Supervised Machine Learning Models for House Price, Customer Churn and Census Income Prediction

Deepak Kumar Swain, x19216769@student.ncirl.ie, MSc Data Analytics, National College of Ireland

Abstract— **When the economies of countries decrease, various areas like jobs, businesses, supply chain management, hospitality departments get impacted. Out of these scenarios, three areas are covered in this paper are 'House price', 'Customer churn' and 'Annual income' with respect to several exploratory factors. In this project, five supervised machine learning algorithms are used to understand the effect of features on the response variable. For the 'House price prediction', multiple linear regression is used as the considered dependent variable is continuous. For the 'Customer churn' dataset, two classification algorithms, k- Nearest Neighbours and Binary Logistic regression, have been used as the dependent variable is dichotomous in nature. On the other hand, for the third dataset, two more classification models, Decision Tree and Random Forest, have been used. After applying the models on the datasets, necessary evaluations are done on each model with proper justification.**

INTRODUCTION

The year 2020 has become an unforgettable year for the world as it has brought drastic changes in our lives. Apart from life-threatening risk, few more areas are hugely affected during this pandemic such as jobs, house sales, hospitality sectors, customer churns in various businesses. This situation inspired me to dig into various dataset that are related to these areas. The selected datasets for the analysis are:

- House price dataset of the King County which is in Texas.
- Census income dataset on adults in several countries.
- Customer churn dataset for a telecom company called Telco.

These datasets are not directly related to each other. However, it can be interesting to see the various factors that have a positive or negative impact on the responses of the datasets. The objective of this project is to showcase the application of various machine learning algorithms for training the models and predicting the values or class of the responses along with necessary evaluation results.

After the use of necessary machine learning algorithms, the following questions need to be answered clearly with justification.

1. Which features are responsible for explaining and predicting the response in an efficient manner?
2. Are the models fitting well with the respective datasets?
3. Are the model evaluation techniques showing acceptable results?

The rest of the sections in the report will be in the following sequence. In the next section, an overview of related work will be demonstrated which will highlight the work done by researchers on these datasets. After this section, we will be discussing about the methodologies that have been used in the current project with necessary information. The next session will give an overview of evaluating each model to check the goodness of fit with necessary measures, visualization and comparison. The final section is the conclusion part which provides a summary of the entire project. Also, it gives information about the future work than can be done on this project. At the end of the report, references for the project are given.

RELATED WORK

Several research papers have been published which are based on the mentioned datasets.

For the Census Income dataset, a research paper was published by [1] Navoneel Chakroborty and Sanket Biswas in which they used Gradient boosting classifier, Principal Component Analysis, XGBoost and Support vector machin alogithms to predict the income class. With Gradient boosting algorithm they got the maximum accuracy of 88.16 percent. The analysis seemed to be done efficiently. However, there was no clarity on how the categorical variables were handled. It was only mentioned in a single line that the categorical values were handled.

In another paper, published by Vidya Chokalingam[2], used various classification algorithms reached to maximum accuracy of 87 percent. However, It was seen that, the columns 'Age', 'Fnlwt' and 'Hrs_per_week' were filtered with certain condition as part of data preprocession. However, this doesn't seem to be reasonable.

In another paper, published by Chet Lemon[3], ariuos classification algorithms were used to predict the income class. The Decision tree model gave the maximum accuracy which was around 83 percent. 83 percent accuracy is indicating that the model performed well with the data. However, few features like 'relationship' and 'Capital gain/loss' were excluded from the model without any proper justification.

In another paper, published by Jayavarthini[4], Logistic regression and Naïve Bayes models were used to classify the income class. The paper was well explained but there was no proof of model evaluation for any model.

In a paper published by Sumit[5], Logistic regression, Random forest, Gradient Boosting, BernoulliNB Classifier and Support vector classifier were used. Gradient Boostig gave the maximum accuracy of 87 percent. Visual explanation of exploratory data analysis was impressive. However, some details were still not cleared even if those were performed.

For King County house dataset, a research paper was published by Abdallah[6] where regression models were used for analysis and prediction. The R squared values received in the analysis was around 0.72 which was a good score. However, some improvements can also be done on this.

Another paper was published by Jiao[7], where Support Vector Regression along with linear regression was used to analyse the data. The SVR gave a R squared value as 0.86 whereas the Linear regression gave an R squared value as 0.77. However, the linear model can be improved to give better R squared value and low RMSE value.

In another paper published by Stephen[8], Linear regression model was used which gave an R squared value as 0.87. This was an impressive value. However, there was no clear picture of how the data preprocessing was done.

For Telco customer churn dataset, a research paper was published by Chinnu[9] where several models like decision tree, K-means were used. However, necessary justifications were not provided properly.

In a paper published by Abdelrahim[10], classification models like XGBOOST, Random forest, Decision tree were used where XGBOOST gave the optimum accuracy with a value of 84 percent. However, random forest and decision tree models gave accuracy as 79 and 76 percent accuracy respectively. An algorithm like Random forest can cause high performance. With the help of simple models like KNN and Logistic regression, high accuracies can be achieved.

In a paper, published by Ammara[11], different churn prediction techniques were summarized. It gives a high level idea about various churn prediction techniques.

In another paper, published by Kriti[12], models like XGBOOST, Decision tree and Random forest were used to perform analysis and prediction of customer churning. In the paper, the model evaluation factor ROC-AUC has been used which was not mentioned in the above mentioned papers. However, data processing section lacks information.

METHODOLOGY

In the field of data science, various types of methodologies are being followed. Two of the most used methodologies are 'Knowledge Discovery in Databases (KDD)' and 'Cross-Industry Standard Process for Data Mining (CRISP-DM)'. CRISP-DM is considered as a complete cycle of data science project building that involved business and business understanding, data preparation, modeling, evaluation and deployment. On the other hand, KDD is a simpler methodology that does not involve any business understanding and deployment. It involves data selection, pre-processing, transformation, data mining and evaluation. In this paper, KDD approach has been followed as there are no business/functional understanding involved. Also, the final step of the process is model evaluation. So, in the coming session, an attempt is made to demonstrate how KDD methodology is used in the project.

A. ABOUT THE DATASETS (SELECTION):

The 'King County House' dataset[1] has 21597 observations and 21 columns. The dependent variable in this dataset is 'Price' which shows the price of the house.

Column Name	Type
id	Numeric
date	Date Type
price	Numeric(Continuous)
bedrooms	Numeric(Categorical)
bathrooms	Numeric(Categorical)
sqft_living	Numeric(Continuous)
sqft_lot	Numeric(Continuous)
floors	Numeric(Categorical)
waterfront	Numeric(Categorical)
view	Numeric(Categorical)
condition	Numeric(Categorical)
grade	Numeric(Continuous)
sqft_above	Numeric(Continuous)
sqft_basement	Numeric(Continuous)
yr_built	Numeric(Categorical)
yr_renovated	Numeric(Categorical)
zipcode	Numeric(Categorical)
lat	Numeric(Continuous)
long	Numeric(Continuous)
sqft_living15	Numeric(Continuous)
sqft_lot15	Numeric(Continuous)

The second dataset is the customer churn[2] data for the company called Telco. It contains 7043 observations and 21 columns. The dependent variable in the dataset is 'Churn' which is a dichotomous variable that tells whether the customer churning is going to happen or not.

Column Name	Type
customerID	Numeric
gender	String(Categorical)
SeniorCitizen	Factor(Binary)
Partner	String(Categorical)
Dependents	String(Categorical)
tenure	Numeric(Categorical)
PhoneService	String(Categorical)
MultipleLines	String(Categorical)
InternetService	String(Categorical)

[1] https://www.kaggle.com/swathiachath/kc-housesales-data?select=kc_house_data.csv

[2] https://www.kaggle.com/blastchar/telco-customer-churn

OnlineSecurity	String(Categorical)
OnlineBackup	String(Categorical)
DeviceProtection	String(Categorical)
TechSupport	String(Categorical)
StreamingTV	String(Categorical)
StreamingMovies	String(Categorical)
Contract	String(Categorical)
PaperlessBilling	String(Categorical)
PaymentMethod	String(Categorical)
MonthlyCharges	Numeric(Continuous)
TotalCharges	Numeric(Continuous)
Churn	String(Dichotomous)

The third dataset is about census income dataset[3] which contains 20010 observations and 15 variables. The dependent variable is 'Salary' which is a dichotomous variable that tell whether the annual income of an individual is more than $50000 or not.

Column Name	Type
age	Numeric(Categorical)
workclass	String(Categorical)
fnlwgt	Numeric(Continuous)
education	String(Categorical)
education-num	Numeric(Categorical)
marital-status	String(Categorical)
occupation	String(Categorical)
relationship	String(Categorical)
race	String(Categorical)
sex	String(Categorical)
capital-gain	Numeric(Categorical)
capital-loss	Numeric(Categorical)
hours-per-week	Numeric(Categorical)
native-country	String(Categorical)
Salary	Dichotomous (<=50K, >50K)

B. DATA PREPROCESSING AND TRANSFORMATION:

In practical scenarios, we do not come across a dataset which is ready to be used in a model without performing any exploratory analysis. Data preprocessing is one of the most critical part in machine learning analysis. In some scenarios, the model may show good result even without any data preprocessing but that does not mean that the model is a good fit. Before using the feature and response variables in a model, necessary data cleaning is required. Else, the model can behave strangely or throw errors.

I. Dataset: King County House Price

Data Preprocessing:

After having a close look at the dataset, it is observed that the columns 'id' and 'date' were irrelevant data for the model analysis. The id column contains nothing but a unique observation number whereas the date column shows the date parameter. Hence, these two columns were excluded from the analysis.

One of the most critical tasks in the data preprocessing is to check if there are any missing values. Fortunately, there were no missing values present in the dataset. Hence, there was no need to handle the missing values.

One more critical aspect of the data preprocessing is to validate if the features are in the correct form to be used in a model. It was observed that there were few features which were having numeric data (data type numeric or integer) but those were holding categorical values. Hence, the data types of these columns were changed to factor so that it would be easier to create the dummy columns.

But there was a column (Name: waterfront) which was having binary data in it. So, it was natural that the column was showing datatype as factor. But here is a problem. As per the R functionality, dummy columns are created for the columns which are having data type as 'Char' or 'Factor'. As a thumb rule, if a feature has n number of categories (where n >2), we need n-1 dummy columns to explain the feature. For all the factor type features, this was good but in case of waterfront, this will create an issue. To avoid that, the data type of 'waterfront' was changed to integer as binary categories do not need to be transformed into dummy columns. They can directly enter into a linear regression equation. Once the dummies were created, the original columns were excluded from the analysis as these columns would create high collinearity.

After dummy creation, a correlation matrix was created to see which features were having significant relation (p-value < 0.05) with the dependent variable. The columns which were having a p-value more than 0.05 indicated that they were not having a significant contribution in deciding the outcome of the feature. Hence, these columns were removed.

After the data cleaning, the dataset was split into two parts (train and test). The train data was used to build the model whereas the test data was used to do predictions and necessary evaluation.

Data Modelling:

The initial multilinear regression model was applied on the training dataset by regressing the dependent column (Price) with the features. In the model summary, it was found that there were some features which were not having a significant contribution to the model. These columns were excluded from the analysis and the model was applied again. This process was repeated a couple of times until we end off with a model with significant columns only. The next step was to check if there was any multicollinearity among the features. This was checked with the help of 'Variation Inflation Factor (VIF)'. The columns which were having VIF value more than 5 were indicating high multicollinearity.

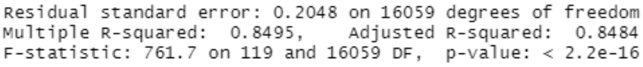
[3] http://archive.ics.uci.edu/ml/datasets/Adult

Hence, these columns were excluded from the analysis and the regression model was applied again.

However, the model was violating certain assumption which will discussed in the next section. To fix this assumption, the regression model was applied by regressing the log of dependent variable (log(Price)) with the remaining features.

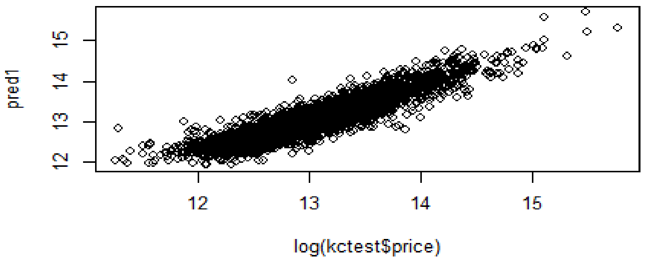
Model Evaluation:

Three important measuring factors are used to validate the performance of a multiple linear regression model. These are 'R squared', 'Adjusted R squared' and 'Root mean squared error'. All these measures range from 0 to 1. For a good model, 'R squared' and 'Adjusted R squared' values should be high whereas the RMSE should be as low as possible. For the final model, 'R squared' value was 0.8495. This means, 84.95 percent of variation in the dependent variable were being explained by the independent variables. This indicates that the model is going well wit the data. However, one drawback in 'R squared' measure is that the number of independent variables in a multilinear regression makes the R squared value larger. To balance this effect, adjusted R squared value is used. The adjusted R^2 increases only when there is any improvement to the model. As we can see in the following details, the R^2 and adjusted R^2 are almost equal.

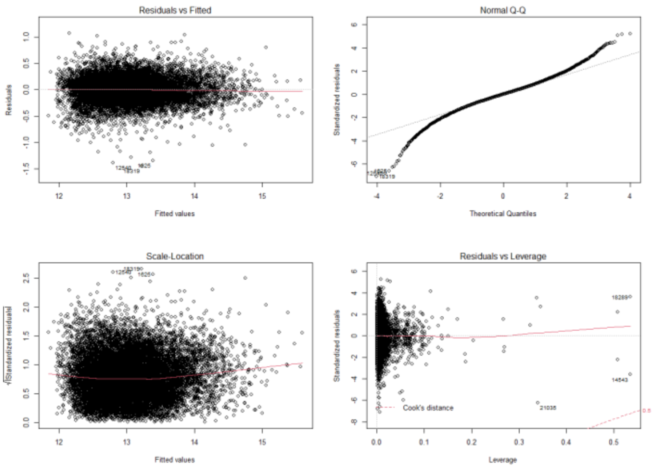


```
Residual standard error: 0.2048 on 16059 degrees of freedom
Multiple R-squared:  0.8495,	Adjusted R-squared:  0.8484
F-statistic: 761.7 on 119 and 16059 DF,  p-value: < 2.2e-16
```

On the other hand, the RMSE value of the model should be as low as possible. For the current dataset, RMSE value came as 0.204 which is good for the model. With the help of these measures, it can be considered that the model is fitting well with the data. Also, we can see from the following plot that there is a linear relationship between the actual value and predicted value. This indicates that the prediction is done with high accuracy.

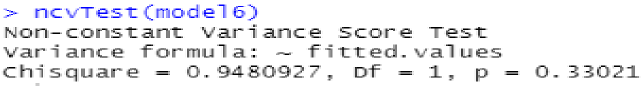


But, we must not forget that linear regression models are built on certain assumption. The model needs to satisfy all the assumption with respect to data point. These assumptions can be validated using diagnostics plots in R.



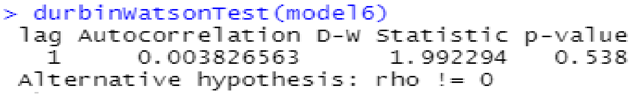
The first assumption is about correct functional form (In this case it is linearity). To satisfy the correct functional form, there should not be any systematic pattern between residuals and fitted values. As we can see in the 'Residuals vs Fitted' plot, the residuals seem to be random. Hence, the assumption was not violated.

The second assumption is that the errors should have constant variance without any signs of heteroscedasticity. To validate this assumption, 'Scale-Location' plot was referred. As we can see in the plot, the squared roots of standardized residual points are not showing any funnel kind of shape (No fan-in or fan-out patterns). In fact, the points are random. Hence, no sign of heteroscedasticity was noticed. Another method of checking the heteroscedasticity is to perform the 'Nc Test' which is a test on non-constant variance. This test gave the p-value as 0.33021 which is more than 0.05. Hence, the assumption of constant variance is not violated.



```
> ncvTest(model6)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.9480927, Df = 1, p = 0.33021
```

The third assumption states that there should be no auto correlation between errors. The Durbin-Watson statistics was used to validate this assumption. If the D-W statistics value is less than 1 or greater than 3, the assumption gets violated. A value close to 2 indicates that the assumption is satisfied.



```
> durbinWatsonTest(model6)
 lag Autocorrelation D-W Statistic p-value
   1     0.003826563      1.992294   0.538
 Alternative hypothesis: rho != 0
```

For the current dataset, D-W statistics value was 1.992294 which is close to 2. Hence, the assumption was not violated.

The fourth assumption states that the errors should be normally distributed. This assumption was validated by one of the diagnostic plots called 'Normal Q-Q' plot which is a probability plot standardized residual against the values that are expected under normality. To satisfy the assumption, the points on the graph should fall on a straight line. For the current dataset, the initial models were violating this assumption. To fix this, a logged version of the response variable was taken in the final model. As we can see in the plot, most of the points are coming in a straight line. We also notice here that there are still some data points which are not coming on the line. But this happens when the model is a perfect model which normally does not happen in real world. Hence, the model still can be considered as a good fit with the data as is showing the R squared value as 0.8495.

The fifth assumption states that there should not be high multicollinearity between the features. This assumption was validated during the modelling process with the help of VIF.

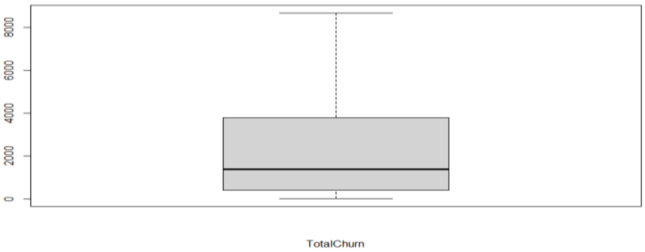
The last assumption states that there should be no influential data points which are having negative impact on the model. To validate this assumption, a measuring parameter called ‘Cook’s Distance’ was used. As we can see in the ‘Residuals vs Leverage’ plot (4th plot in the diagnostic plots), no data points are having cook’s distance more than 1. Hence, the assumption is not violated.

After looking at the performance measures and assumptions, it can be stated that the multiple linear regression model fits well with the King county dataset.

II. Dataset: Telco Customer Churn

Data Preprocessing:

In data preprocessing, one of the most important tasks is to handle missing data if there are any. Three of the most popular techniques are replacing the missing values with mean, median of mode of the column values. For the ‘Customer Churn’ dataset, it was found that the column ‘TotalCharges’ was having 11 missing values. This column was having continuous values that were varying from 18.8 to 248.9. After having a closer look at the column values, it was observed that most of the values were transpiring once or twice. Only a few columns (less than 10) were being repeated around 6 times, but those values (example: 19.7, 19.9,20.05,20.2) were too less as compared to the maximum available value in the column. Also, with the help of whisker plot, we found that there were no outliers. So, it was not suitable to replace the missing values with median or mode. Hence, the eleven missing values were replaced with the mean



In the second step, the column ‘CustomerID’ was removed as it was containing random unique numbers belonging to customer. This column had no meaningful use in the model analysis.

Another important task in data preprocessing is to handle the features with categorical values. For most of the machine learning algorithms, the column values must be in numerical form. Else, the values can not be entered into model equation. For the current dataset, there were some dichotomous columns containing values like ‘Yes’, ‘No’, ‘Male’, ‘Female’. To handle these values, ‘Yes’ and ‘Male’ values were replaced with 1 whereas ‘No’ and ‘Female’ values were replaced with 0. Apart from these dichotomous columns, there were eleven more columns were present with multiple categorical values. Most of these columns were having string values but these values didn’t need to be replaced with numerical values. For these multi-categorical features, dummy columns were created. After creation of the dummy variables, the original columns were removed as those were not required for the model analysis.

Since, the dataset is for classification problem, the dependent variable ‘Salary’ was having dichotomous values (‘Yes’ and ‘No’). But these values needed to be converted to numerical forms. So, the ‘Yes’ and ‘No’ values were replaced with 1 and 0 respectively.

The final task performed in the data preprocessing was to split the data into train and test datasets. As part of this process, 5250 rows out of 7043 were selected as training dataset whereas the remaining rows were used in the test dataset.

Data Modelling:

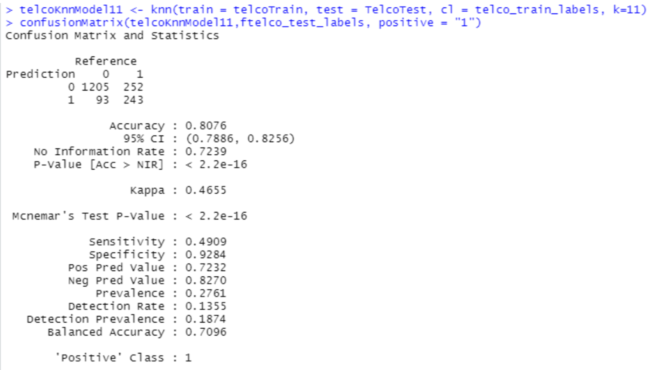
For the Telco customer churn dataset, KNN and Logistic regression models were used. For KNN model, several k values were used to get optimum result whereas in Logistic regression model, insignificant measures were dropped to get the optimum model.

For KNN, K-fold cross validation was also used where the data was split into 10 folds and then the average accuracy was taken for these folds. This was helpful in analysing if the model overfits the data.

Model Evaluation:

Like every model, the first thing that is checked in a model is the accuracy.

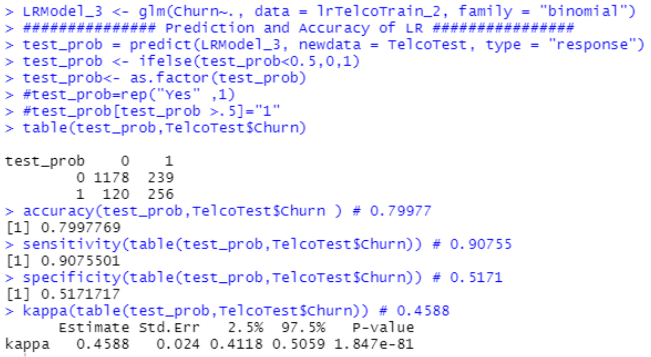
In case of KNN model, several k values were tried to get the optimum accuracy and avoid overfitting. Ideally in KNN model, the optimal value of K is found as the square root of number of observations. For the current dataset, the training data was having 5250 observation. So, the optimal value of K was supposed to be near 72 (square root of 5250 = 72.4). However, the model did not give the best accuracy with k=72. Also, it is not advisable to take a high k value like 72 for a model. To keep the k value as low as possible, the model was run by taking the k values from 3 to 12. It was noticed that for k=11, maximum accuracy was received. Also, the Kappa value was 0.4655 (better than other k values) which indicated the model fitted well with the data.



```
> telcoKnnModel11 <- knn(train = telcoTrain, test = TelcoTest, cl = telco_train_labels, k=11)
> confusionMatrix(telcoKnnModel11,ftelco_test_labels, positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1205  252
         1   93  243
                                          
               Accuracy : 0.8076          
                 95% CI : (0.7886, 0.8256)
    No Information Rate : 0.7239          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.4655          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.4909          
            Specificity : 0.9284          
         Pos Pred Value : 0.7232          
         Neg Pred Value : 0.8270          
             Prevalence : 0.2761          
         Detection Rate : 0.1355          
   Detection Prevalence : 0.1874          
      Balanced Accuracy : 0.7096          
                                          
       'Positive' Class : 1               
```

So the optimum accuracy given by KNN model was 80.76 percent.

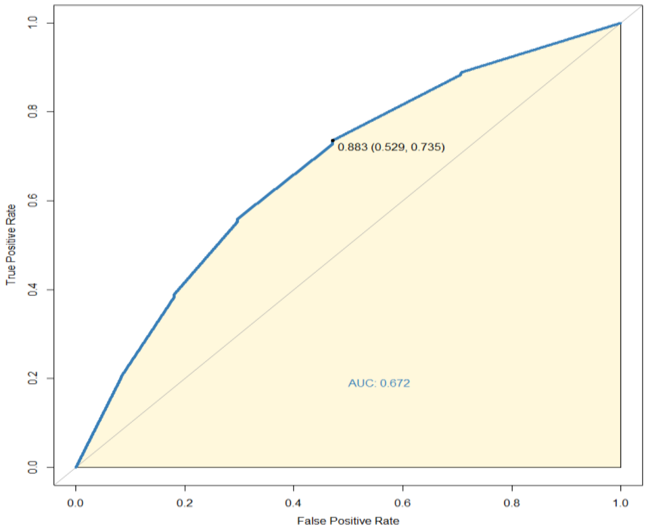
In case of Logistic regression model, the values for accuracy, sensitivity, specificity and kappa were 0.7997, 0.9075, 0.5171 and 0.4588 respectively. As per the above values, it can be stated that the logistic model gives fairly good results is classifying the response.



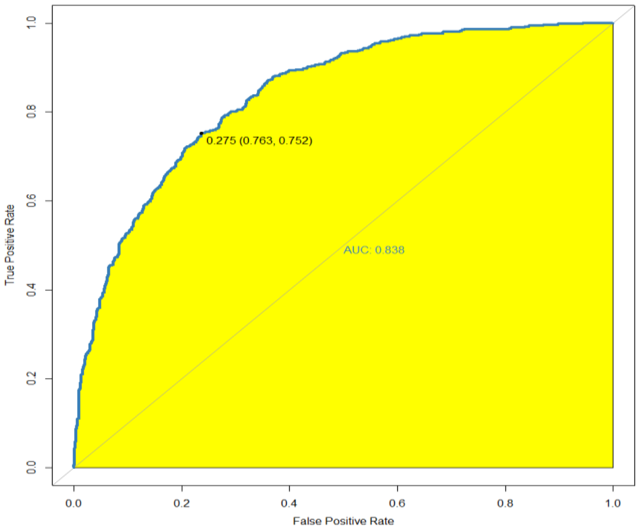
```
> LRModel_3 <- glm(Churn~., data = lrTelcoTrain_2, family = "binomial")
> ############### Prediction and Accuracy of LR ################
> test_prob = predict(LRModel_3, newdata = TelcoTest, type = "response")
> test_prob <- ifelse(test_prob<0.5,0,1)
> test_prob<- as.factor(test_prob)
> #test_prob=rep("Yes" ,1)
> #test_prob[test_prob >.5]="1"
> table(test_prob,TelcoTest$Churn)

test_prob    0    1
        0 1178  239
        1  120  256
> accuracy(test_prob,TelcoTest$Churn ) # 0.79977
[1] 0.7997769
> sensitivity(table(test_prob,TelcoTest$Churn)) # 0.90755
[1] 0.9075501
> specificity(table(test_prob,TelcoTest$Churn)) # 0.5171
[1] 0.5171717
> kappa(table(test_prob,TelcoTest$Churn)) # 0.4588
      Estimate Std.Err   2.5%  97.5%   P-value
kappa   0.4588   0.024 0.4118 0.5059 1.847e-81
```

Apart from above measuring evaluation parameters, ROC-AUC curve was also used to measure the goodness of fit.

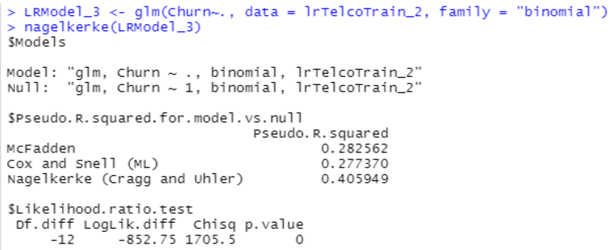
In case of KNN, the ROC and AUC values were 0.883 and 0.672 respectively. As 67 percent of area is covered by the ROC curve, thd model can be considered as a good model for the dataset. However, the area doesn't seem to be very effective.



In case of Logistic regression, the ROC nd AUC values were 0.275 and 0.838 respectively. As we can see here, 83.8 percent of area is covered by the ROC curve. An interesting thing to be noticed here is that the area under the curve is high even if the roc value is low. This explains why the concept of AUC came into picture.



For Logistic regression, there was another factor which is used to measure the goodness of the fit. Like linear regression, logistic regression also has the concept called R squared value. However, the way of calculating R squared value in Logistic regression is different. That's why is being called as 'Pseudo R squared'. There are various method to calculate the pseudo r squared value but one of the most used method is the Nagelkerke method. If the Nagelkerke's pseudo R squared value is more than 0.05, the model can be considered as a good model. As we can see in the following results, Nagelkerke's R squared value came as 0.405 which can be considered as good value.



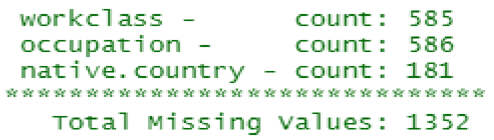
```
> LRModel_3 <- glm(Churn~., data = lrTelcoTrain_2, family = "binomial")
> nagelkerke(LRModel_3)
$Models

Model: "glm, Churn ~ ., binomial, lrTelcoTrain_2"
Null:  "glm, Churn ~ 1, binomial, lrTelcoTrain_2"

$Pseudo.R.squared.for.model.vs.null
                             Pseudo.R.squared
McFadden                             0.282562
Cox and Snell (ML)                   0.277370
Nagelkerke (Cragg and Uhler)         0.405949

$Likelihood.ratio.test
 Df.diff LogLik.diff  Chisq p.value
     -12     -852.75 1705.5       0
```

Overall, the KNN model was showing higher accuracy than Logistic regression model. But, after looking at the ROC curve, we can conclude that Logistic regression model should be considered for the dataset instead of KNN model.

III. Dataset: Census Income

Data Preprocessing:

Like other datasets, the missing values were checked at the very beginning. The dataset was not having any missing values. However, there was a strange value present in three columns that has no meaning. It was denoted as ' ?'.



```
workclass -       count:  585
occupation -      count:  586
native.country - count: 181
**********************************
   Total Missing Values: 1352
```

As this value was not having any significant meaning, it was treated as missing value. The above mentioned columns were having categorical data. One of the suitable method of handling the missing values in a categorical data is to replace it with the mode value. But for the current dataset, it wasn't justified to replace the values with modes. Lets cosider the column ‘occupation’ which has ctegorical values like ‘Armed-Forces’, ‘Farming-Fishing’, ‘Sales’, ‘Tech-support’. All these ctegorical values have sinificant meaning in deciding the income of a person. So, it wasn’t justified to replace the missing values with the modes. Hence, the missing values were replaced with a new category called ‘Unknown’.

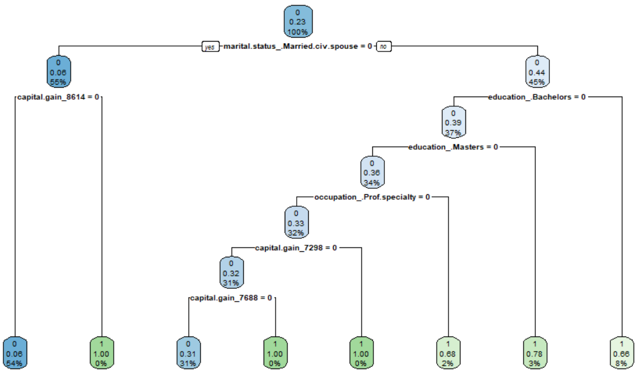
After handling the missing values, the next step was to analyse the dependent variable (Salary). It was seen that the target variable was dichotomous in nature, but it was having values as ‘<=50K’ and ‘>50K’. The value ‘<=50K’ meant that the individual has annual income less than or equals to $50000 whereas the other value was representing the income more than $50000. However, for the model analysis, these values needed to be converted into numerical form. The values ‘>50K’ and ‘<=50K’ were replaced by 1 and 0 respectively.

The features having multiple categorical values were needed to be replaced with dummy columns as they couldn’t directly be used in model equations. But the dummies can be created only for those columns which have data type as ‘char’ or ‘factor’. Hence, the datatypes of necessary columns were changes to factor and then dummy columns were created for respective features. Later the original columns were dropped from the dataset.

The final task performed as part of data preprocessing was to split the data into train and test datasets. 75 percent of the dataset was used in the train dataset to train the models while the remaining 25 percent data were used in the test dataset for prediction and model evaluation purposes.

Data Modelling:

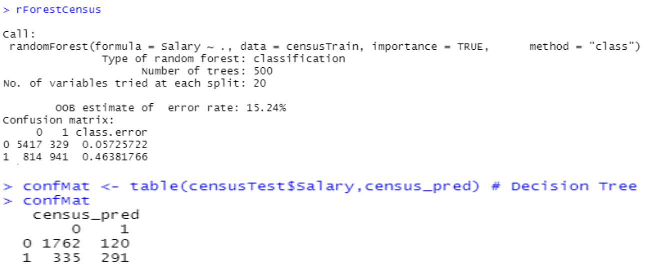
For the census dataset, two classification machine learning models, Decision Tree and Random forest, were used. With the help of ‘rpart’ library decision tree model was applied. For random forest, ‘randomForest’ library was used. One critical thing was noticed while applying random forest algorithm. The random forest algorithm was applying a regression technique even if the method parameter was passed as classification. On further analysis, it was noticed that the dependent variable must be of factor type for a classification. The following plot represents the decision tree for the model.



Decision tree uses some technques to calculate entropy, gini impurity and information gain based on which it decides the root and brances of the tree. Entropy values of each feature are combined to get the information gain which helps in spliting the branches in best possible way. Gini impurity helps in hadling the impure splits(example: One parent node having 2 Yes and 2 No). These techniques are also used by random forest which is an ensemble technique. The main difference in random forest is that it randomly takes subsets of the data and builds multiple decision trees. Then it takes the average of the output of all the decision trees to give the final ouput.

Model Evaluation:

One of the first model evaluation parameter in a classification model is evaluating the accuracy. The accuracies given by decision tree and random forest models were 81.85% and 84.37% respectively. Both the models were performing well but random forest gave slightly better accuracy. However, this was not enough to decide whether a model is a good fit with the data.



```
> rForestCensus

Call:
 randomForest(formula = Salary ~ ., data = censusTrain, importance = TRUE,      method = "class")
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 20

        OOB estimate of  error rate: 15.24%
Confusion matrix:
     0   1 class.error
0 5417 329  0.05725722
1  814 941  0.46381766

> confMat <- table(censusTest$Salary,census_pred) # Decision Tree
> confMat
   census_pred
       0    1
  0 1762  120
  1  335  291
```

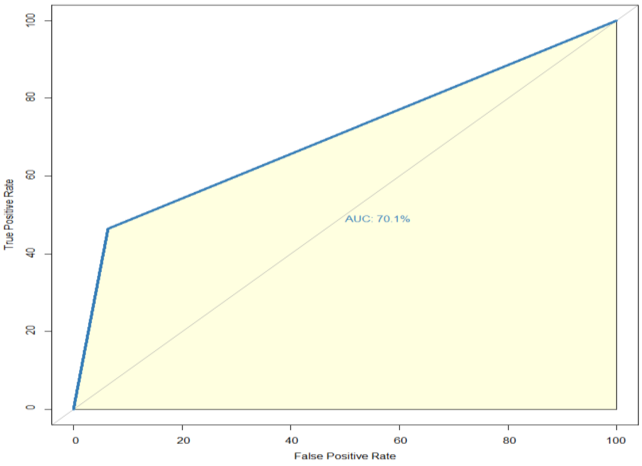
The above figures show the confusion matrices for decision tree and random forest. For decision tree, recall, precision and F-measure values were 0.70, 0.46 and 0.88 respectively. As the F-measure value is more than 0.5, the model is a good fit but not the best one.

For Random Forest, rcall, precision and F-measure values were 0.75, 0.55 and 0.90 respectively. This shows that Random forest performs better than Decision tree model.

Two more factors which are widely used to evaluate the goodness of fit are ‘Receiver Operating Characteristics (ROC)’ curve and ‘Area Under Curve (AUC)’. ROC-AUC curve is widely used for binary classification models to measure how well a model is able to distinguish between classes. The ROC curve is plotted by taking ‘True Positive Rate’ against ‘False Positive Rate’. The percentage of AUC tells how well the model is performing. An AUC value of 0.5 (50 percent) indicates that the model has no capacity performing the classifiation operation fairly. If the area under the curve is more than 0.5, it indicates that the model is performing well with classification. An AUC value of 1 indicates a perfect model.

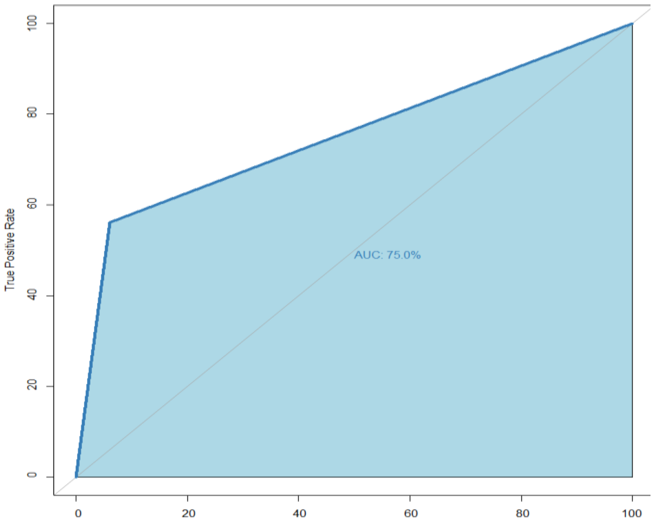
ROC and AUC for Decision Tree:

As we can see in the following ROC-AUC plot, the model is giving and AUC value as 0.701 (70.1 %) which is a resonable value.



ROC and AUC for Random Forest:

In the ROC-AUC plot for Random Forest, the percentage of area under curve was 75 which is reasonably good.



Overall, both the models were performing reasonably well for the classification problem. But there is one more factor which needed to be evaluated. It is always good to validate whether a model is underfitting or overfitting with the data. K-fold cross validation method was used to check if the model is overfitting or underfitting. It was found that neither of the models were overfitting or underfitting.

CONCLUSION AND FUTURE WORK

In summary, we saw that the multiple liner regression model was fitting well with the king county dataset. Also, it was not violating any assumptions based on which the linear regression model is built. As we have seen that the in the model, log of the response variable was taken. Hence, during prediction, it the values need to be converted to the actual value to know the predicted price of the model. To perform further improvements, we can also check the interaction factor in the analysis. If there are any significant interaction factors, they need to be used in the model.

For the Telco customer churn dataset, both KNN and Logistic regression models gave good accuracy. However, in terms of model evaluation, Logistic regression model outperformed KNN model. It will be interesting to see whether any of these models overfit or underfit the data.

For the census income dataset, both Decision tree and Random forest performed good. However, Random forest model was performing slightly better than Decision tree. One problem that was faced for random forest was that it was taking 40 to 60 minutes to run (can run faster in a high-performance system). But the decision tree was taking only 5 minutes to run. If time is crucial, decision tree can be a good candidate for analysis.

REFERENCES

[1] A Statistical Approach to Adult Census Income Level Prediction, Navoneel Chakrabarty, Sanket Biswas

[2] Vidya Chockalingam, Sejal Shah and Ronit Shaw: “Income Classification using Adult Census Data

[3] Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques, Chet Lemon,Chris Zelazo, Kesav Mulakaluri

[4] Analysis and prediction of adult income, C. Jayavarthini, Ishu Todi

[5] Classification Algorithms for the prediction of Income from Adult Census Income Dataset, Sumit Mishra

[6] King County House Prices Prediction Model, Abdallah Alsaqri, Sree Inturi

[7] Housing Price prediction Using Support Vector Regression, Jiao Yang Wu

[8] Comparison of Data Mining Models to PredictHouse Prices, Stephen O’Farrell

[9] Customer Churn Prediction, Ms. Chinnu P Johny,Mr. Paul P. Mathai

[10] Customer churn prediction in telecom using machine learning in big data platform, Abdelrahim Kasem Ahmad,Assef Jafar,Kadan Aljoumaa

[11] A review and analysis of churn prediction methods for customer retention in telecom industries, Ammara Ahmed, D. Maheswari Linen

[12] Customer churn: A study of factors affecting customer churn using machine learning, Kriti