MACHINE LEARNING

ST3189

COURSEWORK

UNIVERSITY
OF LONDON

UOL STUDENT NUMBER: 200618238

Total Pages: 10 excluding cover page, table of content, and references

Table of Contents

1. Introduction 3
2. Country Development (Unsupervised Learning) 3
2.1 Substantive Issue 3
2.2 Methodology 3
2.3 Dataset and Variables 3
2.4 Analysis 4
2.5 Results 6
3. Insurance Charge (Regression) 6
3.1 Substantive Issue 6
3.2 Methodology 6
3.3 Dataset and Variables 6
3.4 Analysis 7
3.5 Results 8
4. Telco Customer Churn Prediction (Classification) 9
4.1 Substantive Issue 9
4.2 Methodology 9
4.3 Dataset and Variables 9
4.4 Analysis 10
4.5 Results 11
5. References 13

1. Introduction

Machine Learning is a technique to understand the datasets, build algorithms, and statistical models to improve the systems' performance on a specific task through training. They are classified into supervised, unsupervised, and reinforcement learning. We will only discuss the first two. With supervised learning, a statistical model is trained to predict target/output variable using provided input variables. Through unsupervised learning, a statistical model is trained on unlabeled data to identify patterns and relationships.

2. Country Development (Unsupervised Learning)

2.1 Substantive Issue

HELP International is a non-governmental organization dedicated to fighting poverty and providing essential amenities to underdeveloped countries. With a funding of $10 million, the organization intends to strategically distribute resources to countries in need. To determine the most pressing areas for financial assistance, it is necessary to identify which countries require aid the most. Therefore, we have established the need to cluster countries based on their socio-economic and health factors to determine their overall development status. Through this approach, we will provide valuable insights into which countries require the most attention and aid from HELP International.

The research questions (RQ) that have been identified for the issue are as follows:

- RQ1: What is the optimal number of clusters for assessing a country's development?
- RQ2: What is the interpretation of the identified clusters?
- RQ3: Which countries are most likely in need of financial assistance?

2.2 Methodology

Principal component analysis (PCA), hierarchical clustering with Euclidean and Manhattan distance, and K-means clustering are machine learning techniques used to analyze the country dataset to understand the pattern and relationship between variables with the aim of clustering countries which need financial help and which not. Silhouette and elbow methods have been used in determining the optimal number of clusters.

2.3 Dataset and Variables

The original country dataset consists of 167 rows and 10 columns. There is no missing value and duplicate data. We remove the country variable (character) to make the data in numerical variables resulting in 167 rows and 9 columns. We keep the outliers as they may represent the country's bad situation and in need of financial help.

Variables	Data Type	Description
Country	Character	Name of the country
Child_mort	Numeric	Death of children under 5 years of age per 1000 live births
Exports	Numeric	Exports of goods and services per capita in %
Health	Numeric	Total health spending per capita in %
Imports	Numeric	Imports of goods and services per capita in %
Income	Integer	Net income per person
Inflation	Numeric	Annual growth rate of the total GDP
Life_expec	Numeric	Average number of years a newborn child would live
Total_fer	Numeric	The number of children that would be born
Gdpp	Integer	The GDP per capita (total GDP/total population)

Table 1 Country Dataset Variables Description

2.4 Analysis

Principal Component Analysis (PCA) is a dimensionality-reduction method and mainly used for variables that are highly correlated. From the country dataset, we plot the correlation matrix to see the relationship between variables. Then, we scaled the dataset for the model to learn and understand the problem easily. The country data analysis involved PCA to find the correlation between variables. From the Figure 1, we select PC1, PC2, and PC3 as their variance are more than 1. PC1 represents “quality of life” since *child_mort, total_fer, life_expec, gdpp,* and *income* load heavily on PC1. PC2 represents “trading condition” since *imports* and *exports* load heavily on PC2. PC3 represents “inflation rate” since *inflation* and *health* load heavily on PC3. They account for 76.14% of the total variance (cumulative proportion) in the data. Since PCA does not show the relationship between variables in each principal component, we will conduct cluster analysis to show a clear understanding of variables in each component using hierarchical clustering and K-mean clustering.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
child_mort	0.4195194	0.192883937	-0.02954353	0.370653262	-0.16896968	0.200628153	-0.07948854	-0.68274306	-0.32754180
exports	-0.2838970	0.613163494	0.14476069	0.003091019	0.05761584	-0.059332832	-0.70730269	-0.01419742	0.12308207
health	-0.1508378	-0.243086779	-0.59663237	0.461897497	0.51800037	0.007276456	-0.24983051	0.07249683	-0.11308797
imports	-0.1614824	0.671820644	-0.29992674	-0.071907461	0.25537642	-0.030031537	0.59218953	-0.02894642	-0.09903717
income	-0.3984411	0.022535530	0.30154750	0.392159039	-0.24714960	0.160346990	0.09556237	0.35262369	-0.61298247
inflation	0.1931729	-0.008404473	0.64251951	0.150441762	0.71486910	0.066285372	0.10463252	-0.01153775	0.02523614
life_expec	-0.4258394	-0.222706743	0.11391854	-0.203797235	0.10821980	-0.601126516	0.01848639	-0.50466425	-0.29403981
total_fer	0.4037290	0.155233106	0.01954925	0.378303645	-0.13526221	-0.750688748	0.02882643	0.29335267	0.02633585
gdpp	-0.3926448	-0.046022396	0.12297749	0.531994575	-0.18016662	0.016778761	0.24299776	-0.24969636	0.62564572

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	2.0336	1.2435	1.0818	0.9974	0.8128	0.47284	0.3368	0.29718	0.25860
Proportion of Variance	0.4595	0.1718	0.1300	0.1105	0.0734	0.02484	0.0126	0.00981	0.00743
Cumulative Proportion	0.4595	0.6313	0.7614	0.8719	0.9453	0.97015	0.9828	0.99257	1.00000

Figure 1 Principal Component Analysis

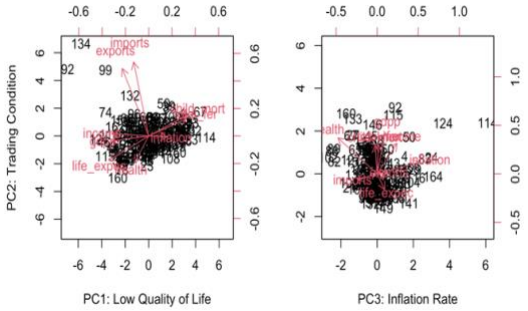

Figure 2 Principal Component Plot

In hierarchical clustering, we use Euclidean and Manhattan distance with four dissimilarity measures, which are single, complete, average, and ward method. From the aggregation coefficient which measures the degree to which the clustering structure is identified (the closer to 1, the stronger the clustering structure), Manhattan Hierarchical Clustering combined with the Ward's Linkage has the highest performance at 97.55%.

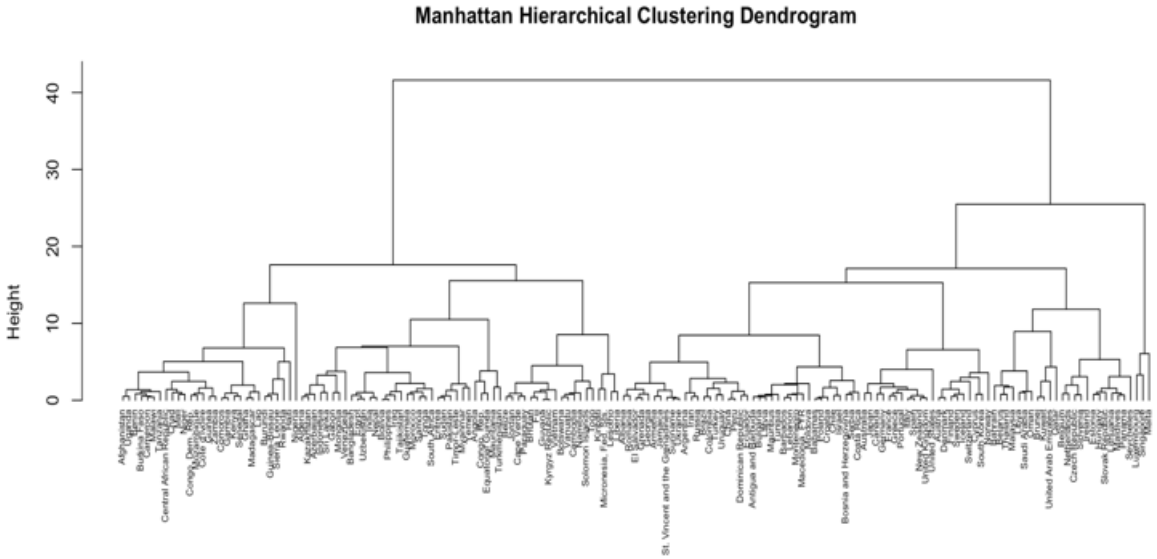

Figure 3 Manhattan Hierarchical Clustering with Ward’s Linkage Dendrogram

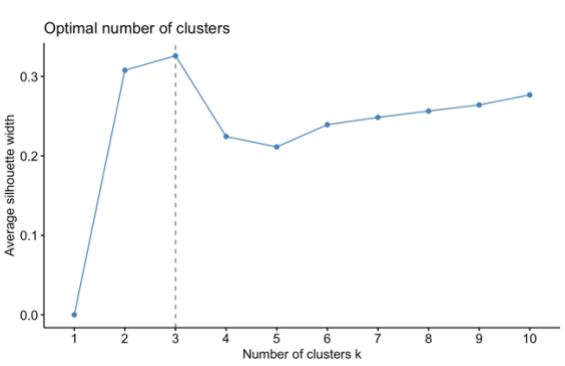

Figure 4 Silhouette Method

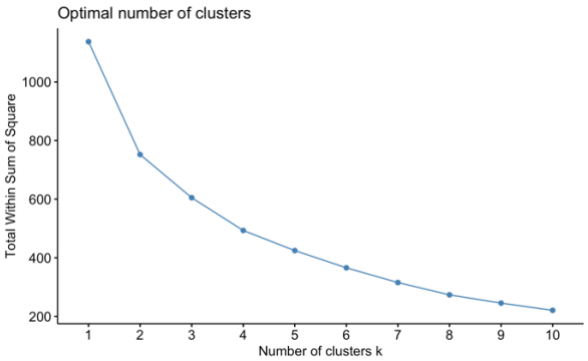

Figure 5 Elbow Method

To determine the optimal number of clusters required, Elbow and Silhouette methods were employed. From both methods, we have determined that forming three clusters is the optimal number of clusters.

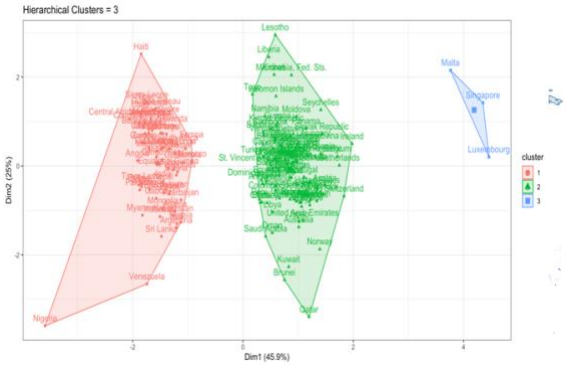

Figure 6 Hierarchical Clusters

Figure 7 Country Clusters by Hierarchical

Hierarchical clustering is a machine learning algorithm that groups similar data points together in a hierarchy. From Figure 6, the 1st cluster (red) represents countries that need help, 2nd cluster (green) represents countries that might need help, and 3rd cluster (blue) represents countries that do not need help. In Figure 7, we present a visualization of the countries clustered in a world map for better visualization. We use PC1 score, which combines factors such as child mortality rate, low income, low GDP, low life expectancy, and high fertility rate, to determine the cluster in greatest need of financial help.

In K-means clustering, we aim to divide n observations into k clusters with the nearest mean. From the sum of squares within groups plot, we decide to divide into 3 clusters. The representation of each cluster in Figure 8 is the same as the hierarchical clustering.

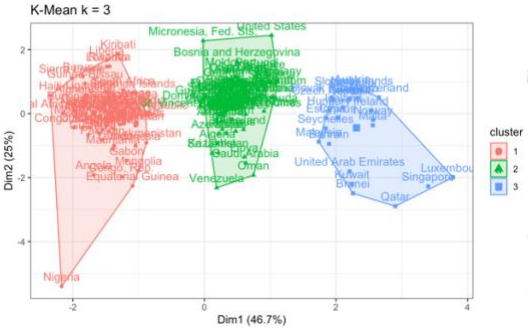

Figure 8 K-means cluster

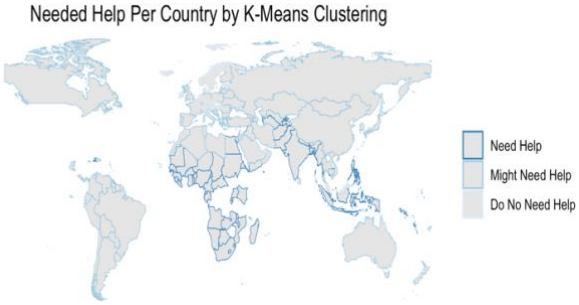

Figure 9 Country Cluster by K-means

2.5 Results

RQ1: What is the optimal number of clusters for assessing a country's development?

- Based on the silhouette and elbow methods, we decided that the optimum number of clusters are three which are “need help”, “might need help”, and “do not need help”.

RQ2: What is the interpretation of the identified clusters?

- From the PCA analysis, variables *child_mort, total_fer, life_expec, gdpp,* and *income* show high loadings in PC1 which clustered into “Quality of Life”. Next, variables *import* and *export* show high loading in PC2 which clustered into “Trading Condition”. Lastly, we clustered *inflation* and *health* into “Inflation Rate” as they are loaded highly in PC3. This analysis has enabled us to interpret the previously identified clusters.

RQ3: Which countries are most likely in need of financial assistance?

- From Figures 7 and 9, we observe that most African countries are clustered under countries that need help. In fact, most African countries are in need for financial help (World Bank, 2021).

3. Insurance Charge (Regression)

3.1 Substantive Issue

Maintaining good health is an essential aspect of life, and health insurance plays a crucial role in providing coverage for necessary health benefits, including treatment for illness and accidents. The profitability of a health insurance company is dependent on collecting more funds than what it spends on the medical care of its beneficiaries. However, predicting medical costs can be challenging, given that a significant portion of expenses arises from rare conditions. Therefore, it is imperative to develop a robust model that can effectively forecast insurance costs based on available data and optimize its performance. The research questions (RQ) that have been identified for the issue are as follows:

- RQ1: Which machine learning model performs the best in predicting charges?
- RQ2: What are the most important factors that influence charges?

3.2 Methodology

Exploratory Data Analysis (EDA) is conducted on the insurance dataset to comprehend the correlation between input and target variables. Multiple Linear Regression, Polynomial Regression, Random Forest, Support Vector Machine (SVM), and Xtreme Gradient Boost (XGBoost) are regression techniques used to predict the insurance charge. Furthermore, we will compare the models’ performance to find the best model and discover the most important variables that influence the prediction of charges.

3.3 Dataset and Variables

The insurance dataset consists of 1337 observations and 7 columns after removing 1 duplicate data, originally 1338 observations. There is no missing value, and the variables contain both numerical and categorical data.

Variables	Data Type	Description
Age	Integer	Age of primary beneficiary
Sex	Character	Insurance contractor gender: female and male
BMI	Numeric	Body mass index
Children	Integer	Number of dependents/children
Smoker	Character	Is the person a smoker or not
Region	Character	The beneficiary's residential area in the US
Charges	Numeric	Individual medical costs billed by health insurance

Table 2 Insurance Dataset Variables Description

3.4 Analysis

Initially, we import and clean the insurance dataset. Exploratory Data Analysis is performed to understand the relationship between variables, and we find that *sex* is not significant in predicting *charges*. From the correlation plot (Figure 10), *smoker* has a high correlation with *charges* and there is almost no correlation between other features with *charges*. We divide the dataset into train set and test set at ratio 70:30 before applying the regression techniques. In multiple linear regression, at first, we use all variables in the model (linear regression 1) and study the variables' significance. As expected and suggested in the exploratory data analysis, *sex* is not significant in predicting charges. The second model (linear regression 2) is formed without *sex*, resulting in all variables becoming significant. Next, we want to transform the target variable to correct the model inadequacies by using log transformation in the third model (log linear regression), resulting in higher adjusted R-squared at 0.7654. From Figure 11, the residual plots show a parabolic trend, which indicates a strong indication of non-linearity in the data.

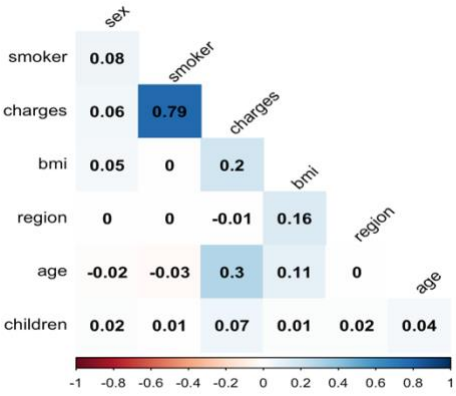

Figure 10 Correlation Plot

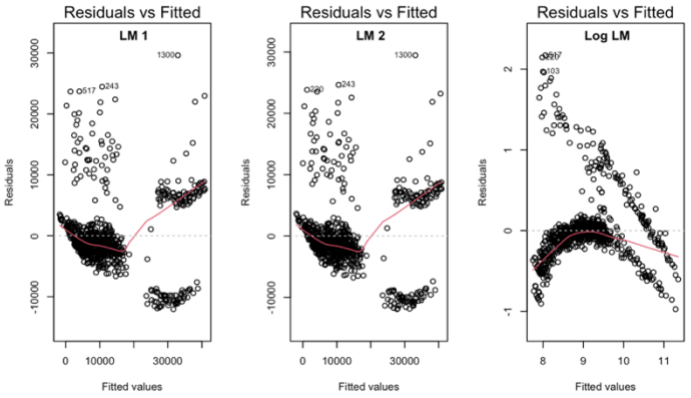

Figure 11 Compare Linear Regression's Residuals

Thus, we choose to employ polynomial regression to enhance the correlation of the other features and improve our model. However, this technique is vulnerable to outliers, and the existence of even a few outliers can have a significant negative impact on the model's performance. Thus, we plan to create a new feature matrix that contains all second-degree polynomial combinations of the features. In the polynomial regression model, we start with all features and use "step" function with backward elimination until only significant variables are left to be used in the model, which are *age, BMI, children, smoker, age², bmi², children², age𝑥region, BMI𝑥smoker, BMI𝑥region, and children𝑥smoker*. The result from this model is that all variables are significant in predicting charges and the adjusted R-squared is even better at 0.8519 than the previous log linear model.

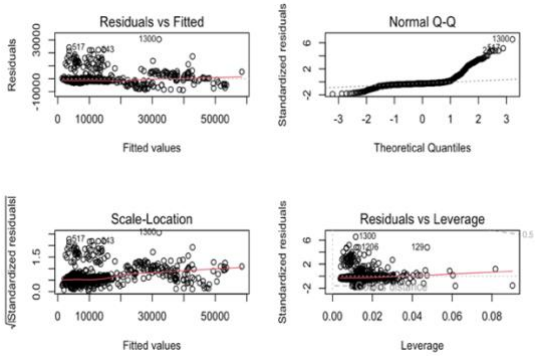

Figure 12 Polynomial Regression Diagnostic Plot

From the diagnostic plot, the Residual vs Fitted plot still indicates a non-linear relationship as it is not equally spread. The Normal Q-Q plot still indicates that the residuals are not normally distributed as they are not on the straight line. The scale-location plot indicates heteroscedasticity as the residuals are not spread equally along the ranges of predictors. The Residuals vs Leverage plot identified some influential outliers beyond the Cook's distance. However, the diagnostic plot's evaluation has improved over the previous regression techniques.

Model	Adjusted R-Squared	RMSE	AIC
Linear Regression 1	0.7540868	6157.362	15476.2490
Linear Regression 2	0.7541757	6152.304	15474.9840
Log Linear Regression	0.7644757	18200.392	945.0203
Polynomial Regression	0.8519360	4958.516	15091.6042

Table 3 Compare Regression Methods

From the table above, polynomial regression plays an important role as it has 85.12% adjusted R-squared, meaning that 85.12% of the variation in charges could be explained by the independent variables. Also, it has the lowest RMSE at 4958.516. However, we would like to explore other techniques such as Random Forest, Support Vector Machine (SVM), and Xtreme Gradient Boost (XGBoost).

Random Forest is a supervised machine learning technique that combines multiple decision trees to improve predictive accuracy and reduce overfitting. In this model, we used 500 trees and 2 variables at each split. The model's RMSE is 4912.366. According to Figure 13, the top three most important variables based on random forest are smoker, age, and BMI.

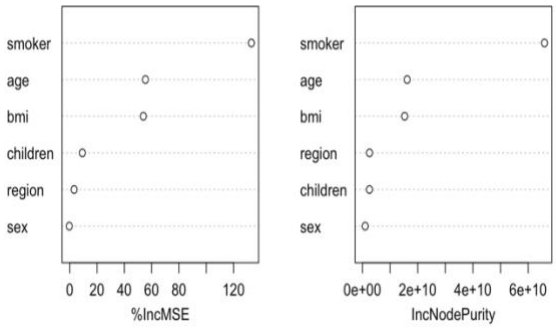

Figure 13 Variable Importance Random Forest

Support Vector Machine model finds the best hyperplane fit that maximizes the margin around predicted values, while minimizing errors and weights assigned to features. It can handle non-linear relationships by transforming features with kernel functions. With number of support vectors of 291, the RMSE of SVM model is 5044.244.

XGBoost combines weak models into a strong predictive model through boosting. It uses regularization to prevent overfitting and can handle missing data and non-linear relationships. It minimizes the loss function with gradient boosting and has achieved state-of-the-art results in many competitions. When training the data, cross validation method is used with ten resampling iterations. The RMSE of XGBoost model is 4767.463 and the top 3 most important variables are smoker, BMI, and age.

3.5 Results

RQ1: Which machine learning model performs the best in predicting charges?

- Comparing the RMSE between the regression techniques employed, we find that XGBoost is the best model with the lowest Root Mean Squared Error (RMSE) at 4767.463. Based on relevant literature review[1], the optimized machine learning algorithms to predict charges are Stochastic Gradient Boosting, XGBoost, and Random Forest respectively. Our analysis shows the same result as XGBoost predicts better than Random Forest, shown from lower RMSE.

[1] Refer to the first reference in the last section.

Model	RMSE
Support Vector Machine	5044.244
Polynomial Regression	4958.516
Random Forest	4912.366
XGBoost	4767.463

Table 4 Regression Model Comparison

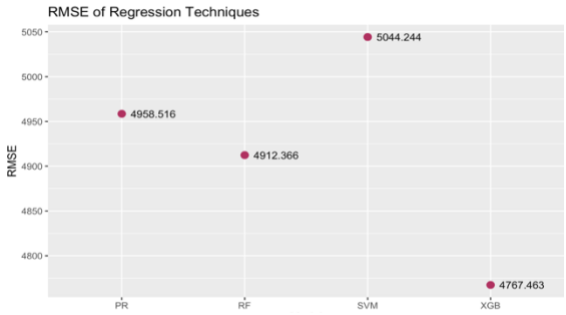

Figure 14 RMSE of Regression Techniques

RQ2: What are the most important factors that influence charges?

- Based on XGBoost and Random Forest, the most important variables in predicting insurance charges are *smoker*, *BMI*, and *age*.

4. Telco Customer Churn Prediction (Classification)

4.1 Substantive Issue

Customer churn, also known as customer attrition, is a crucial factor for companies to monitor as it directly affects their ability to maintain profitability and sustain growth (Gandy, 2019). In the current market landscape, preventing churn and decreasing the churn rate has become an essential aspect of business strategy. However, predicting customer churn can be a complex task due to the variety of human preferences and satisfaction factors. So, creating a reliable churn prediction model using available data is crucial for companies to retain customers and improve their overall business performance.

The research questions (RQ) that have been identified for the issue are as follows:

- RQ1: Which machine learning model performs the best in predicting Telco customer churn?
- RQ2: What are the most important factors that influence Telco customer churn?

4.2 Methodology

Exploratory Data Analysis is performed once more to understand the variables' relationship and distribution. Logistic Regression is used to build a model after one-hot encoding is performed to the variables from the telco customer dataset. Other classification techniques, such as Decision Tree and Random Forest are used to compare the predictive performance and find the best model to predict whether the customers will churn or not. The dataset will be split into train set and test set with a ratio of 70:30 respectively.

4.3 Dataset and Variables

The telco customer dataset consists of 7043 rows and 21 columns. There are 11 missing values which only accounts for 0.156% of the total number of observations. We decide to handle the missing values by removing the 11 rows, resulting in 7032 rows and 21 columns in the final dataset. The variables contain both numerical and categorical data and no outliers for the numerical variables.

Description	Variables
Customers who left	Churn
Basic Information	CustomerID, gender, SeniorCitizen, Partner, Dependents
Services	PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies
Account Information	Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, tenure

Table 5 Telco Customer Dataset Variables

4.4 Analysis

After import and cleaning the telco customer dataset, we check the continuous variables' distribution against customers who churn and those who do not churn, and check their correlations. The correlation plot indicates strong positive correlations between *total charges* and *tenure* at 0.83 and between *total charges* and *monthly charges* at 0.65.

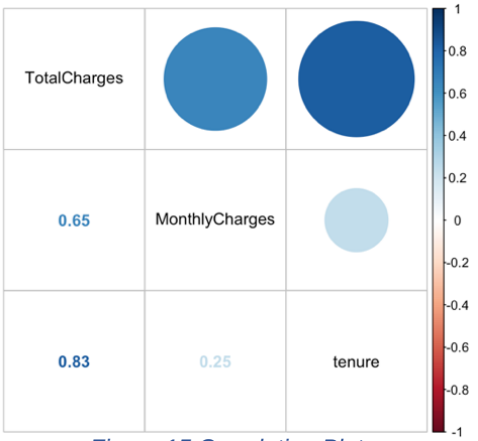

Figure 15 Correlation Plot

Exploratory data analysis is also conducted for categorical variables to determine the variables' importance in predicting churn. It seems that customers with the following traits (senior citizen, do not have dependents/partners, use fiber optic for the internet service, pay with electronic check, use paperless billing, and have shorter contract) have higher churn rate. On the other hand, gender and phone service have low or no influence for the churn rate. The customers who opt-in for *DeviceProtection*, *OnlineBackup*, *OnlineSecurity*, and *TechSupport* experience a lower churn rate. However, the churn rate for *MultipleLines*, *StreamingMovies*, and *StreamingTV* shows no significant disparity between customers who opt-in and those who don't. The overall Telco customer churn rate is 26.58%.

Using logistic regression model, we modify the data into binomial characters and remove the *customerID* variable as it is an identifier variable. Then, we perform one-hot encoding for dummy variables and eliminate the final category of each factor and variables containing "No phone service" and "No internet service" since they lack predictive ability.

After splitting the dataset in a ratio of 70:30 for train set and test set respectively, we check if the churn rate for train set and test set are close to ensure that their distributions are similar and run the model (lrm). Next, we use stepAIC function with backward elimination until significant variables are left (lrm2). Checking the Variance Inflation Factor (VIF) which measures the amount of multicollinearity, *InternetServiceDSL*, *InternetServiceFiber*.optic, and *MonthlyCharges* have large VIF.

From the correlation plot earlier, *MonthlyCharges* and *TotalCharges* are highly correlated. Thus, we removed the *TotalCharges* and *InternetServiceFiber.optic* for the subsequent model (lrm3), resulting in all variables having VIF value lower than 5 which indicates non-multicollinearity. However, the p-value for *StreamingTVNo* and *StreamingMoviesNo* are

more than 0.05. Thus, we removed these 2 insignificant variables and ran the updated model again (lrm4). All variables in the final model became significant.

Area Under Curve (AUC) is a score that shows how well a binary classification model performs, ranging from 0.5 to 1.0 (perfect classification). When using a threshold of 0.5, the training set achieves an accuracy of 0.80 and an AUC of 0.85, while the test set has an accuracy of 0.79 and an AUC of 0.83. Since there is not much difference between the accuracy and AUC in the train set and test set, we can say it is a good model. However, the specificity for both train and test are low at 0.52 and 0.50 respectively.

The Decision Tree Model is capable of processing categorical variables without requiring one-hot encoding. Previous analysis has revealed that *TotalCharges*, *MonthlyCharges*, and *tenure* exhibit strong correlations and consequently might influence the model's performance. Thus, we run the model without *TotalCharges* to avoid multicollinearity resulting in the accuracy of 0.78 and AUC of 0.785.

In Random Forest model, we utilized the identical train set and test set data that were previously applied in the decision tree. We used 500 trees and 4 variables at each split. We did not standardize the dataset because random forest is not sensitive to unscaled data. The accuracy from this model is 0.79 with AUC of 0.818.

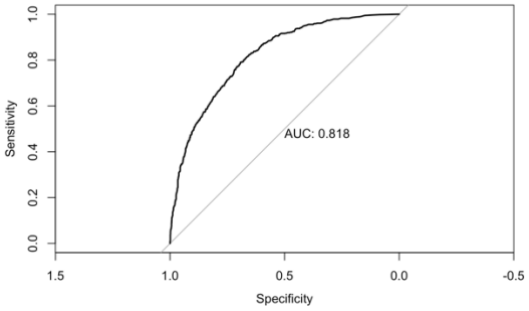

Figure 16 Random Forest AUC

4.5 Results

RQ1: Which machine learning model performs the best in predicting Telco customer churn?

- Based on the ROC plot shown below, the Random Forest model exhibits the highest performance, followed by the Logistic Regression and Decision Tree models. This is indicated by the ROC curves being in closer proximity to the upper left corner. The accuracies for each model are 78.2% for Decision Tree, 78.6% for Logistic Regression Model, and 79.0% for Random Forest. Thus, we can conclude that Random Forest is the best model in predicting Telco Customers Churn as it has the highest accuracy.

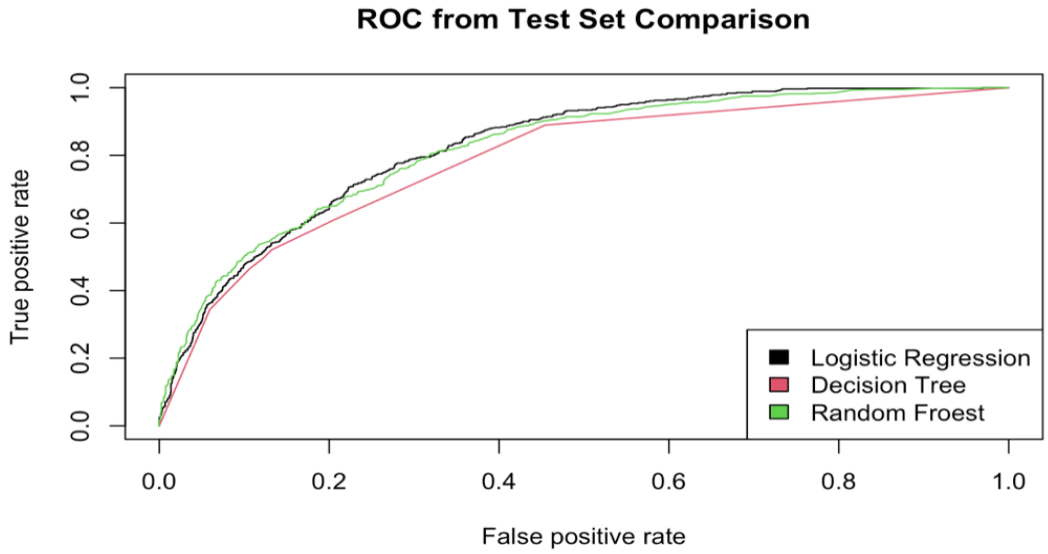

Figure 17 ROC from Test Set Comparison

Model	Accuracy	AUC
Logistic Regression	0.7856148	0.8249681
Decision Tree	0.7819026	0.7851893
Random Forest	0.7902552	0.8184854

Table 6 Model Comparison with Test Set

RQ2: What are the most important factors that influence Telco customer churn?

- According to the Random Forest model, which is considered the best model, *TotalCharges*, *MonthlyCharges*, *Contract*, *year*_tenure, and *PaymentMethod* are the most influential variables in predicting Telco Customer Churn, while other variables have a relatively low impact, as shown in Figure 18.

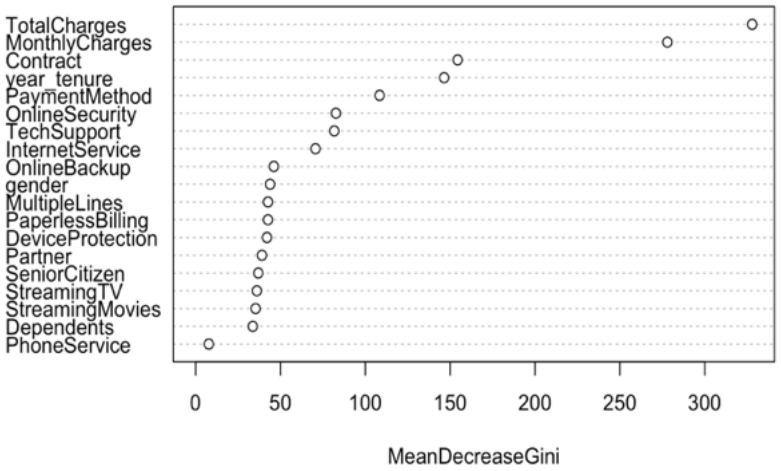

Figure 18 Variable Importance (Random Forest)

5. References

Hassan, C.A.ul et al. (2021) A computational intelligence approach for predicting medical insurance cost, Mathematical Problems in Engineering. Hindawi. Available at: https://www.hindawi.com/journals/mpe/2021/1162553/ (Accessed: March 30, 2023).

Narkhede, S. (2021) Understanding AUC - roc curve, Medium. Towards Data Science. Available at: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (Accessed: March 30, 2023).

World Bank. (2021). Poverty and Shared Prosperity Report 2021: Reversals of Fortune. Retrieved from https://openknowledge.worldbank.org/handle/10986/35996

Gandy, A. L. (2019). Customer churn: How to identify it, and what to do about it. Harvard Business Review. Retrieved from https://hbr.org/2019/11/customer-churn-how-to-identify-it-and-what-to-do-about-it.