Exploring Data Mining and Machine Learning Models

Ananya Pratap Singh Chandel
School Of Computing
National College Of Ireland,
Dublin, Ireland
x19237529@student.ncirl.ie

***Abstract*—The objective of this paper is to explore and implement Data mining and Machine learning across various domains. Seven Machine learning models have been applied to three datasets from different domains to answer specific research questions. Identifying what factors can lead people for not showing up for their medical appointments. Which factors could contribute for prediction of price of an old car? How many bikes would be required at a particular timestamp for a bike rental service? Different Machine learning models have been applied and evaluated to answer these questions. Few techniques like 10-fold Cross validation, stratified random sampling, Laplace estimation, tuning have been explored. Random forest classifier with an accuracy of 72.41% and F score of 0.838 has been identified best for predicting a No-Show. To predict Price for the old vehicles Multiple linear regression with log transformation on price and predictor variables as Miles per gallon, Engine size and Age of the vehicle has been identified best for prediction with and accuracy of 83.5%(R square value). For predicting the count of bikes for a bike rental service Random forest regressor has been identified to perform the best with an accuracy of 90% and lowest RMSE of 230.**

Keywords—KDD, Multiple linear regression, Naïve Bias, Random Forests, SVM, Decision tree regression.

I. INTRODUCTION

Earlier computers were limited to perform complex calculations with limited storage and computational abilities. Later, computers evolved to perform more complex tasks by merely following instructions. The next stage was to learn from experience where computers can gain insights from the data and establish some rules or even perform predictions. This marked the beginning of machine learning.

With outpouring data produced and recorded every day we have entered an era of Big data. Data mining involves identifying relationships, patterns, or anomalies by teaching the computer on large datasets to gain insights for the business. This involves human intervention for decision making. Whereas Machine learning focuses on developing computer algorithms that allows machines to analyze large datasets and learn the patterns from them and predict on the new data.

Today, Machine learning has wide range of applications in almost every sector. All the business or companies who capture data apply machine learning to some extent to gain insights from the data. This knowledge helps in formulation of business processes and be future ready.

This project aims at applying five different machine learning models to three different datasets.

A. Medical Appointments No-show

Healthcare services shall be readily available for the people when in need. People not showing up for their appointments are considered No-show appointments, these have a significant impact on the healthcare providers. There is a lot of research undertaken to analyze factors that influence No-Show behavior. Results from a research showed that on an average No-shows are 23%, highest being for the African subcontinent which is 43%. [1]. Another research shows that missed appointments cost one billion pounds annually as per UK National Health Service.[3] Increase in No-Snows for medical appointments result in poor health outcomes and poses problems for service providers. The health care services can righteously allocate by the health care if they could predict the people who might not show up for their appointments. By collecting data on the appointments and predicting the attendance of patients could be of great help in healthcare resource allocation. This project aims on predicting the No-Show by trying to analyze different factors from data set taken from Kaggle.[2]

B. UK Used car dataset

With the rapid increase in the number of private cars the market for preowned cars has also developed.[5] There is a significant increase in number of people Buying preowned cars. There are several platforms that list used cars and help both the buyer and sellers. It would be helpful for the buyers if they can assess and compare the price of used vehicles and make a right decision. Similarly, it would be helpful for the Sellers of the secondhand car, if they could determine the right price for their vehicle to sell. In this assignment Price of pre-owned cars is predicted on the data set for used cars in UK from Kaggle.[4] It has the price of the used cars of more than 100,000 cars in UK along with several makes. For this analysis we used car data for Audi manufacturer to predict the price of the used Audi vehicles in UK. An accurate price prediction for a used vehicle can help the buyers and sellers for evaluating the appropriate price for their vehicle.

C. Bike Rental

Due to rapid development and increase in population there has been a significant growth in the number of Vehicles this poses stress on environment by causing pollution and result in traffic problems in large cities. With the increase in the awareness about these problems Bike sharing systems have evolved. Bike rental systems allows a person to temporary rent a bike and return it to a docking station. There are several bike rentals services that have started throughout the World. For example, Dublin has a bike rental system named Dublin bikes. There are several other bike rentals that have started recently as people are steering towards a healthier and eco friendly mode of transport. This also helps in curbing the pollution and traffic problems for a City. If we could predict the count or the number of bikes that are required at a given time and place this could help the bike rental services in delivering better services and maintain enough bikes at their Kiosk. In this assignment the count of bikes is predicted using the data from London Bike sharing system.[6]. The prediction of the count is done considering different parameters like season, Weather,

Temperature, Holidays. This would help the companies providing bike sharing services to forecast the bike demand and maintain the count of required bikes based on the holidays, temperature, and season information.

II. Related Work

The following are the related works on Medical Appointment No shows, Used car price prediction and Bike rental systems.

A. Medical appointments

Low attendance and no shows of medical appointments deprive the needy patients of the required Healthcare services. Accurate prediction of the medical appointment can result to better staff allocation less waiting time and adequate allocation of hospital resources. An analysis was performed on the data extracted from a data warehouse that contained the data for multiple Hospitals in Singapore captured by their IT systems, 42 variables were extracted on consultation with domain experts. This contained millions of records on which text mining was performed to extract useful information and XGBoost was applied to gain an AUC of 0.793 on 15 important features like days since last visit, prior no show, appointment waiting time etc. [7] These parameters were helpful for determining No show for this assignment while applying different models. A similar analysis that was performed on data obtained from famous Brazilian hospitals which encountered large number of applications every day. The data contained columns like No-show (yes or no), Patient ID, Alcoholism, Handicap, Sms received etc. All the mentioned factors were considered for maintaining records of the patients. Exploratory analysis was performed on all the variables to predict no shows using independent variables. Like a person who received reminder message for the appointment did not show up for the appointment 23% of the time. After considering all the aspects logistic regression model was applied to the data with an accuracy of 0.86. Later, a Decision tree Classifier was applied on the data which resulted in improved accuracy of 0.89.[8] The data used in [8]is similar to the data used for this assignment. This assignment intends to explore the data further by applying different machine learning models apart from logistic regression and decision tree.[2] A similar kind of work was done in a research to predict No shows specific to the radiology hospitals. 16 years of patient data from radiology department of a multi-site hospital was combined with patient income obtained from U.S Census data. Different Logistic model were developed with respect to feature groups to access the likelihood of no show. Five fold cross-validation was performed on the feature set of 554,611appointment data. This resulted in the AUC of 0.77.[9]This model can be improved by exploring and applying other machine learning algorithms like random forests or Naïve Bias. Another study was performed on Data from public hospital in Singapore where 19 Demographics related variables were taken like Referral type, appointment day of the week, department etc. Later Logistic regression model and recursive partitioning was applied on the data after splitting into train and test data. The logistic regression model could predict around 70% of the no show with Kappa coefficient of 0.41 on validation data.[10]We will be following the same approach of splitting the data into train and test and building different machine learning models apart form the one already applied in the above literature with the goal to achieve more accuracy by overcoming the limitations of the cited literature.

B. UK used car price prediction

As the demand for private car has increased, demands for second hand cars have increased significantly all around the world. Second hand cars are affordable for the buyer. A model that could evaluate price for the used cars will be helpful for both the buyer and the seller. Keeping this in mind in a Study comparative studies were performed to build a model for predicting the price of an old vehicle. The models were build on the data that was web scraped from an German ecommerce website for used car market. Different machine learning models like multiple linear regression, gradient boosted regression trees and Random forest were used to build a Price prediction model. As a result, the best results were obtained from gradient boosted regression with the mean absolute error of 0.28 while multiple linear regression showed MSE of 0.55 and random forests to be 0.35.[11]We will be exploring these models on our data and try to achieve better results by comparing the performance of different models. A similar study explained that predicting car price is a challenging problem as there are many different factors that affect the car price like models, year of manufacture, engine etc. This study focused on quantifying the qualitative data before applying a model. To achieve this One-Hot encoding was applied. The data was collected from four different automobile websites in Vietnam. Machine learning Models like Random Forests, XGBoost, Light Gradient Boosting Machine were applied to achieve R2 performance of 0.75, 0.64 and 0.70 respectively.[12]For our Analysis we try to apply these model on our data[4] and compare the results also tried Quantifying variables in preprocessing to check for improved results. Another related study that aimed at predicting the MPG i.e miles per gallon based on the vehicle data from American, European, Asian origin. Various factors like horsepower, weight and age of the vehicles was used to predict the MPG of a vehicle using Linear Regression model with an RMSE of 3.26 and R^2 of 0.82.[13] This study is similar to the studies for building Price prediction models. For this assignment we will be using the MPG to analyze its relationship with price while building a price prediction model.

C. Bike Rental Systems

The use of bike rental services has gained popularity recently. The bikes rental services are an efficient mode of transport for short distance. It saves time and is faster and convenient. But if the bike stations do not have required number of bikes it would be inconvenient for people to use the service. As per a statistic, In Taiwan, Taipei city, around over a million people rented bike from each station every month. In a study conducted on dataset that contained bike rental data for Porto, Portugal. Thirteen features were selected to apply machine learning models like Random Forest and Support Vector Regression and combining both to predict the price of the rental service. It was observed that Random forests Performed the best and the SVM had least accuracy having error of 68%. It was concluded that SVM is weak for predicting Regression data.[14] In this assignment we will be using a similar approach but instead of price we predict the count of bike, With feature selection we try to improve the performance of SVM for predicting regression data. Another study on Bike sharing data for Washington, D.C was performed by combining weather data in order to forecast bike

demand throughout the city. The models that were applied on this data included Multiple Linear regression using SPSS software and Random Forest and Decision tree with GMB packet. The linear regression model showed the least accuracy, Random forest improved the accuracy significantly Lastly, Decision tree was applied with GBM package to improve performance and an accuracy rate of 82% was obtained.[15]Similar analysis was performed on the bike data from a bicycle sharing system in Barcelona. The CEO of the company Bicing observed that the customers could not find empty spots while returning the bikes making the service inconvenient for use. This problem could be solved by predicting the number of bikes that may be expected at a Kisok and managing its capacity. To Apply the machine learning models the data is taken from Bicing system API that can be accessed online, along with the Holiday and weather data for Barcelona. ARIMA and Random forests models were applied on the data. Random forest outperformed the ARIMA models. While ARIMA models could only predict 40% statuses for the bikes While Random Forests returned with an accuracy of 86%.[16] ARIMA models can be used for time series data and are also efficient for forecasts. But it can be concluded that Random forests performed better than the ARIMA models, it is inferred not applying ARIMA models for this assignment will be appropriate.

III. KDD Methodology

For this assignment Knowledge Discovery in Databases (KDD) methodology has been followed throughout. KDD methodology can be seen in Figure 0.

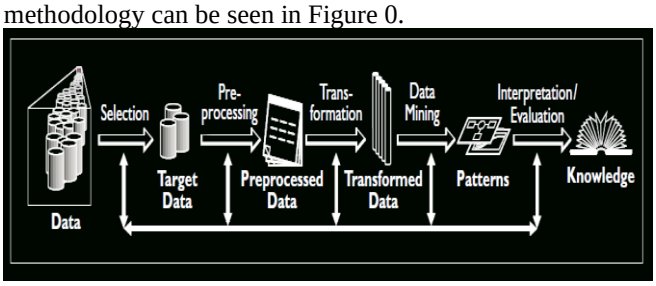

Figure 0, KDD Methodology

A. *Medical Appointment Data*

1. Data selection:

Medical Appointment dataset is obtained from Kaggle website.[2] This dataset contains the appointment information of the patients. It also contains No-Show data for the patients, who do not show up for their appointments. It has 110,527 observations of medical appointments and 14 variables containing attributes like patient ID, Appointment ID, Age, Sex, Appointment time.[2]

2. Data preprocessing:

The structure of the data was checked and from the summary it was observed that Age contained negative values, the negative values for Age were removed. The data was checked for the NA values no missing values were found. Missing values were visualized using Amelia Package in R using Missingness map function. As can be seen in Figure 1.

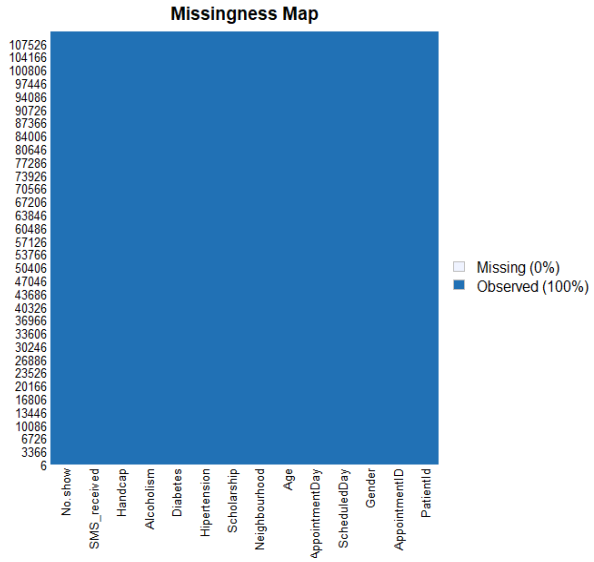

Figure 1: Missing values in Noshow

3. Data Transformation:

Once the NA values have been removed and Age with negative values were removed, few attributes were transformed as factors like, Gender, Scholarship, Sms_received that were Dichotomous. The target column to be predicted i.e No_Show was also transformed into factor Yes/No. As inferred from the previous work Waiting time could have effect on dependent variable Noshow. It is observed to be an important factor, therefore its calculated using the columns Appointment day and Scheduled day columns. The waiting time less than zero is filtered using Dplyr package in R. The unnecessary columns like Patient ID, Appointment ID, Appointment and Scheduled day are dropped as they do not help in model building. The name of the columns is changed to make the data more understandable. After the data is cleaned and structured, we visualize the Dependent variable against the different Independent Variables. Using ggplot library we visualize Age and similarly other variables to check for relationships with the dependent class i.e no show data. As can be seen in figure 2.

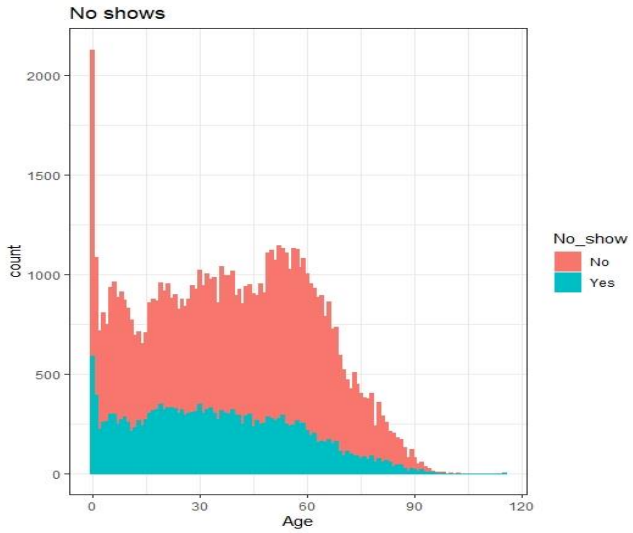

Figure 2: Age with fill factor as no show

4. Data Mining Model:

Three different Models Logistic regression, Naïve Bias and Random Forest Classifier are applied on the data and evaluated.

- Logistic Regression:

Logistic Regression is applied by taking all the factors into consideration. Later, only the variables that were significant towards predicting No shows have been selected. The data has been split into train data and test data using createDataPartition function from R's Caret package which randomly samples the data. The data is split into train data and test data. The data was split into 75/25 ratio for train and test. Later the Logistic regression has been applied on the train data using the glm function. In which No Show is predicted against Age, Scholarship, Waiting time, Alcoholism, Diabetes and Hypertension.

- Naïve Bias Classification:

It is used to compute the probabilities of a categorical variable using the independent variables as per the Bayes rule. Naïve bias is applied using the R's library 'e1071'. The model is applied on the train data like the logistic regression. x is supplied as a numeric vector containing all the variables except no show and y is imputed as a class vector, the column containing the noshow data.

- Random Forest Classifier:

Random forest Algorithm can be used for both classification and regression. For this data set random forest classifier has been used to classify the No shows on the appointment data. The neighborhood column has been dropped before applying the random forest as it contained way too many factors. The data is similarly split into test and train data and the model is applied using the randomForest function in R. The data frame containing the predictor variables is provided along with the response vector Noshow in this case. The number of trees has been provided as 500 to ensure every row executes few times.

5. Interpretation and Evaluation:

On applying Logistic Regression model and evaluating it using confusionMatrix, CrossTable, and F score, we get Accuracy of 71.47%, F1 score of 0.833 and the value of Precision and recall being 0.71 and 0.99. We observe that the model mostly predicts NO 71.5% times i.e the person not showing up for the appointment, but it poorly predicts the person showing up for the appointment. This shows class imbalance. As can be seen from figure 3.

```
Confusion Matrix and Statistics

          Reference
Prediction    No   Yes
       No  12851  5125
       Yes     8     5

               Accuracy : 0.7147
                 95% CI : (0.708, 0.7213)
    No Information Rate : 0.7148
    P-Value [Acc > NIR] : 0.5235

                  Kappa : 5e-04

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9993779
            Specificity : 0.0009747
         Pos Pred Value : 0.7148976
         Neg Pred Value : 0.3846154
             Prevalence : 0.7148257
         Detection Rate : 0.7143810
   Detection Prevalence : 0.9992773
      Balanced Accuracy : 0.5001763
```

Figure 3 Logistic model Results

Similarly, On Evaluating Naïve Bias model we get Accuracy of 71.45%, F1 score of 0.832, Precision of 0.7162 and Recall of 0.99. Naïve Bias model works better than Logistic model for predicting TN (Yes) i.e number of people showing up for the appointment. As can be seen from figure 4.

```
Confusion Matrix and Statistics

          Reference
Prediction    No   Yes
       No  12790  5067
       Yes    69    63

               Accuracy : 0.7145
                 95% CI : (0.7078, 0.7211)
    No Information Rate : 0.7148
    P-Value [Acc > NIR] : 0.5432

                  Kappa : 0.0098

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99463
            Specificity : 0.01228
         Pos Pred Value : 0.71625
         Neg Pred Value : 0.47727
             Prevalence : 0.71483
         Detection Rate : 0.71099
   Detection Prevalence : 0.99266
      Balanced Accuracy : 0.50346
```

Figure 4 Naïve Bias results

Lastly, On Applying Random Forests Accuracy increased to 72.41% but Specificity improved it could predict the TN more than the other models. F score of 0.838, precision of 0.72 and recall of 0.99 was obtained. As can be seen in figure 5.

```
Confusion Matrix and Statistics

          Reference
Prediction    No   Yes
       No  12840  4945
       Yes    19   185

               Accuracy : 0.7241
                 95% CI : (0.7175, 0.7306)
    No Information Rate : 0.7148
    P-Value [Acc > NIR] : 0.003064

                  Kappa : 0.0486

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.99852
            Specificity : 0.03606
         Pos Pred Value : 0.72196
         Neg Pred Value : 0.90686
             Prevalence : 0.71483
         Detection Rate : 0.71377
   Detection Prevalence : 0.98866
      Balanced Accuracy : 0.51729
```

Figure 5 Random Forests Results

6. Improving model performance:

To improve performance of Logistic model k-fold cross-validation was used by keeping value of k as 10. k-fold cross validation has become an industry standard for estimating model performance. Using 10-fold cross validation improved the performance of the model. As can be seen in the figure 6.

```
> log_model3
Generalized Linear Model

53970 samples
    7 predictor
    2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 48573, 48573, 48573, 48573, 48573, 48573, ...
Resampling results:

  Accuracy   Kappa
  0.7146007  0.00104174
```

Figure 6 10-fold Cross-Validation

To improve the performance of the Naïve bias model we added laplace estimator as 1 this helped in reducing number of false positives for our model. The Random Forest model is an ensemble-based method that focuses on ensembles of decision trees. This model already gave the best accuracy by defining number of trees as 500.

B. *UK Used Car data*

1. Data selection:

This data set has been taken from Kaggle.[4] It has data of around 100,000 cars in UK. Data of cars made by Audi is taken for this assignment, it has 10,668 observations and 9 variables. This data is read as csv and converted to data frame(cardf).

2. Data preprocessing:

Data is checked for NA values using is.na(cardf) function in R. There were no null values in the data. The structure of the data can be seen in figure 7 below.

```
> #checking structure of df
> str(cardf)
'data.frame':   10668 obs. of  9 variables:
 $ model       : chr  " A1" " A6" " A1" " A4" ...
 $ year        : int  2017 2016 2016 2017 2019 2016 2016 2016 2015 2016 ...
 $ price       : int  12500 16500 11000 16800 17300 13900 13250 11750 10200 1200
 $ transmission: chr  "Manual" "Automatic" "Manual" "Automatic" ...
 $ mileage     : int  15735 36203 29946 25952 1998 32260 76788 75185 46112 22451
 $ fuelType    : chr  "Petrol" "Diesel" "Petrol" "Diesel" ...
 $ tax         : int  150 20 30 145 145 30 30 20 20 30 ...
 $ mpg         : num  55.4 64.2 55.4 67.3 49.6 58.9 61.4 70.6 60.1 55.4 ...
 $ engineSize  : num  1.4 2 1.4 2 1 1.4 2 2 1.4 1.4 ...
```

Figure 7 Structure of car df

3. Data Transformation:

Data is transformed to fit regression models for predicting price. Age of the vehicle has been calculated by subtracting current year with the year of manufacture. Then model and year attributes were dropped from the data frame. Categorical variables like fuel type and transmission have been converted to factor rest price, mileage and tax were converted to numeric data. On visualizing price using ggplot it was observed to be right skewed which could be corrected by using log transformation on price but on removing the outliers the price variable showed normal distribution. As can be seen in figure 8.

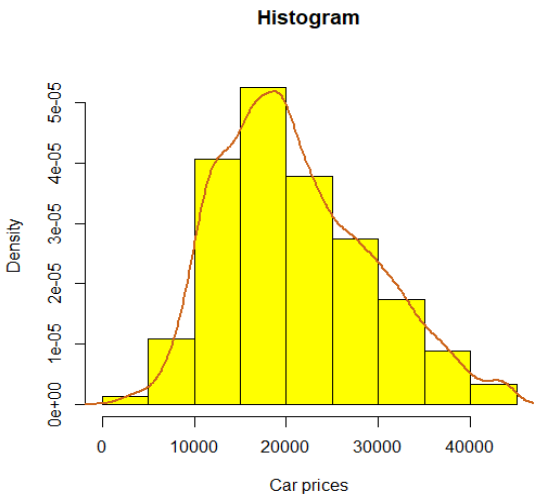

Figure 8: Car price distribution

On checking the correlations in price and other numeric attributes we found that price is correlated with tax, Age, milespergallon, enginesize and mileage. Correlations were visualized using corrgram function in R As can be seen in figure 9.

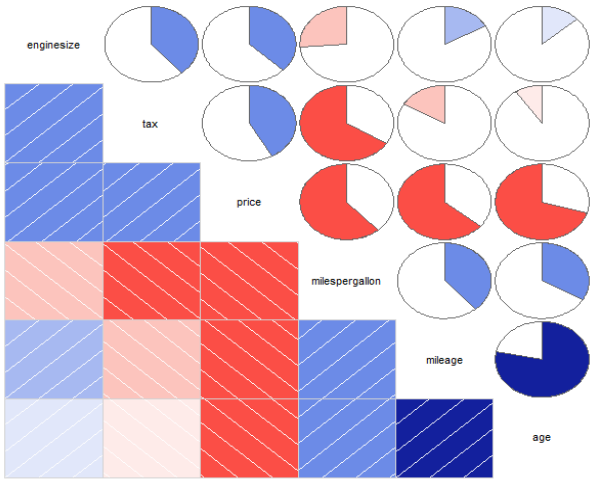

Figure 10. Correlations plot

Lastly, variable names were changed to enhance the interpretability.

4. Data Mining Model:

Two machine learning models that are applied and evaluated for this data are Multiple linear Regression and Decision tree regressor.

- Multiple Linear Regression:

Data was split into train and test data, the ratio of 70:30 was maintained using sample.split function from R's caTools package. Best subset regression was applied using Regsubset function from leaps package in R. it was used to visualize the most influencing predictor variables for price. As can be seen in figure 11.

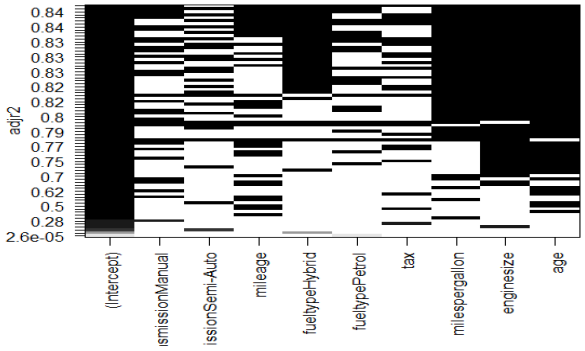

Figure 11: Regression Subset

Mileage, Milespergallon, enginesize and age were selected to build the model.

- Decision Tree Regression:

The data was split into train test data with 80:20 ratio using stratified random sampling. Decision tree regression was applied on the train data to predict price using R's rpart function from rpart library.

5. Interpretation and evaluation:

On Applying the Multiple linear regression model for predicting price using Mileage, Milespergallon, enginesize and age as the predictors an RMSE of 3816 was obtained and the R square value of 80% was obtained. This means that model is able to predict the price of the old

vehicle with 80% accuracy while the root mean square error being high with value of 3816. As can be seen in figure 12.

```
> car.fit1.lm <- lm(price ~ mileage+age+milespergallon+enginesize,data=car_tra
> summary(car.fit1.lm)

Call:
lm(formula = price ~ mileage + age + milespergallon + enginesize,
    data = car_train)

Residuals:
   Min     1Q Median     3Q    Max
-20780  -2248   -445   1690  41938

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.401e+04  2.976e+02   80.68   <2e-16 ***
mileage        -7.470e-02  3.116e-03  -23.97   <2e-16 ***
age            -1.981e+03  3.286e+01  -60.30   <2e-16 ***
milespergallon -1.592e+02  4.028e+00  -39.52   <2e-16 ***
enginesize      7.130e+03  9.155e+01   77.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3839 on 7568 degrees of freedom
Multiple R-squared:  0.8007,    Adjusted R-squared:  0.8006
F-statistic:  7600 on 4 and 7568 DF,  p-value: < 2.2e-16
```

Figure 12: Linear regression Model summary

Decision tree regressor was evaluated by checking the correlation between the predicted price and the actual test price, correlation value of 0.88 shows that they are highly correlated which is a positive aspect for our model. An RMSE of 3990 and MAE of 3036 was obtained for the model. By this we can Interpret that our Linear regression model performs better that the regression tree. Plot of the decision tree can be seen in Figure 13.

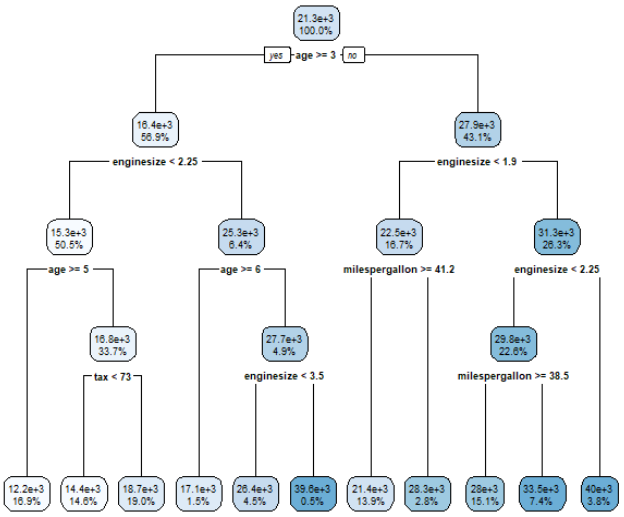

Figure 13: Decision tree Plot

6. Improving model performance:

To improve the model performance, we built multiple linear models by using different combinations of attributes and checking residual plots to verify the fit of the model. Model with log transformation on the price, predicted over age, milespergallon and enginesize resulted in the highest R square value of 83.52%.

Residual plots for this model also provide evidence for a decent fit of the model. As can be seen in figure 14.

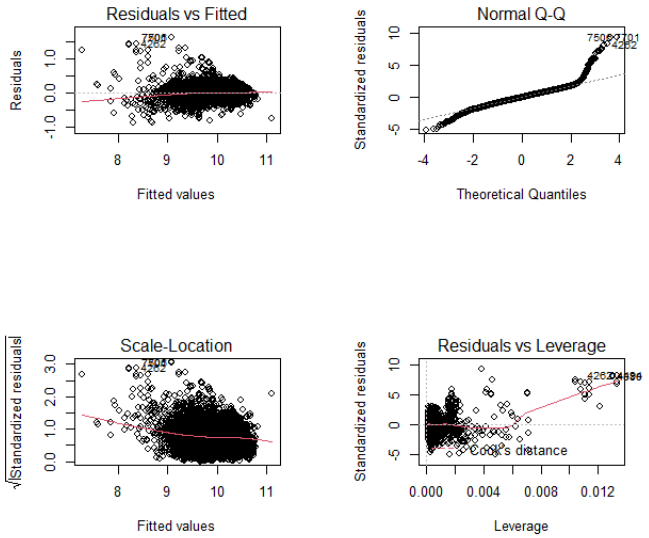

Figure 14: Residual plots

C. *Bike Rental Data*

1. Data selection:

The data for bike rental service in London is downloaded from Kaggle website.[6] Originally Data was obtained by combining bike rental service data, weather data and bank holiday data in UK. Data set contains 17,414 rows and 10 attributes like timestamp, temperature, is_holiday and Count that has been used as dependent variable. Data is downloaded as csv and read as data frame using readr library in R.

2. Data preprocessing:

The data has been checked for the null values using is.na() function in R. There are no NA values in data, this was also visualized using Amelia package in R. Structure of the data can be seen in the figure 15.

```
> #step2 preparing and exploring data
> str(bike_data)
tibble [17,414 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ timestamp   : POSIXct[1:17414], format: "2015-01-04 00:00:00" "2015-01-04
 $ cnt         : num [1:17414] 182 138 134 72 47 46 51 75 131 301 ...
 $ t1          : num [1:17414] 3 3 2.5 2 2 2 1 1 1.5 2 ...
 $ t2          : num [1:17414] 2 2.5 2.5 2 0 2 -1 -1 -1 -0.5 ...
 $ hum         : num [1:17414] 93 93 96.5 100 93 93 100 100 96.5 100 ...
 $ wind_speed  : num [1:17414] 6 5 0 0 6.5 4 7 7 8 9 ...
 $ weather_code: num [1:17414] 3 1 1 1 1 1 4 4 4 3 ...
 $ is_holiday  : num [1:17414] 0 0 0 0 0 0 0 0 0 0 ...
 $ is_weekend  : num [1:17414] 1 1 1 1 1 1 1 1 1 1 ...
 $ season      : num [1:17414] 3 3 3 3 3 3 3 3 3 3 ...
```

Figure 15: structure of bike_data

3. Data Transformation:

Order of the columns have been changed to keep the dependent variable count first. Names of the variables have been changed to increase the understandability of the attributes. Different factors have been visualized against Count to explore and visualize the dependencies of the independent variables against our dependent variable i.e Count. The timestamp attribute has been processed to create new attributes like Date, Day, Month, Year, Hour. Later, the timestamp and date columns are dropped to avoid redundancy. Attributes like Month, Day, Season, Workday and holiday were converted to factors and labels were added. The weather attribute is checked and filtered using dplyr to reduce levels with less data. All the numeric

columns have been checked for correlation, High correlation was observed between temperature and Feel temperature, hence, feel temperature was dropped. Data for year was visualized against count and data for year 2017 has been dropped as there is not enough data for year 2017. As can be seen in Figure 16.

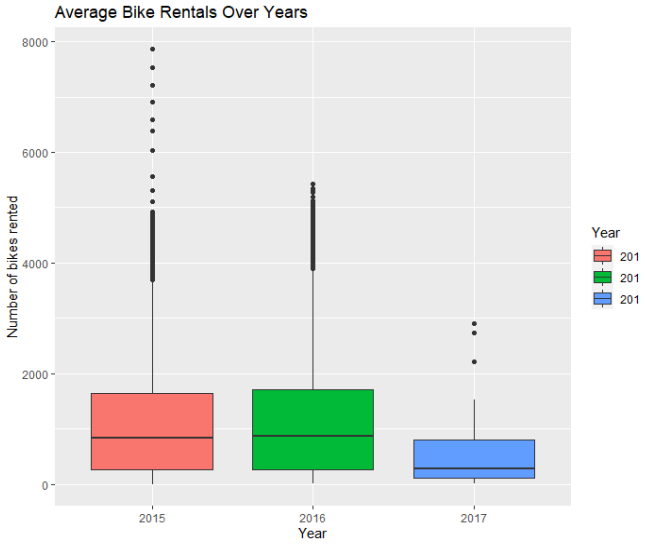

Figure 16: Bike rentals over years

4. Data Mining Model:

- Decision tree Regression and Random Forest Regression:

Initially decision tree regression was used Later, Random Forest regression is used as it builds multiple decision trees to get a more accurate decision. Random forest combines prediction from each decision tree to give us a more accurate result. Random forest is applied for regressing Count over the train data. Data has been divided into train test splits in the ratio of 75/25 using caTools library in R.

- Support Vector Machine:

Support Vector Regression has been applied in the data using the similar train test data as mentioned in the previous section. It is a supervised learning method, and it can be used for both classification and regression problems, also there is no need to check for the normal distribution of the target variable when applying Support vector regression. SVM can be found in 'e1071' library in R, by using function svm, support vector regression has been applied.

5. Interpretation and evaluation:

The individual performance of the model has been measured using the MSE, RMSE, MAE and R2 functions to calculate Mean square error, Root mean square error, Mean absolute error and R square value. For this Metrics and MLmetrics library were used.

- Decision tree Regression and Random Forest Regression:

On applying Decision tree Regression(bike_model0) an RMSE of 498, MAE of 354.20 and R^2 value 0.55 i.e 55% was obtained. The lower R square and high RMSE value indicates that the model performs poorly and can only explain 55% variance in the target variable. On the other hand, on applying random forest regression RMSE of 230, MAE of 147 and R square value of 0.90 i.e 90% is obtained. Random forest regressor performs well in fitting the data with lower RMSE and improved accuracy of 90%.

- Support Vector Machine:

On Applying Support vector regression we obtained RMSE of 596, MAE of 322.83 and R square value of 0.50 i.e 50%. Model performed poorly as could be indicated by the high RMSE of 596 and lower R square of 50%.

6. Improving model performance:

The performance of the decision tree regression model was increased using Random forests regression which uses multiple regression trees to enhance the accuracy of the model and it was observed the random forest regressor model performed the best with the least RMSE of 230 and high R square value of 0.90 i.e 90% accuracy.

Performance of the SVM regression model has been increased by tuning the model. Model was tuned on a sample obtained from data using sample_frac function from dyplr library in R. tune function has been used to tune model. After tuning, tuned parameters are applied in the SVM, kernel as radial, gamma value of 0.5 and cost value 4 gave an enhanced accuracy of R square 0.85 i.e 85%, MAE of 199.35 and reduced RMSE of 286.78. As can be seen in the Figure 17.

```
> #applying tuned parameters
> bike_m3 <- svm(formula = Count ~ .,
+           data = bike_train,
+           type = 'eps-regression',
+           kernel = 'radial',
+           gamma = 0.5,
+           cost = 4)
> ################### MODEL 3 EVALUATION ################
##
> pred3 = predict(bike_m3,bike_test)
> MSE3 = MSE(bike_test$Count, pred3)
> MAE3 = MAE(bike_test$Count, pred3)
> RMSE3 = RMSE(bike_test$Count, pred3)
> r23 = R2(bike_test$Count, pred3, form = "traditional")
> cat(" MAE:", MAE3, "\n", "MSE:", MSE3, "\n",
+     "RMSE:", RMSE3, "\n", "R-squared:", r23)
 MAE: 199.3502
 MSE: 82244.54
 RMSE: 286.7831
 R-squared: 0.850268
```

Figure 17: Tuned SVM Results

IV. Knowledge Discovery and Overall Evaluation

A. *Medical Appointment Data:*

For classifying no shows in medical appointment data total three classification models were used Logistic regression, Naïve Bias and Random forest. Total 6 models were built, 3logistic, 2 naïve bias and a random forest model. Three different Logistic models were built, among these three models, the one with Age, Scholarship, Hypertension, Alcoholism, Diabetes, Sms_received, Waitingtime as predictors and on performing 10-fold cross validation gave the best fit with Accuracy of .715, precision of 0.71, recall 99 and F score of 0.83. But the logistic model was only able to correctly classify only 5 true negatives i.e lowest specificity. Naïve Bias model with laplace as 1, gave results as accuracy, 0.7145, Precision of 0.716, F score of 0.832 and it was able to categorize 63 True negatives i.e improved Specificity. Lastly, It was observed Random Forest model gave accuracy

of 0.7141, F score of 0.838 precision of 0.72 but it was able to classify 185 cases of True Negative that is the best Specificity. From the evaluations it was observed that the Random forest model showed the best Specificity and almost equal Sensitivity i.e correctly classifying true positives (No shows). If we look at the tradeoff between the Sensitivity and the Specificity Random forest performed the best when compared to other models.

B. UK cars data:

Total 5 multiple linear models and a Decision tree regressor were implemented to estimate the price of old vehicles. Use of regression subset for linear regression gave the results as RMSE of 3816 and R square value of 80%. Price was regressed over transmission, mileage, fueltype, tax, milespergallon, enginesize, age to created models. The best model with R square of 83.52% was obtained by performing log transformation on the price the model and predictor variables as Age, Engine size and mile per gallon. The decision tree showed high correlation between the original and predicted price to be 0.88 and an RMSE of 3990. The multiple linear models performed well in this scenario, this can also be seen from the section above Q-Q plots of the residuals show normal distribution and there are no abnormalities in the plots making this model a good fit.

C. Bike Rental data:

Total 4 models have been fitted with London bike sharing data to predict the count of the rental bikes based on weather data, holiday data and bike sharing data. Decision trees Random forests and support vector machines were used. Decision tree model with the R square value of 0.55 and RMSE of 498 fitted the model poorly and was improve using random forest regression with RMSE of 230 and R square value of .90. This has been observed as the best model in predicting the count of London bike rental service. Support vector regression was used to build 2 models first model, gave RMSE of 596 and R square of 0.50 which was poor result. To improve the result of the SVM model Tuning has been used and the model has been rebuilt using the tuned parameters to get an enhanced accuracy with R square being 0.85 and RMSE of 286.78. The random forest regressor gives the best results when compared to the other models its fairly able to predict the counts for the bike rental service based on the weather data, holidays in UK information and bike sharing data.

V. CONCLUSION AND FUTURE WORK

To conclude, seven different machine learning models were applied on three different data sets. By applying these machine learning models we tried to explore the performances of the various machine learning models and tried answering three different research questions. First, Identifying the reasons that influence people not showing up for the appointment, three classification models Logistic regression, Naïve bias, Random forests were built to answer this question in which combination of various attributes like Neighborhood, Waiting time, Sms received, Existing Ailments were tried. Random Forest model with an accuracy of 71.41% performed the best as it was able to correctly justify Specificity and Selectivity as compared to other models. Second, Identifying the appropriate price for selling a used car, by building various models and applying Multiple linear regression and Decision tree regression, Model with log price transformation regressed over predictors like Age, Miles per gallon and engine size gave the best R square value of 83.52%. This model was able to predict the price of the old car with the accuracy of upto 83.52% based on its mpg, age and enginesize. Third, Identifying the count of bike needed by bike rentals services at a given time. To identify the count of bikes based on weather conditions like temperature windspeed, seasons, holidays multiple models were built using decision trees, SVM and random forest regression. Random forest regression with an RMSE of 230 and R square value of .90 performed the best, it was able to predict the count of the bikes with an accuracy of 90%.

All the predictions by the models are based on limited number of parameters or attributes that were available in the data. As the future scope we could try to predict by taking into consideration a larger number of parameters with few extra attributes. To achieve this extra information and attributes will have to be recorded for analysis. If given time, time series analysis and ARIMA models could be applied and explored on the bike rental data to forecast the future requirement for the bikes. The classification problem dataset was slightly imbalanced, given time it can be balanced by using ROSE or SMOTE over sampling techniques to improve the model performances.

REFERENCES

[1] Dantas, L.F., Fleck, J.L., Cyrino Oliveira, F.L. & Hamacher, S. 2018, "No-shows in appointment scheduling – a systematic literature review", Health Policy, vol. 122, no. 4, pp. 412-421.

[2] "Medical Appointment No Shows", *Kaggle.com*, 2020. Available:https://www.kaggle.com/joniarroba/noshowappointments.

[3] Incze, E., Holborn, P., Higgs, G. & Ware, A. 2021, "Using machine learning tools to investigate factors associated with trends in 'no-shows' in outpatient appointments", *Health and Place,* vol. 67.

[4] "100,000 UK Used Car Data set", *Kaggle.com*, 2020.Available:https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes.

[5] "Price evaluation model in second-hand car system based on BP neural network theory - IEEE Conference Publication", *Ieeexplore.ieee.org*, 2019.Available:https://ieeexplore.ieee.org/document/8022758.

[6] "London bike sharing dataset", *Kaggle.com*, 2016. Available:https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset/discussion/114708.

[7] G. Lee et al., "Leveraging on Predictive Analytics to Manage Clinic No Show and Improve Accessibility of Care," 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, 2017, pp. 429-438, doi: 10.1109/DSAA.2017.25.

[8] M. D. A. Praveena, J. S. Krupa and S. SaiPreethi, "Statistical Analysis Of Medical Appointments Using Decision Tree," 2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2019, pp. 59-64, doi: 10.1109/ICONSTEM.2019.8918766.

[9] R. J. Mieloszyk, J. I. Rosenbaum, P. Bhargava and C. S. Hall, "Predictive modeling to identify scheduled radiology appointments

resulting in non-attendance in a hospital setting," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Seogwipo, 2017, pp. 2618-2621, doi: 10.1109/EMBC.2017.8037394.

[10] N. L. Ma, S. Khataniar, D. Wu and S. S. Y. Ng, "Predictive Analytics for Outpatient Appointments," 2014 International Conference on Information Science & Applications (ICISA), Seoul, 2014, pp. 1-4, doi: 10.1109/ICISA.2014.6847449.

[11] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, 2018, pp. 115-119, doi: 10.1109/ICBIR.2018.8391177.

[12] D. Van Thai, L. Ngoc Son, P. V. Tien, N. Nhat Anh and N. T. Ngoc Anh, "Prediction car prices using quantify qualitative data and knowledge-based system," 2019 11th International Conference on Knowledge and Systems Engineering (KSE), Da Nang, Vietnam, 2019, pp. 1-5, doi: 10.1109/KSE.2019.8919408.

[13] V. Shirbhayye, D. Kurmi, S. Dyavanapalli, A. S. Hari Prasad and N. Lal, "An Accurate Prediction of MPG (Miles Per Gallon) using Linear Regression Model of Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-5, doi: 10.1109/ICCCI48352.2020.9104138.

[14] Y. C. Shiao, W. H. Chung and R. C. Chen, "Using SVM and Random forest for different features selection in predicting bike rental amount," 2018 9th International Conference on Awareness Science and Technology (iCAST), Fukuoka, 2018, pp. 1-5, doi: 10.1109/ICAwST.2018.8517237.

[15] Y. Feng and S. Wang, "A forecast for bicycle rental demand based on random forests and multiple linear regression," 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, 2017, pp. 101-105, doi: 10.1109/ICIS.2017.7959977.

[16] G. M. Dias, B. Bellalta and S. Oechsner, "Predicting occupancy trends in Barcelona's bicycle service stations using open data," 2015 SAI Intelligent Systems Conference (IntelliSys), London, 2015, pp. 439-445, doi: 10.1109/IntelliSys.2015.7361177.