
moncode$CODE
code=12 + (12*x) + moncode$CODE
#12+12*1+4=28
#mon
```

Reading the input from console form users:-

```
car= readline(prompt="Enter  ModelName :")
year= readline(prompt  = "Enter the year(YY fromat eg- 18 for 2018) :")
mon= readline(prompt  = "Enter the month(MMM  format eg- JAN for January) :")
#car
```

8.5 ALGORITHMS USED

1.Linear Regression

2.Logistics Regression

3. Decision Trees

4. Random Forest

5. Time Series (ARIMA)

8.6 REGRESSION

A technique for determining the statistical relationship between two or more variables where a change in a dependent variable is associated with, and depends on, a change in one or more independent variables.

8.6.1 LINEAR REGRESSION

Linear regression is a basic and commonly used type of predictive analysis.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.

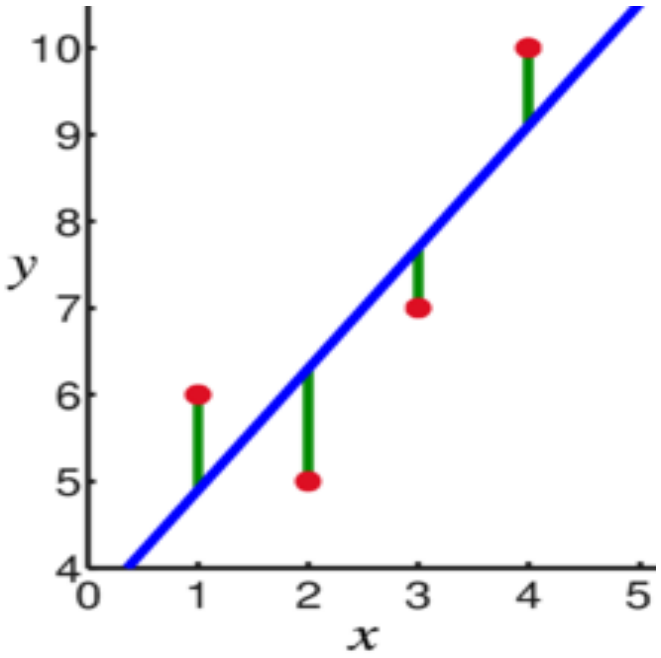


Figure 8: 8.6.1 LIREAR REGRESSION

A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable.

The slope of the line is b, and a is the intercept (the value of y when x = 0)

Model1<-lm(Sale~GDP+Unemployment+Petrol_price,data = train)

8.6.2 LOGISTIC REGRESSION

Logistic Regression is a basic and commonly used type of predictive analysis.

Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm).

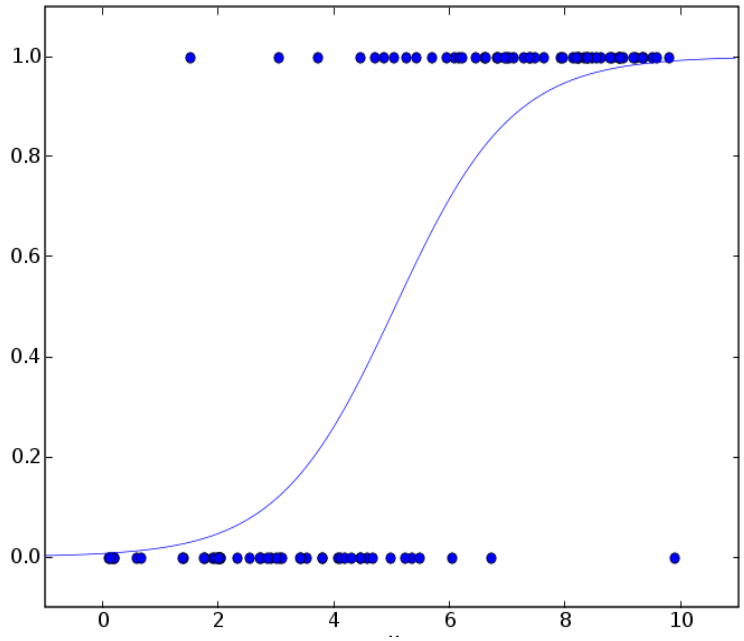


Figure 9: 8.6.2 LOGISTIC REGRESSION

The Fundamental Equation of generalized linear model is:

$g(E(y)) = \alpha + \beta x1 + \gamma x2$

Here, g() is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor (α,β,γ to be predicted).

```
Model2=glm(Total~.,data = training_data)
```

8.6.3 DECISION TREE

A decision tree is a tree in which each branch node represents a choice between a number of alternatives and each leaf node represents a decision.

It is a type of supervised learning algorithm (with a predefined target variable) that is mostly used in classification problems and works for both categorical and continuous input and output variables. It is one of the most widely used and practical methods for inductive inference.

```
tree<-ctree(Total~.,data=testing_data)
plot(tree,inner_panel=node_inner(tree,pval=F,id=F),terminal_panel=node_terminal(tree,id=F))
```

8.6.4 RANDOM FOREST

Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of overcoming over-fitting problem of individual decision tree.

In other words, random forests are an ensemble learning method for classification and regression that operate by constructing a lot of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

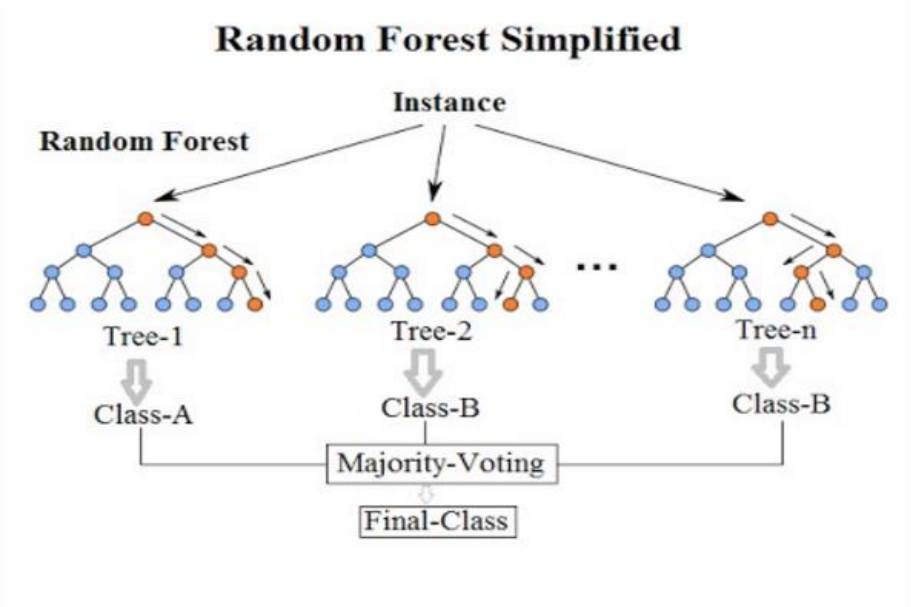


Figure 10: 8.6.4 RANDOM FOREST

Model4=randomForest(Total~.,data = training_data)

8.6.5 TIME SERIES (ARIMA)

Time' is the most important factor which ensures success in a business. It's difficult to keep up with the pace of time. I'm talking about the methods of prediction & forecasting. One such method, which deals with time based data is Time Series Modeling. As the name suggests, it involves working on time (years, days, hours, minutes) based data, to derive hidden insights to make informed decision making.

```
library(tseries)

fit<-arima(log(data.ts),c(1,0,1),seasonal = list(order=c(1,0,1),period=12))
```

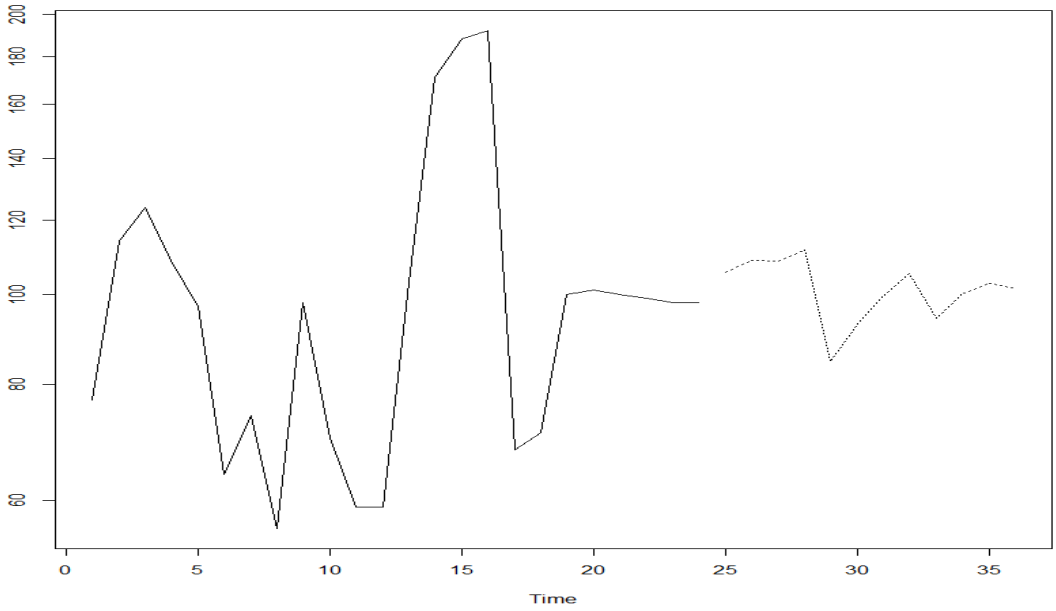


Figure 11: 8.6.2 GRAPH SHOWING SALES FORECAST WITH TIME

9. PROJECT LEGACY

9.1 CURRENT STATUS OF PROJECT AND REMAINING AREAS OF CONCERN

Currently all the functional requirements are fulfilled except the accuracy of our model is only 90-95%. The data is now being saved in the local database and we are waiting for the full access of the company's sales data so that we can train our model better and the results will be more accurate. This functionality is been tested though the 30% of the data that we have is used for testing.

Another function as to be added in the next phase that is showing the patterns, trends, seasonality in the sales with more visualized and optimized way so that it is easy to understand by Vendors and may help them keep track of sale and help them increase the sale and be aware of the upcoming ups and downs in the market.

9.2 TECHNICAL AND MANAGERIAL LESSONS LEARNT

Working in this project was a challenge as the Machine learning and its working was new to me and also a new field in the market as well. CRAN provides a package which can be found on it but everything needs to be understood reading the classes and no other documentation can be found on the web.

It was because of my Mentor and my team who helped me at every step that I was able to grab the knowledge as quick as possible. Now I know the architecture of Machine learning and how the data flows between different layers of the system.

I can now develop system by which one can easily predict and forecast the values based on the data they have which can help them get an idea how the trends will go on for the sales, purchase etc.

10. SYSTEM SNAPS

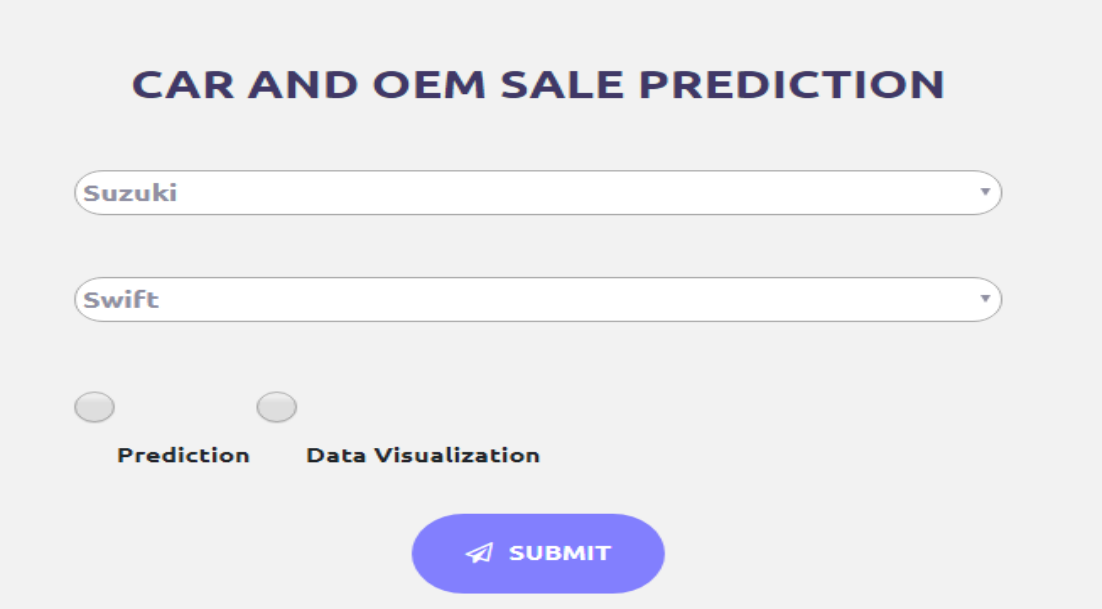


Figure 12: INTERFACE

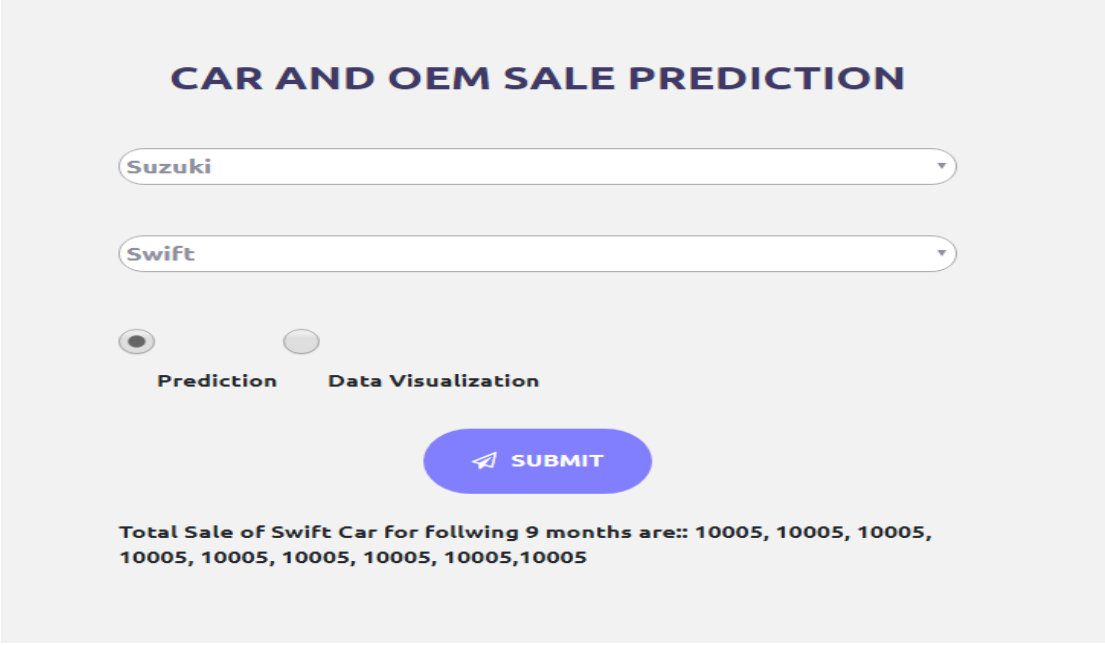


Figure 13: INTERFACE WITH OUTPUT

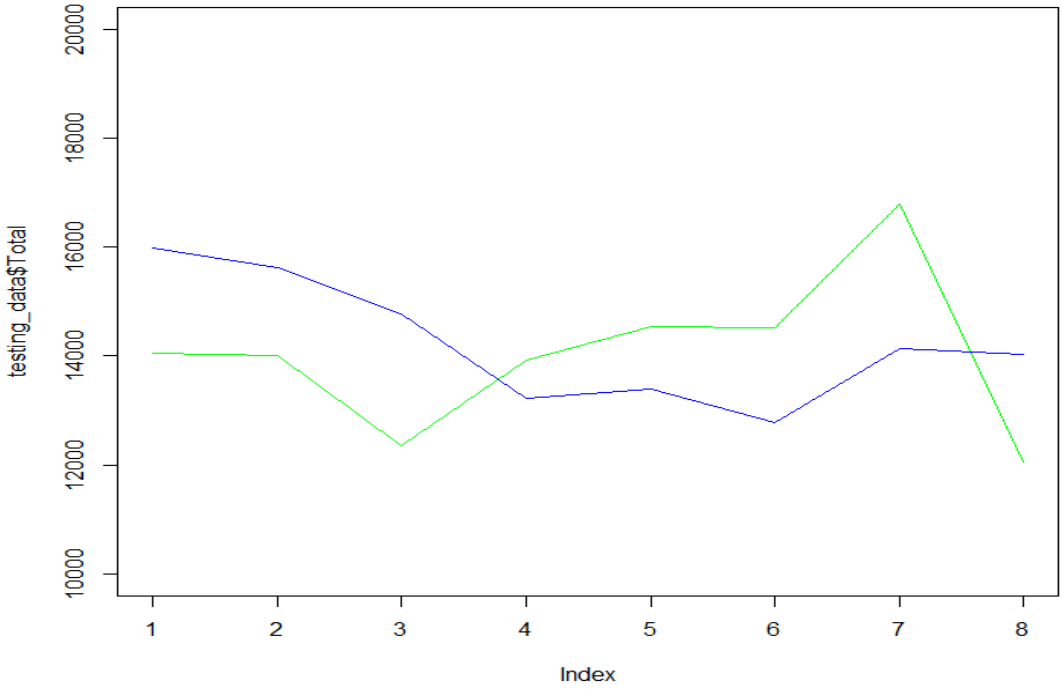


Figure 14: PREDITICTION USING LINEAR REGRESSION

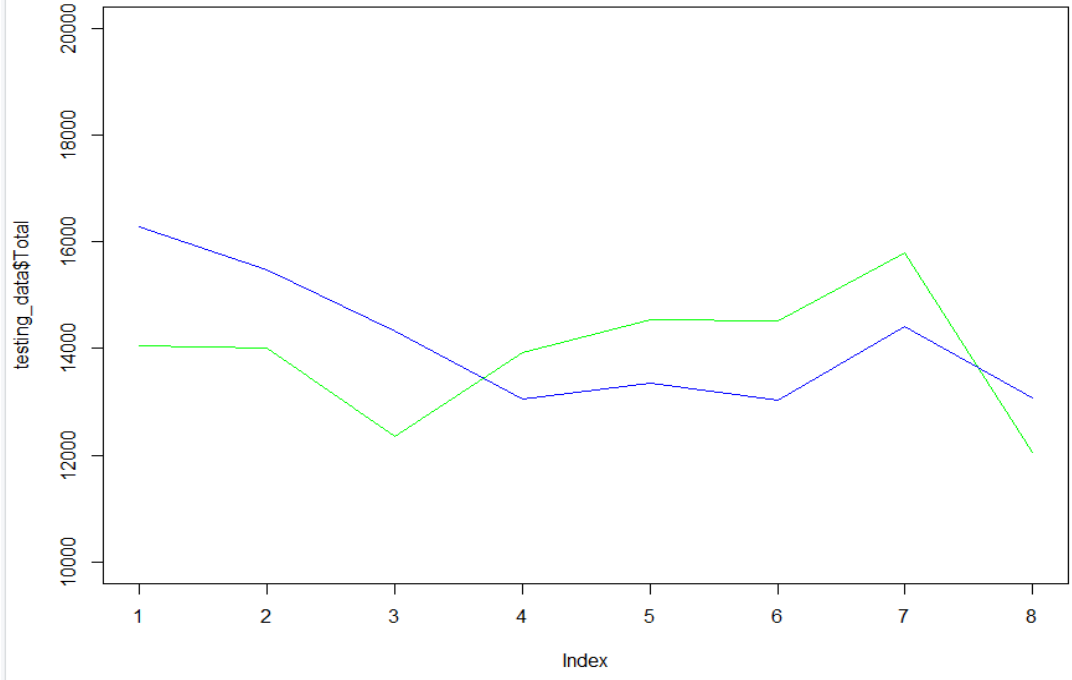


Figure 15: PREDICTION USING LOGISTIC REGRESSION

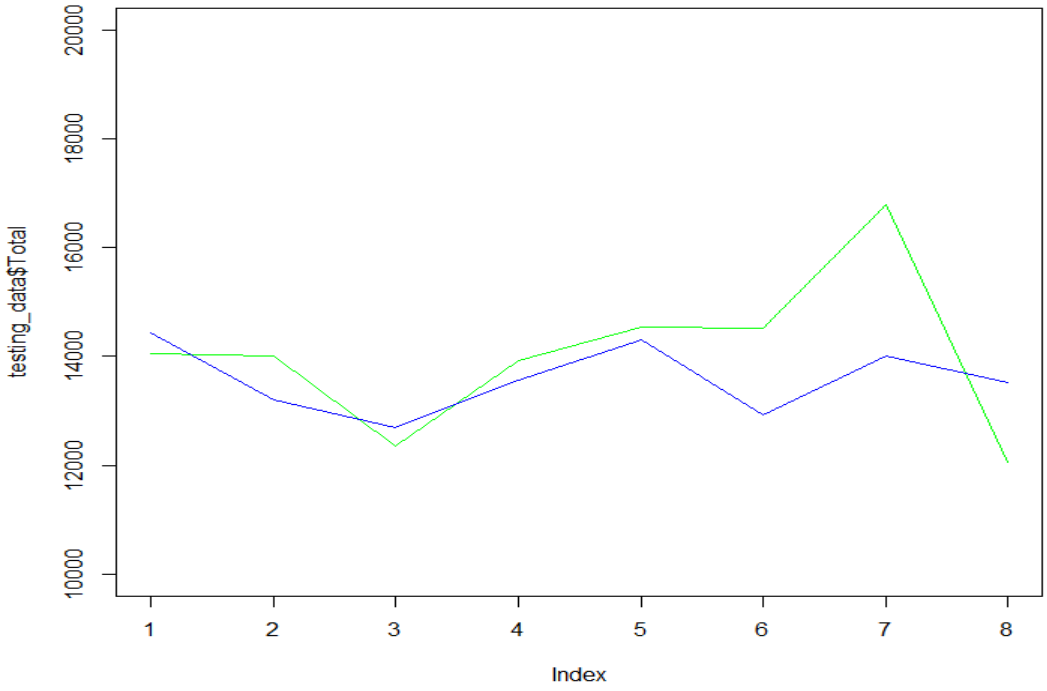


Figure 16: PREDICTION USING RANDOM FOREST

```
Swift requires :

  Diesel.System = 0
  Breaking.System = 1
  Auto.Electricals = 1
  Batteries = 0
  Automotive.Belt = 2
  Horns = 1
  Filters = 6
  Gasoline.Systems = 0
  Auxiliary.Lamps = 2
  Automotive.Bulbs = 9
  Spark.Plugs = 1
  Wiper.Blades = 2
  Lightning.Coils = 0
  Charging.Coils = 1
  Ignition.Coils = 1
```

Figure 17: BOSCH OEM USED IN SWIFT

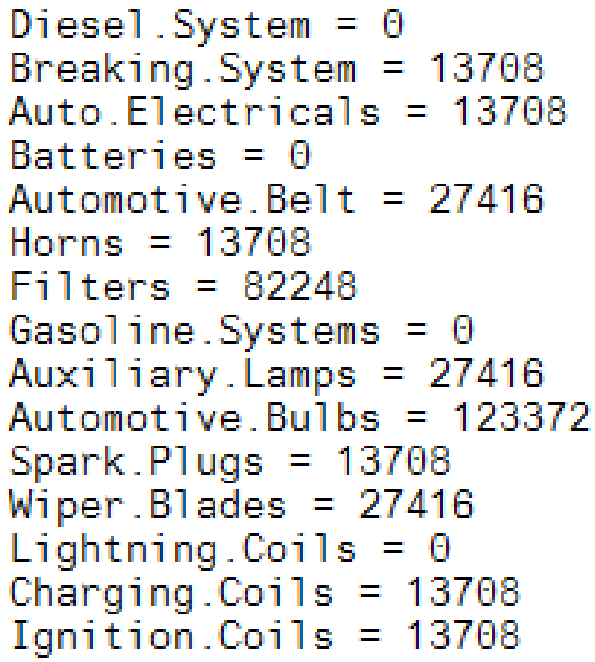


Figure 18: FORECAST OF BOSCH OEM USED IN SWIFT

10.1 VISUALIZATION OF DATASET

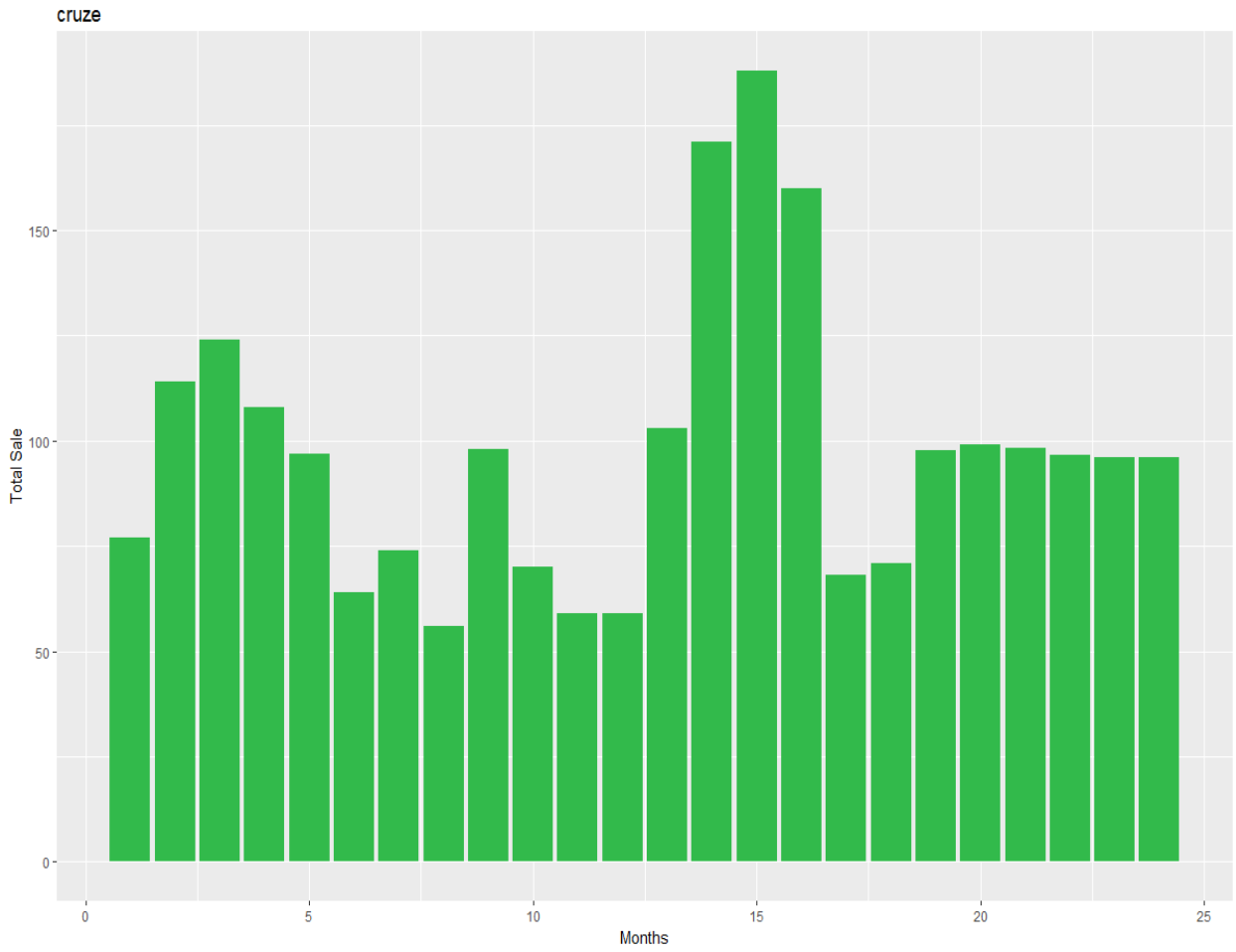


Figure 19: BAR GRAPH FOR TOTAL SALES AGAINST MONTHS

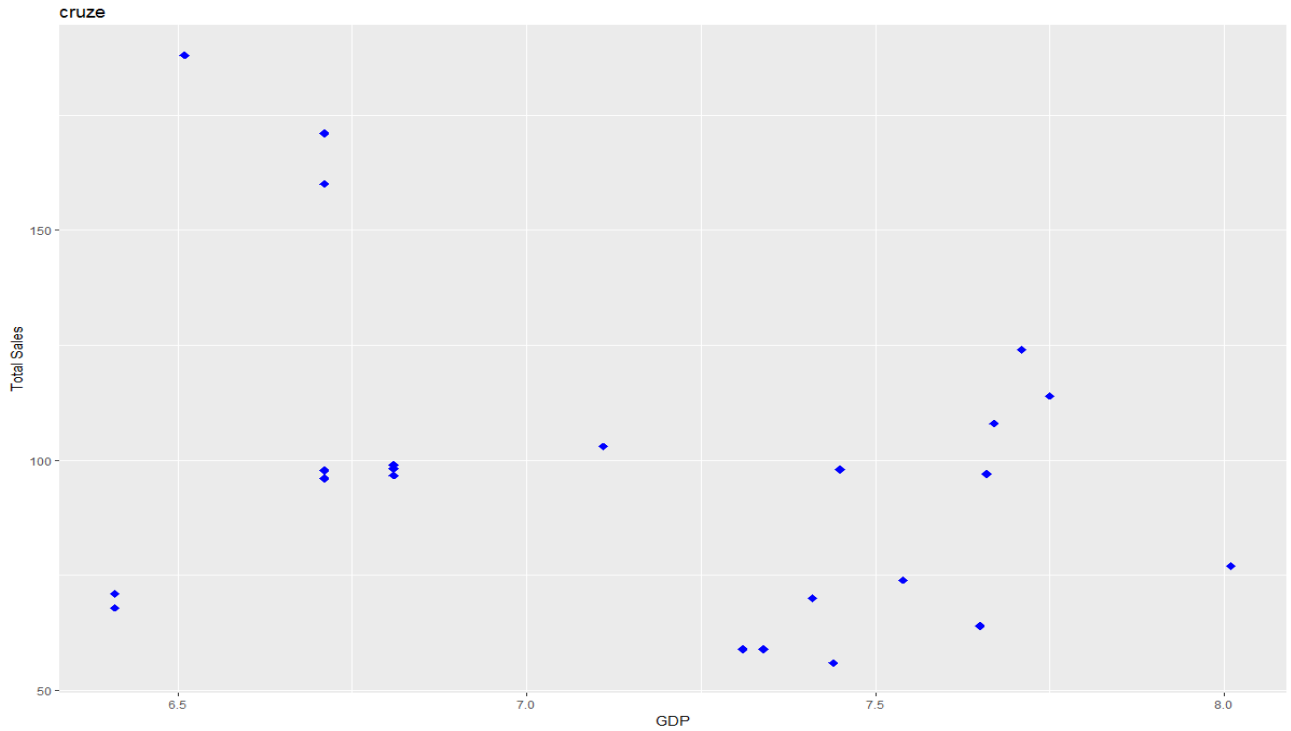


Figure 20: POINT GRAPH FOR TOTAL SALES AGAINST GDP

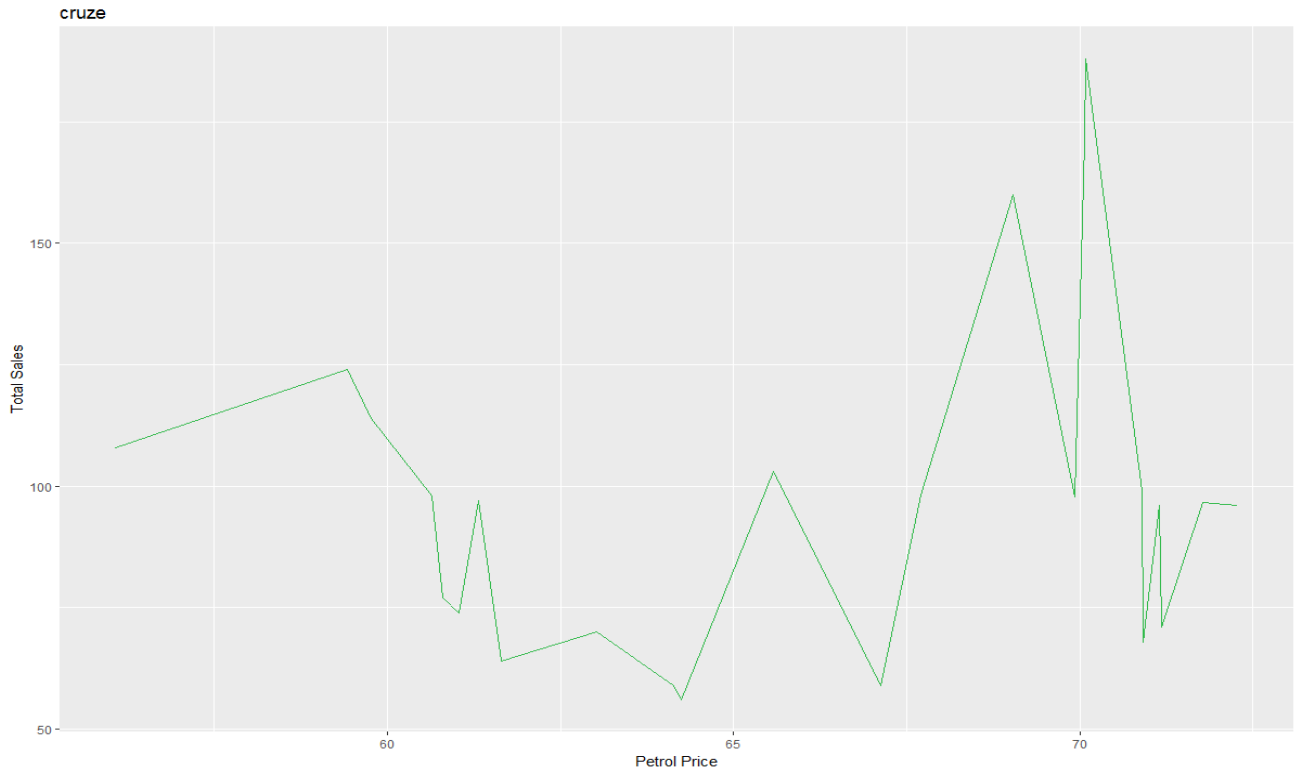


Figure 20: LINE GRAPH FOR TOTAL SALES AGAINST PETROL PRICE

11. BIBLIOGRAPHY

1. https://www.talentsprint.com/
2. https://www.theguardian.com/technology/2016/jun/28/google-says-machine-learning-is-the-future-so-i-tried-it-myself/
3. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
4. https://www.kdnuggets.com/2016/08/10-algorithms-machine-learning-engineers.html/
5. https://blog.statsbot.co/machine-learning-algorithms-183cc73197c
6. https://www.embitel.com/blog/embedded-blog/no-android-auto-is-not-an-infotainment-os-make-way-for-android-automotive
7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2845248/