Analysis and Prediction on price of Commodities and prevent Customer churn using Machine Learning techniques.

Sarthak Bhatnagar
Msc in Data Analytics
National College of Ireland
Dublin, Ireland
x21185352@student.ncirl.ie

***Abstract*— The main goal of this paper is to execute and implement 5 machine learning models on 3 datasets (or possibly related) whose target variable is imbalanced. Our main objective is to create the most accurate models which could predict and analyse the task given in hand. The problem statement for first dataset is to predict the sale price of houses for various "House type" in Melbourne, Australia. The dataset is obtained from Kaggle in the name of Melbourne house pricing. Since it's a classification problem we predicted the sale prices using Multinomial logistic and Naïve Bayes classification techniques. Another dataset is Secondhand car sale price prediction analysis, the car listings scrapped from autolist.com and these listings were filtered for cars of sale price between $5000 - $50,000 in the cities of Austin, San Francisco, and Los Angeles. To predict accurate sale prices Multlinear regression algorithm is used to fit the model. Lastly, the third dataset related to Customer churn problem, to predict the factors responsible for customer churn in telecom industry and develop programs for customer retention. The dataset is obtained from Kaggle and formatted in .csv file type. To understand the potential reasons behind customer churn Logistic regression and *** is used to fit the model which could increase customer retention. Keywords— Machine learning, Predictive modelling, Explanatory analysis, Linear regression, Logistic regression, Decision Tree, Multilinear regression and Naive-Bayes classification.**

I. INTRODUCTION

1. Telecom Customer Churn

In the telecom industry, Customer churn or attrition is the rate at which clients opted out for the products or services from the provided and customer churn analysis is the method to determine the rate of churning. Customers have lots of option to choose services from various other active providers. Customers easily switch from one operator to other if they feel dissatisfaction with the services provided. Ideally, the acceptance customer churn rate in telecom industry could be 5\% to 7\% annually which means that lose 1 out of 200 customers per month but due to rigid competition in the current market, the average churn rate in telecom businesses has upscaled from 15\% to 22\% in recent years. At the telecom industry level, the company suffer losses due to poor customer experience and low customer loyalty. In Telecommunication sector, it has been found that cost of acquiring a new customer is 8-10 times is much more expensive as compared to keeping an existing customer. One of the challenging factor for telecom industry is customer retention as customers are always triggered for better interest rates or benefits offered by the competitors. This makes logically correct that customer retention should be higher in priority than getting new ones. This could be a valid reason why companies paid special attention to highly profitable customers. To keep retaining the profitable customers, companies should analyse and predict the customers of high risk of churning.

2. Melbourne Housing Data

Melbourne is the top in the list of world's liveable city index by the Economist Intelligence Unit for past 6 years. It's also predicted that its population will exceed the population of Sydney by 2030. According to market study, Melbourne's property market is observed to be flat patched from last several years. However, the decline of Melbourne housing market is slowing down and slightly improved in the values has been observed after covid situation and thus the prices of houses been increasing drastically over a period of time than any other cities in Australia.

Melbourne real estate properties provide wide variety of accommodation options to people as it provides better amenities to people and it popular among international students as being a hub of world class universities. Being such nice place to live, Melbourne has observed very ups and downs in the real estate industry which is my prime motive to examine the variation in prices of house type in Melbourne. In this data we are implementing classification technique for house type in Melbourne and the factors associated with the house type and its price. We have several attributes such as price, number of bedrooms, bathrooms and location which could directly impact the house type classification.

3. Secondhand Car price prediction

The paper is introduced to support second-hand car owners' decision-making capabilities in the automobile industries. As per surveys, the growth of new car for next 5 years is average 3.5\% whereas the market for used car is 5\%. It's observed that the market for used car is increasing drastically as compared to new car buyers. As per the government law, to maintain the CO2 emission in the country the lifecycle of a car decided to be 15 years for light commercial vehicles.

\\The used cars can be purchased from auction, online website and directly from owners. Buying a new car needs good investment so people rather than buy a new car go for used cars which could offer less price and budget friendly. Resell car market has potential buyers to buy used cars which is almost new, less driven, and technically maintained. Some online platforms deal with selling/reselling cars which means act as a middleman between the seller and buyer such as Car27, CarDekho, etc. Our goal here is to use the machine learning techniques to predict the price of used cars based on

parameters like odometer, car_purchase_date, Engine, Transmission, etc which could make people life easier and prevent them from being duped. The data is csv format and source of the data is Kaggle. The research question for our analysis is:

"What factors causes the secondhand car prices and how to decide?"

Our task is to analyse and implement a machine which could takes up this large amount of data and provide better accuracy and performance in less amount of time. To implement such machine regression and classification techniques are used to implement machine learning prediction analysis.

II Related Work

[1] This article is focused on customer churn analysis in telecom industry using machine learning techniques and compare the models based on accuracy produced. The datasets taken is in csv format which is then processed by removing ambiguities, errors, and redundancy. Explanatory analysis has been performed on the factors using correlation matrix to showcase colinear attributes. They AUC-ROC curve to evaluate the performance of a model which could provide best performance in less time.[1]

[2] In this study, the author worked on the distinguished features of CRM software and segmented the potential customer characteristics by clustering techniques. Further, the result of customer segmentation is supplied to machine learning models to carry out the list of churned customers using backup data of customers from SAS Enterprise miner.[2]

[3] This study focused Customer churn analysis considering the important factor of NPS (Net Promoter Score) in a dataset for Malaysian telecom company and analysis has been done to indicate that any change in NPS triggers the changes in churn determinants or not. Various machine learning techniques applied but the most accurate result is provided by CART (Classification and Regression Trees) algorithm.[3]

[4] This paper discusses about the method to point out the relevant and specific data items which contributes to the analysis. Effectiveness of feature extraction and processing is explained to achieve better results using telecom dataset.

[5] The study is a comparative analysis of Churn prediction Models using machine learning techniques and deep learning algorithms besides the analysis author applied the procedure of Attribution Selection to determine the significant factors contributing to the churn. The result of analysis showed that Random Forest has supremacy over other machine learning algorithms followed by the deep learning models of Convolutional Neural Network and Multilayer Perceptron.

[6] P. Durganjali and M. V. Pujitha, uses classification algorithms to predict the price of resale houses and the value of house marketing mistreatment. The metrics used is here is the accuracy of algorithm and Ada boost provided the best accuracy and better results.

[7] The main purpose of the author in this paper is to predict the house prices using several machine learning techniques and the model accuracy is compared based on mean square error scores. The hypothesis is enhanced further by mixing hyperparameters as accuracy is the most important aspect in this research paper. Moreover, the weak classifiers are calculated in multiple iteration and in the end sinusoidal function is added to get the strongest classifiers for modelling,

[8] The author predicts the price of residential properties in United states by applying machine learning models on house prices and crime incident data. Model applications are carried out using neural networks mainly first feed neural network and cascade forward neural network. With the help of neural network authors can find out which variable with respect to probability is more contributed to assess the values of house price of residential properties.

[9] The primary aim of the author is to predict real estate price by clustering techniques and continuous improvement of generalized linear regression model which focuses on definition and the characteristics of stock market. Additional analysis has been done on nonparametric regression model and generalized linear model to give the estimation of partial linear models.

[10] The analysis is used to predict the price of the housed using strong classifiers of machine learning techniques and it was random forest performed better than any other classifiers like support vector machine and artificial neural networks.

[11] Supervised learning algorithms to estimate and predict the resale price of used cars is discussed in this paper for the online car selling ecommerce platform "Avito" based in Morocco. After Examining all the tested models Gradient boosting regressor is the one providing better accuracy metrics.

[12] This paper primarily focuses on weighted and mixed regression methods in Random Forest and XGBoost algorithms, the properties of weights are mixed to improve the robustness of the models and evaluation index created which has great effects on used car prediction.

[13] This research paper focuses on comprehensive study of Monthly car sales from the web based study or internet. The data is carried out from online car selling platform Mobi101, merged with the database of DGBAS to perform Genetic algorithm and several machine learning algorithms to carry out the analysis.

[14] The author is comparing the price of secondhand car based on muti variate attributes. The weak classifiers are transformed into strong one using XGBoost in which the accuracy of regular supervised models such as svm, ridge, lasso is. then compared with artificial neural networks algorithms

III Data Mining Methodology

To achieve the state of better performance and high accuracy KDD (Knowledge Discovery in Database) methodology is used-.

- Data selection.
- Data processing.
- Data transformation.
- Data Mining.
- Evaluation.
- Knowledge Discovery.

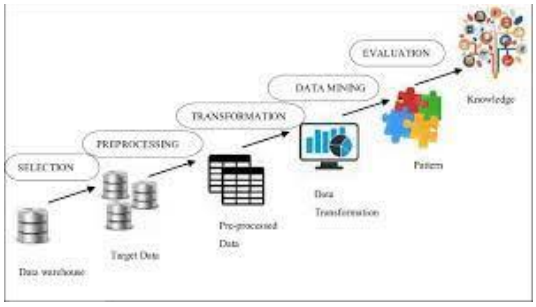


Fig. 1 KDD Process Methodology

A. *Data Selection*

1. Telecom Customer Churn

This dataset is selected from Kaggle.com and it provides information about the customers who opted out for the services. Each row in the dataset represents customer and each column depicts attributes associated to it. One of the major columns named" Churn" is the target column for the implementation of machine learning techniques. This dataset" Telecom Customer Churn" is in .csv format which consists of 7043 rows and 21 columns.

Attributes	Data Types
Customer_id	<factor>
gender	<factor>
SeniorCitizine	<integer>
Partner	<factor>
Dependents	<factor>
Method	<character>
tenure	<character>
PhoneService	<factor>
MultipleLines	<factor>
InternetService	<factor>
OnlineSecurity	<factor>
OnlineBackup	<factor>
StreamingMovies	<factor>
Churn	<factor>
TotalCharges	<double>
MonthlyCharges	<double>
PaymentMethod	<factor>
PaperlessBilling	<factor>
Contract	<factor>
StreamingTV	<factor>
DeviceProtection	<factor>
TechSupport	<factor>

Fig. 2 Attribute Table for Telco Customer churn.

2. Melbourne Housing Data

This dataset is collected from online platform Kaggle.com which provides information about House prices in Melbourne and the attributes responsible for price variation. This data is generated to capture the growth of prices in the city of Melbourne. I am going to apply multi class algorithm on housing type column to perform classification. This dataset" Melbourne Housing Data" is in .csv format which consists of 3 4857 rows and 21 columns.

Attributes	Data Types
Suburb	<character>
Address	<character>
Rooms	<integer>
Type	<character>
Price	<integer>
Method	<character>
SellerG	<character>
Date	<character>
Distance	<character>
Postcode	<character>
Bedroom2	<double>
Bathroom	<integer>
Car	<integer>
Landsize	<integer>
BuildingArea	<double>
YearBuilt	<integer>
CouncilArea	<character>
Lattitude	<double>
Longtitude	<double>
Regionname	<character>
Propertycount	<character>

Fig. 3 Attribute Table for Melbourne Housing Data.

3. Secondhand Car price prediction

This dataset is collected from Kaggle.com which provides information about secondhand car data prices and the attributes of car on which price is dependent. This dataset" Secondhand Car price prediction" is in .csv format which consists of 38506 rows and 18 columns.

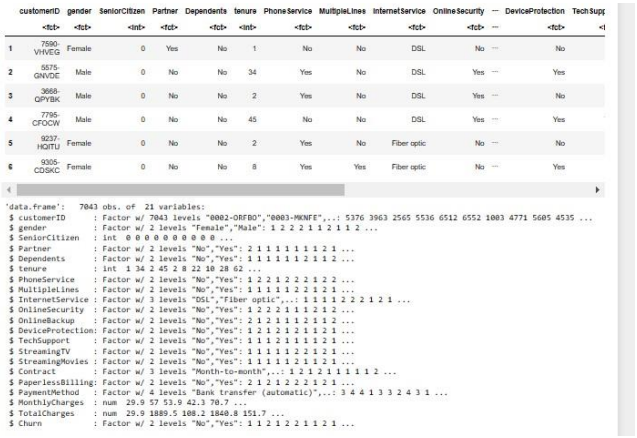
Attributes	Data Types
Index	<integer>
price	<integer>
acquisition_date	<factor>
badge	<factor>
body_type	<factor>
category	<factor>
colour	<factor>
cylinders	<integer>
economy	<double>
fuel	<factor>
last_updated	<factor>
litres	<double>
location	<integer>
make	<factor>
model	<factor>
odometer	<integer>
transmission	<factor>
year	<integer>

Fig. 4 Attribute Table for Secondhand car price prediction.

B. Data Preprocessing and Transformation

1. Telecom Customer Churn

The dataset used for Telecom Customer Churn is in .csv format which is read using read() command and then first 5 rows are showed using head() command and in next step data type is shown using str() command.



	customerID	gender	seniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	…	DeviceProtection	TechSupp
	<fct>	<fct>	<int>	<fct>	<fct>	<int>	<fct>	<fct>	<fct>	<fct>	…	<fct>	<fct>
1	7590-VHVEG	Female	0	Yes	No	1	No	No	DSL	No	…	No	
2	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	…	Yes	
3	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	…	No	
4	7795-CFOCW	Male	0	No	No	45	No	No	DSL	Yes	…	Yes	
5	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	…	No	
6	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	…	Yes	

```
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",...: 5376 3963 2565 5536 6512 6552 1003 4771 5605 4535 ...
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 2 2 2 ...
 $ MultipleLines   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",...: 1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 2 1 2 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 2 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 2 1 2 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 2 1 2 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 2 1 2 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 2 1 2 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

Fig 5. Read csv file for Telecom Customer Churn

II. Checking for Null Values and Outlier Analysis.

In the next step, NA values are checked in the complete dataset, and it was found that Total Charges have 11 NA values. Moreover, Outlier analysis has been performed on numeric attributes to check whether the data outer lying values if we encounter these values, we will remove or replace them with mean values. The outlier analysis is applying on tenure, MonthlyCharges and TotalCharges column. Moreover, Outlier analysis graphs are shown below.

```
### Checking the presence of Null Values.###
sapply(customer_churn, function(x) sum(is.na(x)))
sum(is.na(customer_churn$TotalCharges)) ###Null values in TotalCharges column###
```

```
customerID: 0 gender: 0 SeniorCitizen: 0 Partner: 0 Dependents: 0 tenure: 0 PhoneService: 0 MultipleLines: 0 InternetService: 0 Online Security: 0
OnlineBackup: 0 DeviceProtection: 0 TechSupport: 0 StreamingTV: 0 StreamingMovies: 0 Contract: 0 PaperlessBilling: 0 PaymentMethod: 0
MonthlyCharges: 0 TotalCharges: 11 Churn: 0

11
```

Fig. 6 Code for reading NA values.

```
'data.frame':   7032 obs. of  20 variables:
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 2 2 2 ...
 $ MultipleLines   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 2 1 2 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 2 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 2 1 2 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 2 1 2 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 2 1 2 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 2 1 2 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

Fig.7 describes attributes after replacing null Values.

Fig. 8 Outlier plot for tenure.

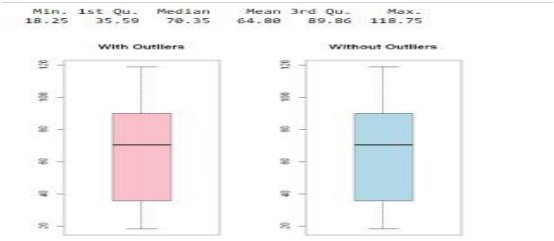


Fig. 9 Outlier plot for Monthly Charges.

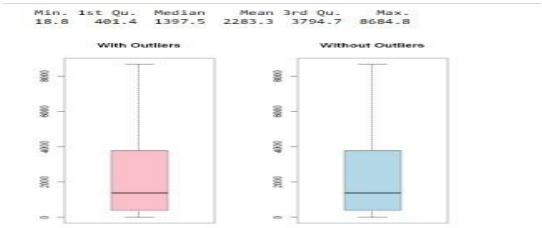


Fig. 10 Outlier plot for Total Charges.

After checking the null values outlier analysis is performed for the numerical data type attributes for columns. Monthly charges column has been removed as it is collinear with Total Charges which can be seen in correlation matrix and plot in next step. Plotting the histogram of tenure and Total Charges.

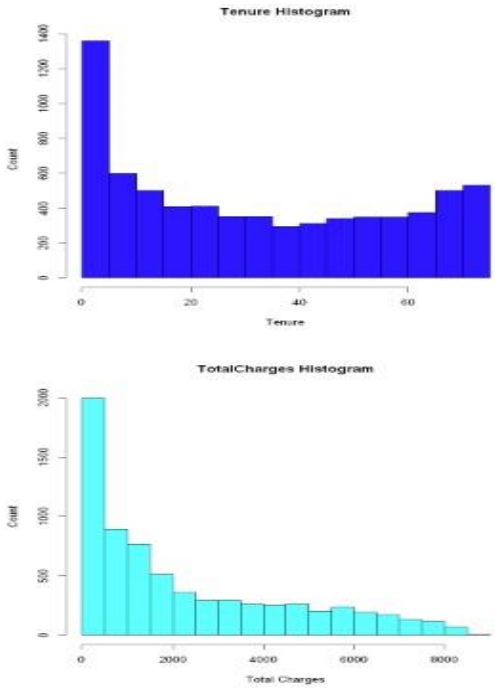


Fig. 11 Histogram for Tenure and Total Charges.

III. Plotting the Correlation Matrix.

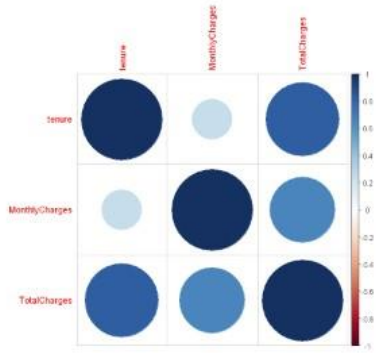


Fig. 12 Correlation Plot.

A matrix: 3 × 3 of type dbl

	tenure	MonthlyCharges	TotalCharges
tenure	1.0000000	0.2468618	0.8258805
MonthlyCharges	0.2468618	1.0000000	0.6510648
TotalCharges	0.8258805	0.6510648	1.0000000

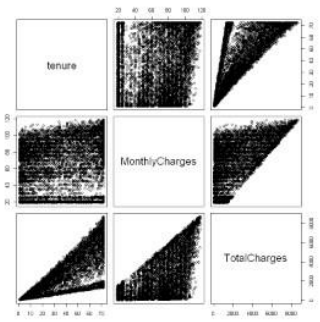


Fig. 13 Correlation Plot and Matrix.

IV. Encoding of Attributes and Data Exploration .

The main reason for encoding the data as the data values consist of either "YES" or "NO" type of values which is transformed to "!" or "0".The encoding of the data is applied to convert the factor data attribute to numeric data type which is required to be best for the machine learning model. This encoding is done using hot encoding methods in which factors are first converted into characters and later to numeric type.

```
str(churn_no_na)
```

```
'data.frame':   7032 obs. of  19 variables:
 $ gender          : num  0 1 1 1 0 0 1 0 0 1 ...
 $ SeniorCitizen   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : num  1 0 0 0 0 0 0 0 1 0 ...
 $ Dependents      : num  0 0 0 0 0 0 1 0 0 1 ...
 $ tenure          : num  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : num  0 1 1 0 1 1 1 0 1 1 ...
 $ MultipleLines   : num  0 0 0 0 0 1 1 0 1 0 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity  : num  0 1 1 1 0 0 0 1 0 1 ...
 $ OnlineBackup    : num  1 0 1 0 0 0 1 0 0 1 ...
 $ DeviceProtection: num  0 1 0 1 0 1 0 0 1 0 ...
 $ TechSupport     : num  0 0 0 1 0 0 0 0 1 0 ...
 $ StreamingTV     : num  0 0 0 0 0 1 1 0 1 0 ...
 $ StreamingMovies : num  0 0 0 0 0 1 0 0 1 0 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: num  1 0 1 0 1 1 1 0 1 0 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : num  0 0 1 0 1 1 0 0 1 0 ...
```

Fig. 14 Data Values after hot Encoding.

Exploratory Analysis is carried out on the data to check which attribute has effect on the churn column which could help us to check the dependencies on Churn attribute. Plotting Graph for attributes.

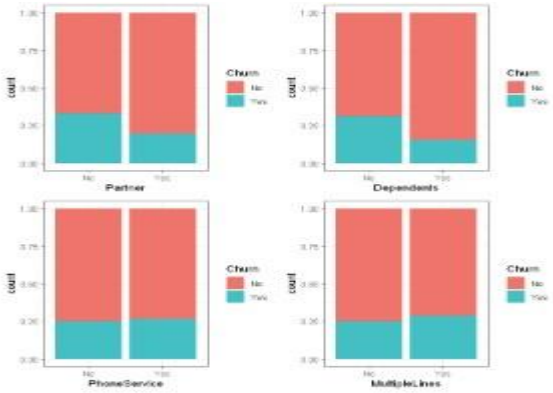


Fig. 15 Churn vs other column attributes.

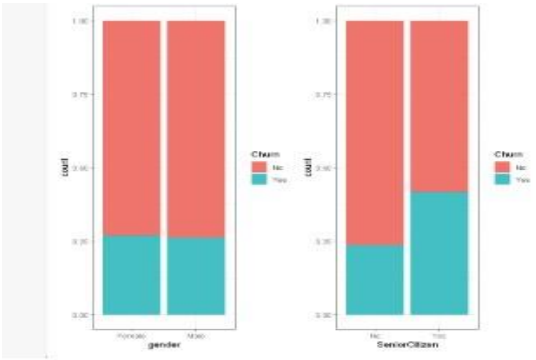


Fig. 16 Churn vs other column attributes.

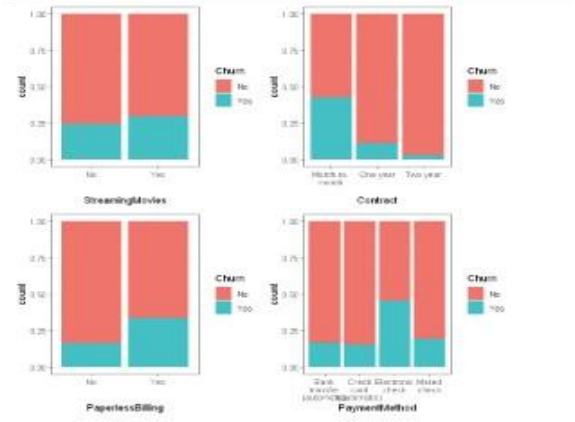


Fig. 17 Churn vs other column attributes.

From the above explanatory analysis, we can conclude that PhoneService, Multiplelines and gender has no effect on churn. The SeniorCitizen customers have higher churn rate. Moreover, for customers who are dependent and have partners have lower churn rate.

V. Final dataset is ready for applying Machine Learning techniques.

2. Melbourne Housing Data

We first read the dataset and display it along with data types.

```
'data.frame':   34857 obs. of  21 variables:
 $ Suburb       : chr  "Abbotsford" "Abbotsford" "Abbotsford" ...
 $ Address      : chr  "68 Studley St" "85 Turner St" "25 Bloomburg St" "18/659 Victoria St" ...
 $ Rooms        : int  2 2 2 3 3 3 4 4 2 2 ...
 $ Type         : chr  "h" "h" "h" "u" ...
 $ Price        : int  NA 1480000 1035000 NA 1465000 850000 1600000 NA NA NA ...
 $ Method       : chr  "SS" "S" "S" "SP" ...
 $ SellerG      : chr  "Jellis" "Biggin" "Biggin" "Rounds" ...
 $ Date         : chr  "3/09/2016" "3/12/2016" "4/02/2016" "4/02/2016" ...
 $ Distance     : chr  "2.5" "2.5" "2.5" "2.5" ...
 $ Postcode     : chr  "3067" "3067" "3067" "3067" ...
 $ Bedroom2     : int  2 2 2 3 3 3 3 3 4 3 ...
 $ Bathroom     : int  1 1 1 2 2 2 1 2 1 2 ...
 $ Car          : int  1 1 0 1 0 1 2 2 2 1 ...
 $ Landsize     : int  126 202 156 0 134 94 120 400 201 202 ...
 $ BuildingArea : num  NA NA 79 NA 150 NA 142 220 NA NA ...
 $ YearBuilt    : int  NA NA 1900 NA 1900 NA 2014 2006 1900 1900 ...
 $ CouncilArea  : chr  "Yarra City Council" "Yarra City Council" "Yarra City Council" "Yarra Ci
 $ Lattitude    : num  -37.8 -37.8 -37.8 -37.8 -37.8 ...
 $ Longtitude   : num  145 145 145 145 145 ...
 $ Regionname   : chr  "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan"
 $ Propertycount: chr  "4019" "4019" "4019" "4019" ...
```

Fig 18. Read csv file for Melbourne Housing Data,

II. Checking for Null Values and Outlier Analysis.

There is only one column which has 11 NA values which is removed using na.omit() function. Outliers for numerical data is checked and plotted on graph.

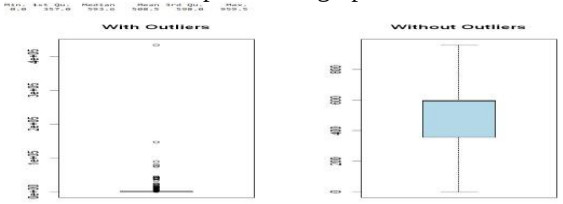


Fig 19. Outlier plot for landsize.

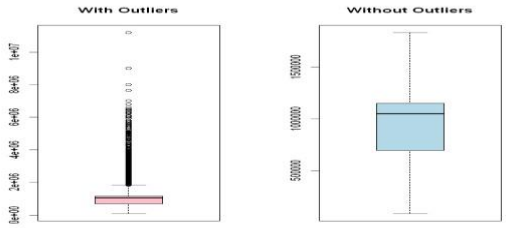


Fig 20. Outlier plot for price.

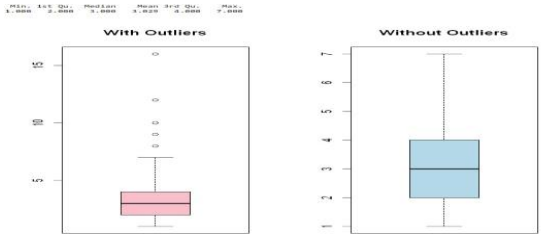


Fig 21. Outlier plot for Room.

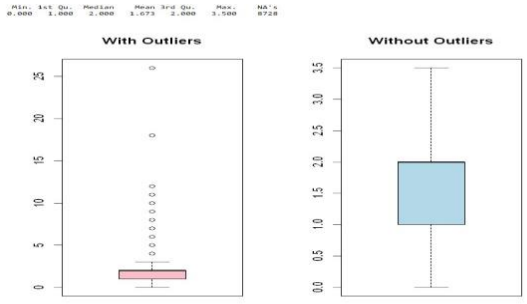


Fig 22. Outlier plot for Car.

III. Plotting the Correlation Matrix.

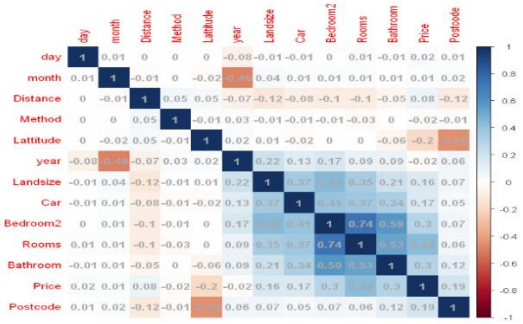


Fig. 23 Correlation Plot and Matrix

IV. Encoding of Attributes and Data Exploration .

In the hot encoding process in this dataset, we have transformed Method attribute from factors to numeric data type for better results.

```
'data.frame':   34857 obs. of  16 variables:
 $ day          : num  3 3 4 4 4 4 6 6 6 ...
 $ month        : num  9 12 2 2 3 3 6 8 8 8 ...
 $ year         : num  2016 2016 2016 2016 2017 ...
 $ Rooms        : num  2 2 2 3 3 3 4 4 2 2 ...
 $ Price        : num  0.1312 1.2258 0.0925 0.1312 1.1876 ...
 $ Method       : num  7 3 3 8 6 1 8 5 3 3 ...
 $ Distance     : num  -0.367 -0.367 -0.367 -0.367 -0.367 ...
 $ Postcode     : num  55 55 55 55 55 55 55 55 55 55 ...
 $ Bedroom2     : num  2.87 2.87 2.87 3 3 ...
 $ Bathroom     : num  1 1 1 2 2 2 1 2 1 2 ...
 $ Car          : num  1 1 0 1 0 1 2 2 2 1 ...
 $ Landsize     : num  -1.62 -1.3 -1.49 -2.15 -1.59 ...
 $ Lattitude    : num  -37.8 -37.8 -37.8 -37.8 -37.8 ...
 $ Longtitude   : num  145 145 145 145 145 ...
 $ Propertycount: num  0.127 0.127 0.127 0.127 0.127 ...
 $ Type         : num  1 1 1 2 1 1 1 1 1 1 ...
```

Fig. 24 Data Values after hot Encoding.

In the explanatory analysis, we plotted pie graph for House type (Target column) and houses sold per year.

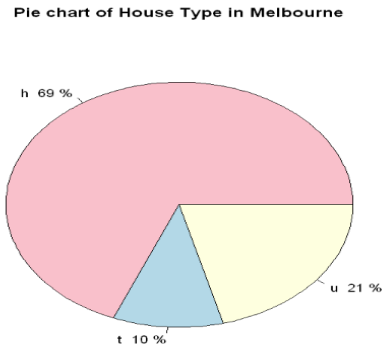


Fig. 25 Percentage of House type in Melbourne.

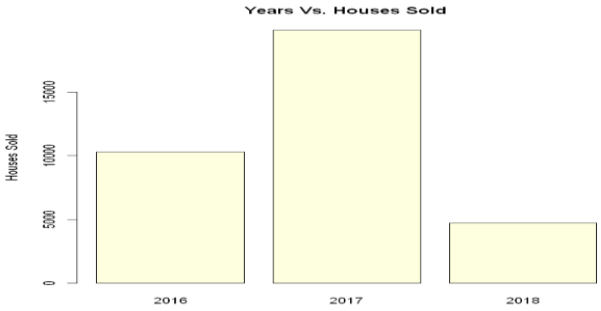


Fig. 25 Number of Houses sold in Melbourne year wise.

V. Final dataset is ready for applying Machine Learning techniques.

3. Secondhand Car Price Prediction.

We first read the dataset and display it along with data types.

```
'data.frame':   38506 obs. of  18 variables:
 $ Index           : int  2 3 4 5 6 8 9 10 11 12 ...
 $ price           : int  8560 17074 8526 10952 33964 18070 12019 10860 12997 30216 ...
 $ acquisition_date: Factor w/ 106 levels "01-02-2018","01-03-2018",..: 74 74 74 74 74 74 74 74 74 74 ...
 $ badge           : Factor w/ 205 levels "","(4x4)","(No Badge)",..: 104 11 98 98 155 16 104 98 104 25 ...
 $ body_type       : Factor w/ 9 levels "","Conv","Convertible",..: 5 5 5 5 6 5 6 6 6 5 ...
 $ category        : Factor w/ 6 levels "","Demo","Other",..: 6 6 6 6 6 6 6 5 6 3 ...
 $ colour          : Factor w/ 31 levels "","/cloth","Beige",..: 28 28 5 4 17 4 17 26 28 26 ...
 $ cylinders       : int  4 4 4 4 4 4 4 4 4 4 ...
 $ economy         : num  8.9 6.8 8.9 8.8 10.5 6.8 8.9 8.8 8.9 6.6 ...
 $ fuel            : Factor w/ 3 levels "","Diesel","Unleaded": 3 3 3 3 3 3 3 3 3 3 ...
 $ last_updated    : Factor w/ 104 levels "01-02-2018","01-03-2018",..: 73 73 73 73 73 73 73 73 73 73 ...
 $ litres          : num  2 2 2 2 2.5 2 2 2 2 2 ...
 $ location        : int  2 3 6 8 3 7 3 2 2 1 ...
 $ make            : Factor w/ 2 levels "Subaru","Toyota": 1 1 1 1 1 1 1 1 1 1 ...
 $ model           : Factor w/ 3 levels "Forester","Impreza",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ odometer        : int  134944 33304 81668 48051 51516 60294 96160 103380 88631 213 ...
 $ transmission    : Factor w/ 3 levels "","Automatic",..: 3 2 3 2 3 2 3 2 3 2 ...
 $ year            : int  2009 2014 2007 2009 2011 2012 2009 2011 2008 2017 ...

38506 · 18
```

Fig. 26 Read csv file for Secondhand Car Price Prediction.

II. Checking for NA values.

The null values in the dataset are replaced by mean of the column and Index column is removed being unnecessary.

```
#### Removing unnecessary columns from dataset such as 'index'.
cols_to_be_removed=c("Index")
car_data = car_data[,!(names(car_data) %in% cols_to_be_removed) ]

#### counting NAs by each columns ####
sapply(car_data, function(x) sum(is.na(x)))
```

price: 3 **acquisition_date:** 0 **badge:** 0 **body_type:** 0 **category:** 0 **colour:** 0 **cylinders:** 2488 **economy:** 3920 **fuel:** 0 **last_updated:** 0 **litres:** 2488 **location:** 0 **make:** 0 **model:** 0 **odometer:** 1550 **transmission:** 0 **year:** 0

Fig. 27 Checking for NA values.

III. Plotting the Correlation Matrix.

Fig. 28 Correlation Plot and Matrix.

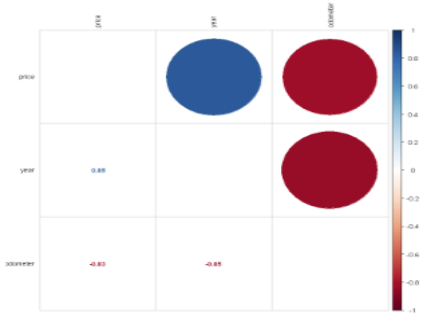


Fig. 29 Correlation Plot.

From the above correlation matrix, it can be deduced that year and odometer is highly co-related. Moreover, Actual age of car is calculated using Difftime() function and converted to numeric data type and avg. yearly usage column is created to check how much has actually driven in a year.\

IV. Encoding of Attributes and Data Exploration.

The acquisition_date and last_updated are common columns with same values. So, one of them is deleted.

Explanatory analysis.

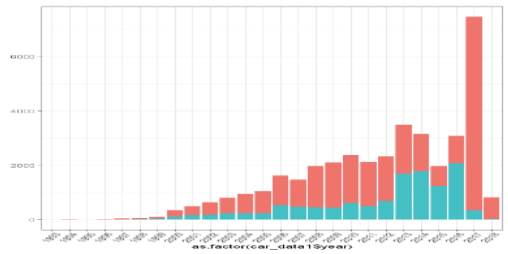


Fig. 30 Car "make" year wise.

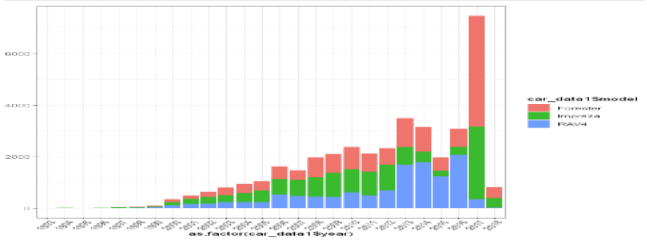


Fig. 31 Car "Model" year wise.

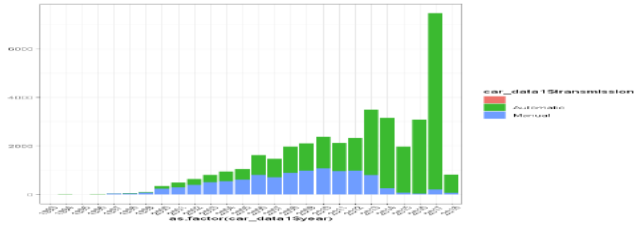


Fig. 32 Car "Transmission type" year wise

```
'data.frame':   33062 obs. of  20 variables:
 $ price           : int  8560 17074 8526 10952 33964 18070 12019 10860 12997 30216 ...
 $ acquisition_date: Factor w/ 106 levels "01-02-2018","01-03-2018",..: 74 74 74 74 74 74 74 74
 $ badge           : Factor w/ 205 levels "","(4x4)","(No Badge)",..: 104 11 98 98 155 16 104 9
 $ body_type       : Factor w/ 9 levels "","Conv","Convertible",..: 5 5 5 5 6 5 6 6 6 5 ...
 $ category        : Factor w/ 6 levels "","Demo","Other",..: 6 6 6 6 6 6 6 5 6 3 ...
 $ colour          : Factor w/ 31 levels "","/cloth","Beige",..: 28 28 5 4 17 4 17 26 28 26 ..
 $ cylinders       : int  4 4 4 4 4 4 4 4 4 4 ...
 $ economy         : num  8.9 6.8 8.9 8.8 10.5 6.8 8.9 8.8 8.9 6.6 ...
 $ fuel            : Factor w/ 3 levels "","Diesel","Unleaded": 3 3 3 3 3 3 3 3 3 3 ...
 $ last_updated    : Factor w/ 104 levels "01-02-2018","01-03-2018",..: 73 73 73 73 73 73 73 7
 $ litres          : num  2 2 2 2 2.5 2 2 2 2 2 ...
 $ location        : int  2 3 6 8 3 7 3 2 2 1 ...
 $ make            : Factor w/ 2 levels "Subaru","Toyota": 1 1 1 1 1 1 1 1 1 1 ...
 $ model           : Factor w/ 3 levels "Forester","Impreza",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ odometer        : num  134944 33304 81668 48051 51516 ...
 $ transmission    : Factor w/ 3 levels "","Automatic",..: 3 2 3 2 3 2 3 2 3 2 ...
 $ year            : int  2009 2014 2007 2009 2011 2012 2009 2011 2008 2017 ...
 $ year1           : Factor w/ 18 levels "03-09-2001","03-09-2002",..: 9 14 7 9 11 12 9 11 8 1
 $ age             : num  7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 ...
 $ avg_yearly_usage: num  17993 4441 10889 6407 6869 ...
 - attr(*, "na.action")= 'omit' Named int [1:5434] 59 88 96 100 109 114 145 157 166 185 ...
  ..- attr(*, "names")= chr [1:5434] "59" "88" "96" "100" ...
```

Fig. 33 Data Values after hot Encoding.

V. Final dataset is ready for applying Machine Learning techniques.

After plotting the graphs and columns created, we finally ready to apply machine learning techniques on our processed dataset.

C. ***Train, Test data ad Model Building.***

1. Telecom Customer Churn.

The dataset is splitted into train and test categories. Train consist of 30% of the dataset and test contain remaining 70%. We used caTool library to segregate them up and implemented accoerdingly,

We have implemented three algorithms.

a. Logistic Regression-

Logistic Regression is a statistical way to predict binary data i.e., “0” or “1”. It predicts relationship between one binary dependent variable and ordinal, nominal and cardinal independent variables.

$$\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

Fig 34. Formula for Logistic Regression.

In our logistic regression model, the target/dependent variable is Churn column which has values “0” and “1” and all independent columns was either “Yes” or “No” are converted to binary numbers except InternetService column which is remains factor type.

b. Decision Tree-

Decision tree is supervised learning algorithm. The name of this algorithm itself suggests that it work like a flowchart in which tree show predictions which results from feature-based splits. Decision tree can be used both in regression and classification. It can be noted that its start with root node and decision made by leaves.

Dependent Variable- Churn Attribute.
Independent Variable – Other Column Attributes.

c. Support Vector Machine Regression-

Support Vector Machine algorithm cab used both in classification and regression. It is similar to linear model but linear model we try to minimize the error where as in SVM we tries to fit the best line within the distance between hyperlane and boundary line.

Dependent Variable- Churn Attribute.
Independent Variable – Other Column Attributes.

2. Melbourne Housing Data-

In housing dataset, the data is splitted into training and testing. Training being 30% and testing rest of it.
We have implemented Multinomial Logistic classification and Naïve Bayes Classification model.

a. Multinomial Logistic Regression.

Multinomial Logistic regression is like logistic regression technique but the difference the target variable can have more than 2 classes.

Dependent Variable = Type Attribute

Independent Variable = Other remaining Attribute.

b. Naïve Bayes.

Naïve Bayes Classifier is a probabilistic machine learning algorithm which assumes that the presence of a particular attribute in a class is not related with any other feature. Thus, often used for classification task. In this algorithm “Type” attribute is used as a dependent variable and other as independent variables

Both the algorithms applied are producing good accuracy.

3. Secondhand Car Price Prediction.

a. Linear regression model-

LM Model is basic regression model with the idea that to obtain a best line that fits the data in which the best line has small total prediction error, In this case, our dependent variable is price and independent variables are other variables.

Formula = Y(pred) = b0 + b1*x

D. Evaluation.

1. Telecom Customer churn-

In this dataset, we implemented 3 different machine learning models keeping the same dependent variable “Churn” and all other attributes as independent column. We build the models and get the accuracy of the model as expected [1].

a. Logistic Regression-

For Customer churn analysis, we used logistic regression model in which we did prediction on customers which are likely to be churned. So, that we can make strategies to retain the existing customers keeping “Churn” attribute as dependent variable and others as independent. The model accuracy is giving accuracy of 79,35% which is average.

```
model_lrm <- glm(Churn~. , family="binomial", data = train)
summary(model_lrm)
```

```
Call:
glm(formula = Churn ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9801  -0.6574  -0.2791   0.7160   3.3933

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    3.264e-02  2.094e-01   0.156  0.87613
gender                         2.953e-02  7.812e-02   0.378  0.70546
SeniorCitizen                  2.472e-01  1.010e-01   2.447  0.01439 *
Partner                       -1.984e-02  9.363e-02  -0.212  0.83219
Dependents                    -1.098e-01  1.081e-01  -1.016  0.30973
tenure                        -5.662e-02  7.500e-03  -7.550 4.37e-14 ***
PhoneService                  -5.065e-01  1.621e-01  -3.124  0.00178 **
MultipleLines                  2.520e-01  9.588e-02   2.629  0.00857 **
InternetServiceFiber optic     7.732e-01  1.182e-01   6.541 6.12e-11 ***
```

Fig. 35 Logistic regression Model

```
           Predictedvalue
Actualvalue FALSE TRUE
          0  1379  179
          1   257  296
```

```
paste("Accuracy:", round((sum(diag(tab))/sum(tab))*100,2))
```

```
'Accuracy: 79.35'
```

b. Decision Tree and Support Vector Machine.

Keeping all the parameters same as above both the decision tree and SVM has worst accuracy for the churn analysis which is 44.2%.

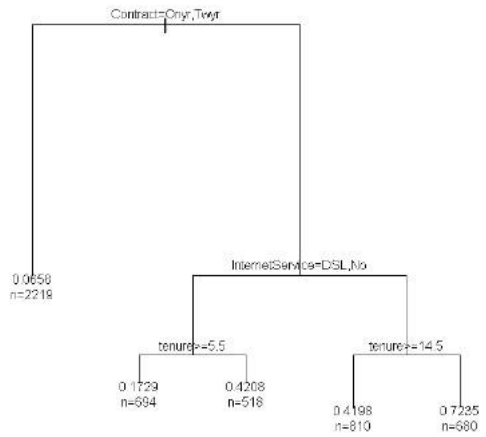


Fig. 36 Decision Tree.

```
Call:
svm(formula = Churn ~ ., data = train, kernel = "linear")


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.04347826
    epsilon:  0.1


Number of Support Vectors:  3718
```

Fig. 37 Support Vector Machine

The AUC-ROC Score is used to know how much model is capable in distinguishing the classes. The higher the AUC Score the better is your classification.

Model	Accuracy	AUC Score
Logistic	79.35	0.834
Decision Tree	44.2	78.6
SVM	44.2	78.8

FIG. 38 AUC Score of the models.

2. Melbourne Housing Data.

In this dataset, we have applied Multinomial Logistic Regression and Naïve Bayes Classification by taking house "Type" as a dependent attribute and all other independent attribute. Our aim is to predict "House type" on the basis of region, number of bedrooms, bathroom, location, etc.

a. Multinomial Logistic Regression.

In multinomial logistic regression, we have selected highly significant attributes to build the model like price, car, bedroom2, etc. Its Accuracy is about 80.56% with AUC score of 0.835 which is good.

```
### Building Classification Table.###

tab <- table(test$Type, test$TypePredicted)
print(tab)
```

```
     1    2    3
1 6824  360   15
2  616 1603   13
3  776  257   16
```

Fig 39. Multinomial Logistic Regression

b. Naïve Bayes Classifier-

Same dependent and independent attribute are used to classify the house type using this machine learning model. Its giving accuracy of 75.68% and AUC score of 0.834.

```
cm <- table(test$Type, pred_y)
cm
```

```
  pred_y
     1    2    3
1 5859  795  545
2  219 1758  255
3  445  290  314
```

Fig 40. Table for Naïve Bayes

Model	Accuracy	AUC Score
Multinomial	80.56	0.835
Naïve Bayes	75.68	0.834

Fig 41 AUC Score for models.

3. Secondhand Car Price Prediction.

In this dataset, we used Linear model to predict the price of secondhand car in which "Price" attribute is target variable else all other are independent variable. Its Adjusted R square is 0.9391

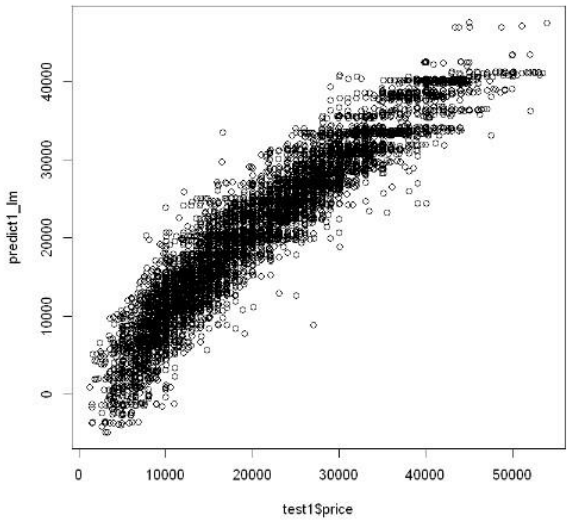
```
###Linear Nodel applies on all numeric attributes.

model_lm = lm(price~., data=train1)
summary(model_lm)
confint(model_lm)
```

```
last_updated29-01-2018
last_updated29-10-2017                ***
last_updated29-11-2017
last_updated30-10-2017                ***
last_updated30-11-2017
litres                                ***
location                              ***
makeToyota                            ***
modelImpreza
modelRAV4
odometer                              ***
transmissionManual                    ***
year                                  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2601 on 24544 degrees of freedom
Multiple R-squared:  0.9397,    Adjusted R-squared:  0.9391
F-statistic:  1494 on 256 and 24544 DF,  p-value: < 2.2e-16
```

```
        studentized Breusch-Pagan test

data:  model_lm
BP = 2951.9, df = 256, p-value < 2.2e-16
```

```
From the above analysis-
Adjusted R-squared:  0.9391
R-squared:  0.9397
BP = 2951.9
p-value < 2.2e-16
```

. Fig 42 BP Test for Linear Regression.



From the BP Test we concluded that our p-value is below 2.2e-16 which means our model is homoscedatic, it means there is noise or random disturbance between the dependent and independent variables.

E. Conclusion

In this paper our task is to implement and execute atleast 5 models on 3 datasets in which telecom customer churn is the one with logistic regression algorithm is applied and the accuracy of it is average while SVM and Decision tree performed worst. For the future work, XGBosst and Random Forest would provide better performance.

For the second data dataset “Melbourne Housing data”, Models are performing pretty well, the scope of betterment is still exists in terms of performance else it correctly justify the house types on the basis of other attributes.

References.

[1] A. Gaur and R. Dubey, "Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), 2018, pp. 1-5, doi: 10.1109/ICACAT.2018.8933783

[2] X. Zhang, G. Feng and H. Hui, "Customer-Churn Research Based on Customer Segmentation," 2009 International Conference on Electronic Commerce and Business Intelligence, 2009, pp. 443-446, doi: 10.1109/ECBI.2009.86.

[3] Nurulhuda Mustafa, Data Curation, Investigation Methodology, Software, Validation Writing – Original Draft Preparation,[1] Lew Sook Ling, Conceptualization, Methodology, Project Administration, Supervision, Validation, Writing- Review & Editing,[a,2] and Siti Fatima Abdul Razak, Conceptualization, Methodology, Software Supervision, Validation, Writing—Review and Editing.

[4] V. E, P. Ravikumar, C. S and S. K. M, "An Efficient Technique for Feature Selection to Predict Customer Churn in telecom industry," *2019 1st International Conference on Advances in Information Technology (ICAIT)*, 2019, pp. 174-179, doi: 10.1109/ICAIT47043.2019.8987317.

[5] A. Siddika, A. Faruque and A. K. M. Masum, "Comparative Analysis of Churn Predictive Models and Factor Identification in Telecom Industry," 2021 24th International Conference on Computer and Information Technology (ICCIT), 2021, pp. 1-6, doi: 10.1109/ICCIT54785.2021.9689881.

[6] P. Durganjali and M. V. Pujitha, "House Resale Price Prediction Using Classification Algorithms," 2019 International Conference on Smart Structures and Systems (ICSSS), 2019, pp. 1-4, doi: 10.1109/ICSSS.2019.8882842.

[7] D. Yu, Z. Wang and W. Wei, "House Price Prediction Based on a Machine Learning Model," 2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), 2021, pp. 391-395, doi: 10.1109/AINIT54228.2021.00082.

[8] Muralidharan S., Phiri K., Sinha S. K., Kim B. Analysis and prediction of real estate prices: a case of the Boston housing market. Issues in Information Systems . 2018;19(2):109–118.

[9] Li X. Prediction PMCID and Analysis of House Price Based on Generalized Linear Reg Model Comput Intell Neuroscience : PMC9536958. 29;2022:3590224 doi: 1 0.1155/2022/3590224. PMID: 36211010; PMC9536958.

[10] D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017, pp. 2998-3000, doi: 10.1109/ICPCSI.2017.8392275.

[11] M. Hankar, M. Birjali and A. Beni-Hssane, "Used Car Price Prediction using Machine Learning: A Case Study," 2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC), 2022, pp. 1-4, doi: 10.1109/ISIVC54825.2022.9800719.

[12] Z. Zhang, "Data Sets Modeling and Frequency Prediction via Machine Learning and Neural Network," 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), 2021, pp. 855-863, doi: 10.1109/ICESIT53460.2021.9696532.

[13] C. Chiu and C. -H. Shu, "Monthly car sales prediction using Internet Word-of-Mouth (eWOM)," 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2017, pp. 345-348, doi: 10.1109/INISTA.2017.8001183.

[14] J. Varshitha, K. Jahnavi and C. Lakshmi, "Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning," 2022 International Conference on Computer Communication and Informatics (ICCCI), 2022, pp. 1-4, doi: 10.1109/ICCCI54379.2022.9740817..

IEEE conference templates contain guidance text for composing and formatting conference papers.