Project Name - Cab Fare Prediction

**Problem Statement :**

You are a cab rental start-up company. You have successfully run the pilot project and now want to launch your cab service across the country. You have collected the historical data from your pilot project and now have a requirement to apply analytics for fare prediction. You need to design a system that predicts the fare amount for a cab ride in the city.

**SOLUTION :**

We are given train and test dataset and our task is to predict the fare\_amount values in test dataset.The target variable in the train dataset contains continuous and integer values which simply implies that target is a numerical variable.

So,this is a Regression problem.And for Regression problem firstly ,we will predict the values using Linear Regression Machine Learning Algorithm.

Before implementing this algorithm,we will determine various characterstics of the dataset for ex- no of observations and variables present,type/class of various variables present,summary of variables and dataset,finding whether missing values are present and if present then undergoing missing value analysis.After this we will perform various Machine Learning Algorithm on our dataset to predict the target variable(fare\_amount) values and depending on the performance of the model, we will choose the best model and those predicted values will be our final result.

Now,starting with our train dataset.Let’s first load the data and determine the characteristics and properties of our data present in train dataset.

**R Code :**

rm(list=ls()) #clean the RAM

setwd("C:/Users/user/Documents") #set working directory

getwd() #get current working directory

df=read.csv("C:/Users/user/Downloads/train\_cab.csv",header=T) #loading train dataset

df2=read.csv("C:/Users/user/Downloads/test.csv",header=T) #loading test dataset

dim(df) #getting no of observations and variables in train dataset

dim(df2) #getting no of observations and variables in test dataset

str(df) #getting structure variables names,their type and data of train dataset

str(df2) #getting structure variables names,their type and data of test dataset

colnames(df) #column names of train dataset

colnames(df2) #column names of test dataset

class(df$target) #type of variable target in train dataset

class(df2$var\_0) #type of variable var\_0 in test dataset

Now,let us start with our Missing Value Analysis :

miss\_val=data.frame(apply(df,2,function(x){

sum(is.na(x))

}))

miss\_val

This will give missing values present in each variable.If missing values found then we can undergo mean,median and KNN imputation method to determine which method is best to depict missing values using this dataset.This would be decided depending on the values nearance to the original value in the dataset.As,this dataset contain the missing values so we need to evaluate missing values through various imputation.Lets just find out which method is best for missing value imputation for this dataset.

df[is.na(df)]=mean(df,na.rm=T) #Mean method

df[is.na(df)]=median(df,na.rm=T) #Median method

require(DMwR) #library to impute KNN Imputation for Missing value Analysis

df=knnImputation(df, k=2) #Knn Imputation

sum(is.na(df))

So,Median method value is nearest to the original value.Therefore,for this dataset it is better to use Median method for Missing Value Analysis.

Now,let us determine the independent variables which affect the target(dependent) variable mostly.Now, using Correlation Analysis for Feature Selection.Correlation Analysis is applicable on only numeric data.So,first convert the target variable into numerical variable.

require(corrplot) #library to plot correlation graph

df1$pickup\_datetime=as.numeric(df1$pickup\_datetime)

df1$pickup\_longitude=as.numeric(df1$pickup\_longitude)

df1$pickup\_latitude=as.numeric(df1$pickup\_latitude)

df1$dropoff\_longitude=as.numeric(df1$dropoff\_longitude)

df1$dropoff\_latitude=as.numeric(df1$dropoff\_latitude)

df1$passenger\_count=as.numeric(df1$passenger\_count)

df$pickup\_datetime=as.numeric(df$pickup\_datetime)

df$pickup\_longitude=as.numeric(df$pickup\_longitude)

df$pickup\_latitude=as.numeric(df$pickup\_latitude)

df$dropoff\_longitude=as.numeric(df$dropoff\_longitude)

df$dropoff\_latitude=as.numeric(df$dropoff\_latitude)

df$passenger\_count=as.numeric(df$passenger\_count)

df1$fare\_amount=as.numeric(df1$fare\_amount)

df$fare\_amount=as.numeric(df$fare\_amount)

M<-cor(df)

head(round(M,2))

corrplot(M, method="circle") #plot correlation graph indicating circles

corrplot(M, method="color") #plot correlation graph indicating colours

corrplot(M, method="number") #plot correlation graph indicating numbers

Through this we will know the correlation among dependent and independent variables.

Now,perform various machine learning algorithms on our dataset in order to predict the values for target variable(fare\_amount) in test dataset.

Firstly,start with basic algorithm i.e. Linear Regression Algorithm.

Linear Regression Algorithm is same as Logistic Regression.The only difference is that if the target variable is categorical and continuous then we go for Logistic Regression and if target variable is continuous and numerical then we go for Linear Regression.It can be used for missing value prediction but only for Regression problems.It assumes that their is little or absence of multicollinearity,no outliers and independence of errors and linear relationship among dependent and all independent variables.

require(rpart)

require(MASS)

require(usdm)

require(tibble)

df1=add\_column(df1,fare\_amount=df$fare\_amount,.before="pickup\_datetime")

lm\_model=lm(fare\_amount~.,data=df)

summary(lm\_model)

pred=predict(lm\_model,df1[,2:7])

pred

require(DMwR)

regr.eval(df1[,1],pred)

Now,we will again predict the values and determine the oerformance using Multiple Regression Algorithm.

lm\_model=lm(fare\_amount~.-pickup\_datetime,data=df)

summary(lm\_model)

confint(lm\_model)

pred=predict(lm\_model,df1[,2:7])

pred

require(DMwR)

regr.eval(df1[,1],pred)

Now,we will again predict the values and determine the performance using KNN Analysis Machine Learning Algorithm.This algorithm stores all available cases and classifies new cases based on a similarity measure.It is local heuristic.Can be used both for Regression and Classification data.

require(class) #library to impute knn analysis

KNN\_pred=knn(df,df2,df$target,k=1) #predict knn values

conf\_matrix=table(KNN\_pred,df$target) #build confusion matrix

sum(diag(conf\_matrix))/nrow(df2) #determine accuracy # 0.8733333

conf\_matrix

(12)/(131+12) #determine False Negative Rate #0.07801418

(130)/(130+11) #determine Recall # 0.9219858

Now,from the performance parameters values for Linear Regression and Knn Analysis ML Algorithm. KNN Analysis is better Algorithm than Logistic Regression because FNR is much more lower in KNN Analysis,which makes the model error free.Also,the Accuracy and Recall of KNN Analysis algorithm is higher than Logistic Regression.

Now,we will again predict the values of test data using Random Forest Algorithm for getting much more accurate values.

Naive Bayes is practical learning method.It is a probabilistic classification model.It Bayes theorem of probability to predict the class of unknown dataset.It assumes each independent variable is contributing independently so all probability for each independent variable is found.

library(randomForest)

RF=randomForest(fare\_amount~.,df,importance=TRUE,ntree=500)

pred=predict(RF,df1[,])

conf=table(observed=df[,1],predicted=pred) #build matrix

caret::confusionMatrix(conf) #print confusion matrix

1/(137+1) #determine FNR #0.007246377

(137)/(137+1) #determine Recall #0.9927536

137/(137+1+12+0) #determine Accuracy #0.9133333

Now,among all the 3 Machine Learning algorithms for predicting values RandomForest and KNN Analysis is the best algorithm because FNR value is lowest in RandomForest algorithm i.e.Random Forest reduce the tendency of errors in the model.Also,at the same time,Accuracy and Recall values are also highest for Random Forest Algorithm.This makes Random Forest Algorithm most accurate and efficient Machine Learning algorithm for predicting the values for this dataset.

So,at last save the predicted values by Random Forest model in the target variable for the test set.

df2$target=pred #save the predicted values in test set-target variable.

**NOTE :**

I have explained the outline of the project using only R code. Similarly, the Python code can also be used for this problem.I have attached both the R and Python code files along with my project.