**Analysis of Churn Rate in Telecom Company**

Sharon Appoline Rosary

**Introduction**

A crucial indicator of client happiness is the churn rate. When churn rates are low, customers are content; when they are high, customers are defecting. Forbes claims that acquiring new clients is five times more expensive for a corporation than maintaining current ones. Because it shows us how many current customers are leaving the company, reducing churn has a hugely positive influence on revenue streams.

In the telecommunications sector (wireless and cable service providers, satellite television providers, internet service providers, etc.), the churn rate is crucial because it reveals the quality of the company, demonstrates customer satisfaction with the good or service, and enables comparison with rivals to determine an acceptable level of churn. In order to anticipate the customer churn rate, we will examine customer data from a telecoms business, Telco.

**Business Objective**

The goal of the business is to develop classification models that can predict the probability of customer attrition. By doing this, we may ascertain whether the client plans to remain with the company or not. The 'churn' variable found in the dataset is the one we are attempting to predict.

**Tools and Packages**

In order to implement the project tools including Anaconda (for Python programming in Jupyter Notebooks) and Tableau (for creating the dashboards). Additionally, packages including numpy, pandas, seaborn, matplotlib, plotly and sklearn were also used.

**Dataset Overview**

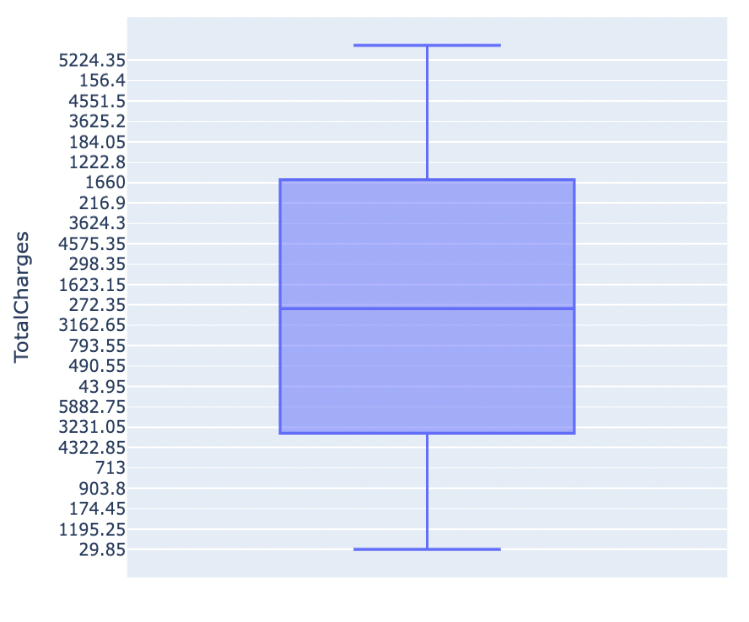
The IBM Developer Platform provided the customer churn statistics for the dataset. The dataset includes 21 characteristics and 7043 instances. 18 of the variables are categorical in nature, and 3 are of a numerical kind. Customer demographics, services each consumer has subscribed to, account information, and our target feature indicating whether the customer left within the last month are all included in this data.

Table, calendar

Description automatically generated with medium confidence

The variables in the dataset are:

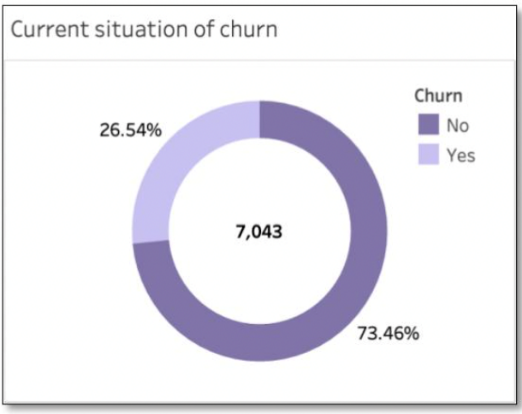
* CustomerID: Customer ID unique for each customer
* gender: Whether the customer is a male or a female
* SeniorCitizen: Whether the customer is a senior citizen or not (1, 0)
* Partner: Whether the customer has a partner or not (Yes, No)
* Dependent: Whether the customer has dependents or not (Yes, No)
* PhoneService: Whether the customer has a phone service or not (Yes, No)
* MultipleLines: Whether the customer has multiple lines or not (Yes, No, No phone service)
* InternetService: Customer’s internet service provider (DSL, Fiber optic, No)
* OnlineSecurity: Whether the customer has online security or not (Yes, No, No internet service)
* OnlineBackup: Whether the customer has an online backup or not (Yes, No, No internet service)
* DeviceProtection: Whether the customer has device protection or not (Yes, No, No internet service)
* TechSupport: Whether the customer has tech support or not (Yes, No, No internet service)
* StreamingTV: Whether the customer has streaming TV or not (Yes, No, No internet service)
* StreamingMovies: Whether the customer has streaming movies or not (Yes, No, No internet service)
* Contract: The contract term of the customer (Month-to-month, One year, Two years)
* PaperlessBilling: The contract term of the customer (Month-to-month, One year, Two years)
* PaymentMethod: The customer’s payment method (Electronic check, mailed check, Bank transfer (automatic), Credit card (automatic)
* Tenure: Number of months the customer has stayed with the company
* MonthlyCharges: The amount charged to the customer monthly
* TotalCharges: The total amount charged to the customer
* Churn: Whether the customer churned or not (Yes or No)

**Data Cleaning**

First the dataset is checked for Null values and no records contained null records. Next, the distribution of the Finished Area is plotted to determine imputation strategy. The boxplot in the figure illustrates the same. The imputation is executed using median values. Also, records with ’No internet service’ in churn data are replaced with 'No.’ Duplicate records are also validated.

**EDA (Exploratory Data Analysis)**

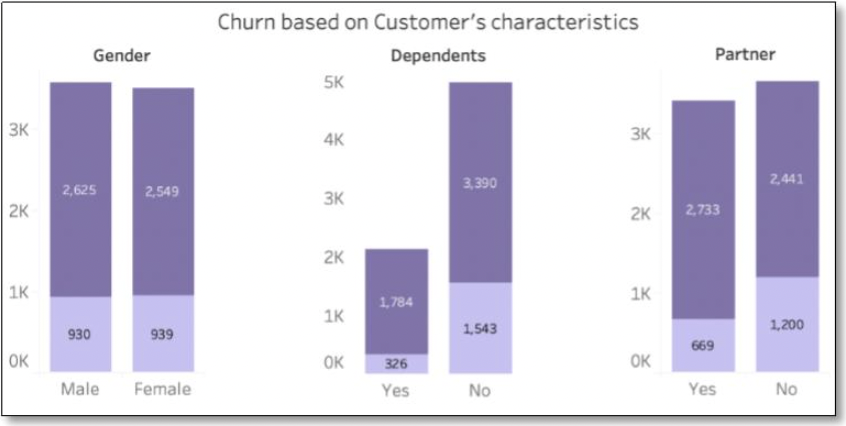
The Customer Churn Analysis includes graphs that show the Customer Churn Rate, Customer Churn based on Customer Characteristics, Customer Churn based on Services Provided, Customer Churn based on Contract, and Customer Churn based on Total Charges. This dashboard's insight enables more straightforward and efficient examination of customer churn and will aid in monitoring and improving client retention.

***Churn Rate***

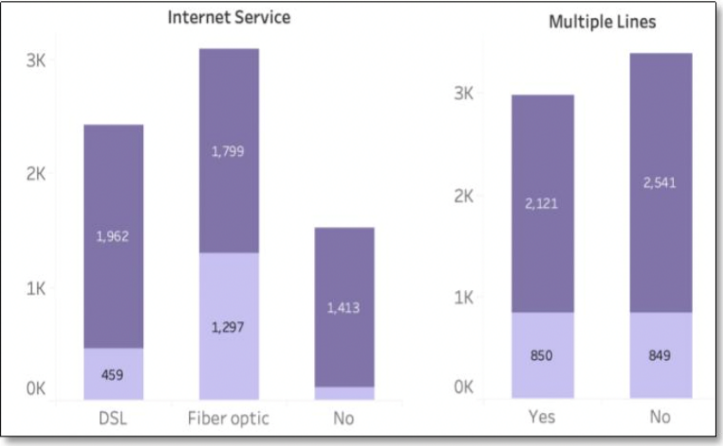
The graph in the image was made to provide a panoramic view of the complete Churn statistic. Only 26.54% of the consumers are classified as having positive churn, according to the data. The attrition rate of clients is frightening.

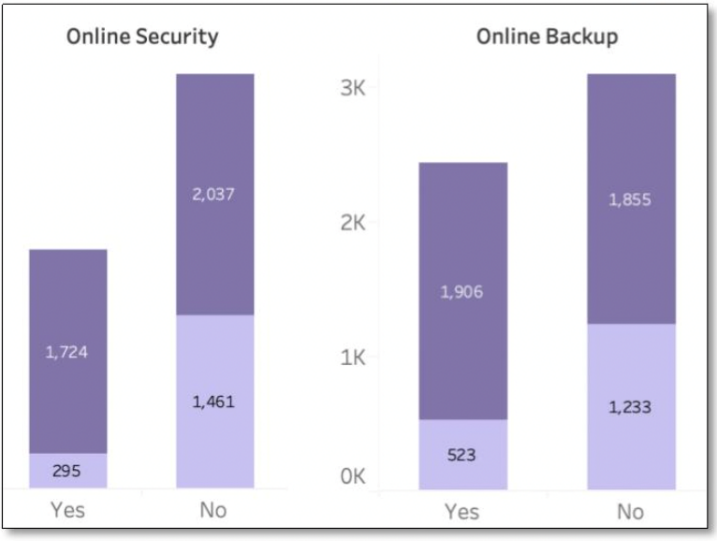
***Churn Based on Customer Characteristics***

The following graph offers a clearer picture of customer-based attrition. The gender of the

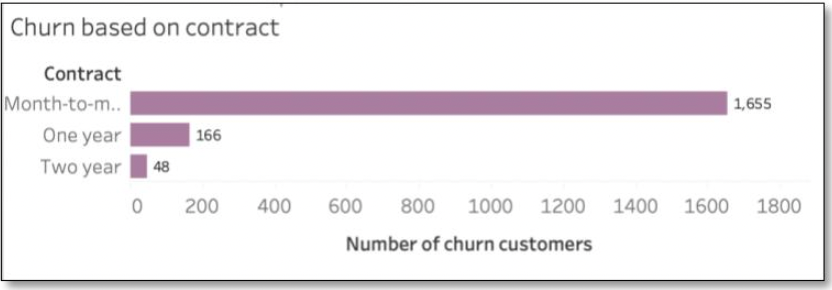
customer has no bearing on the churn rate, as can be shown in the illustration. However, just 15% of customers with dependents fall into the category of positive turnover, compared to 31% of customers without dependents. Similarly, compared to 32.9% of consumers without partners, only 19% of customers with partners have kept using the services.

***Churn Based on Services Provided***

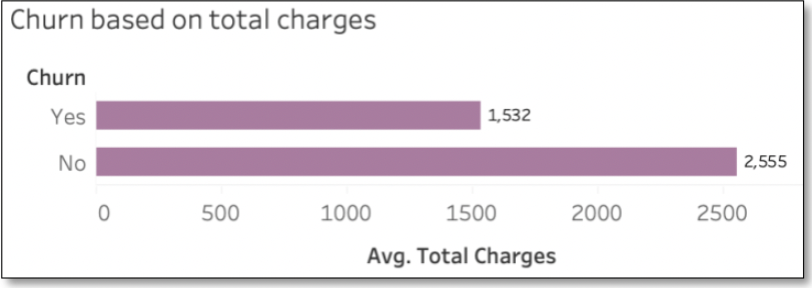
The analysis of the clients' usage of the services comes next. Customers who use fiber optic internet services churn at a higher rate than those who do not, who churn at a lower rate. Additionally, consumers that use multiple lines are retained at a marginally greater rate.

Users that have signed up for Online Backup and Online Security have exceptionally low churn rates, with Online Security customers having the lowest churn. Therefore, while offering promotions and sales deals or executing other marketing techniques, greater attention should be paid to improving these services.

***Churn Based on Contract***

According to the graphs on contract-based turnover, consumers who chose monthly contracts have a far greater rate of churn than those who opted for a 1- or 2-year contract. Given the bigger sum that must be paid upfront for long-term contracts, this may be because there are more users in the monthly arrangement.

***Churn Based on Total Charges***

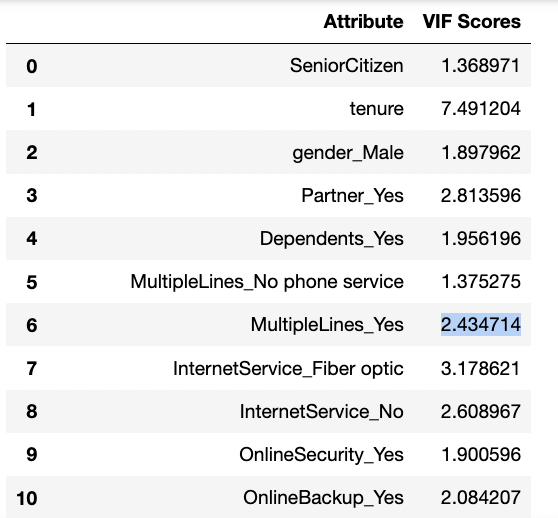
Similar insights can be gained from the graphs plotted to understand churn based on total charges. High negative churn rates among customers who pay more validate the need for competitive service pricing.

Graphical user interface, chart

Description automatically generated***Correlation Plot***

Correlation is the most important relationship that must be examined to develop a solution for the business aim. In order to execute basic Data Cleaning and Integrity Validation, the dataset is loaded. Figure depicts the correlation plot in detail. Most of the variables in the dataset are categorical, hence 30 features in total were evaluated after one-hot encoding was used. The correlation plot shows that, in comparison to other services, customers who used fiber optic service had a stronger positive connection to Churn-Yes. Additionally, it was discovered that Fiber Optic Service and Monthly Charges had a strong positive link, demonstrating the need for this product to be wisely priced. Also, there is a stronger negative link between Churn Yes and Customers who do not use internet services. Therefore, this issue must also be addressed.

**Data Preprocessing**

To prepare the data for modelling, the feature ‘churn’ is converted from categorical value to Binary value. Next, the categorical values are One-hot encoded. In order to check for multi-collinearity, the VIF (Variance Inflation Factor) scores are computed.

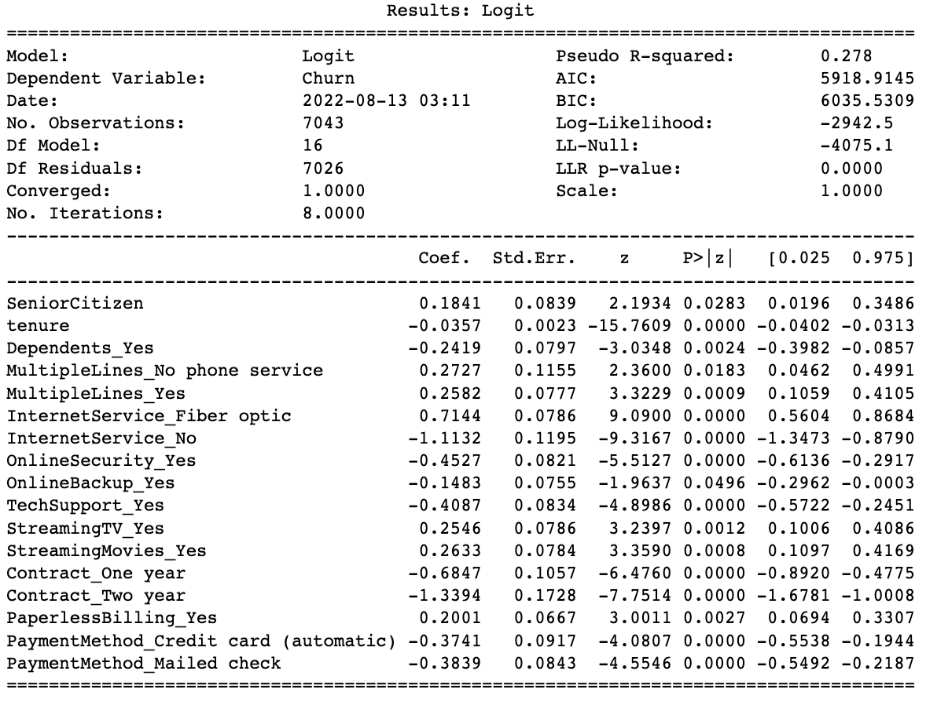
The features with VIF scores above 10 are deleted and the VIF scores are recomputed as shown in the figure.

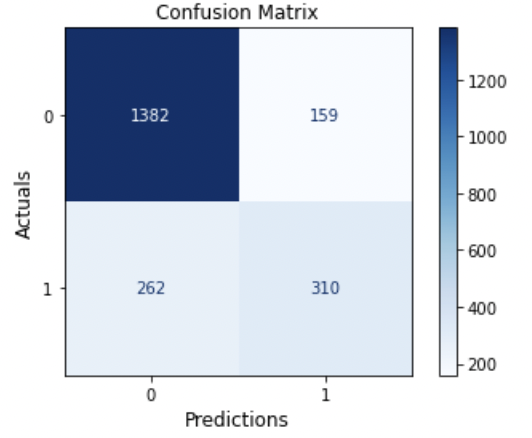
Before creating the model, we must first transform the data into train and test datasets. To build the test and train datasets, we employed a 7:3 ratio (70% training data and 30% test data). The dependent variable is response in our model.

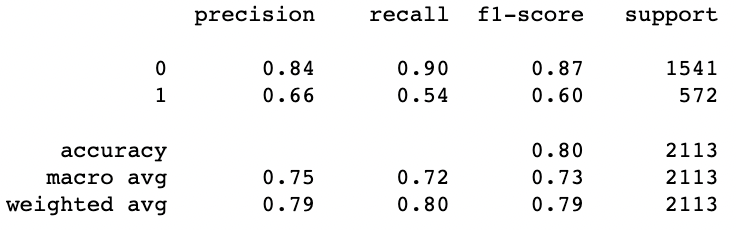
**Data Modelling**

To learn more about the influence of churn rate, we will build Logistic Regression model, Random Forest model and a decision tree model.

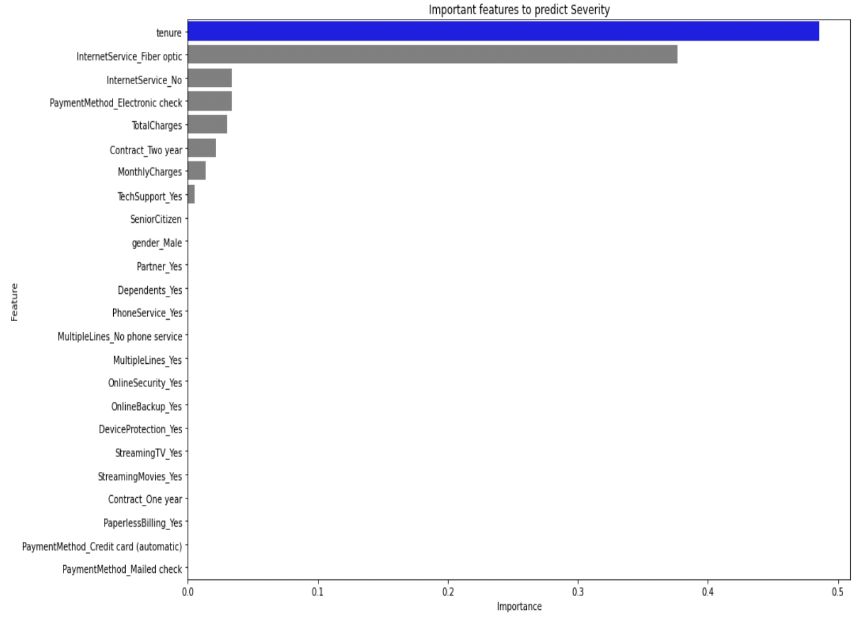
***Logistics Regression Model***

The Logistic Model is then built with the selected variables. The summary of the model is as shown in the image. Internet\_Service\_Fibreoptic, StreamingMovies\_Yes, MultipleLines\_Yes

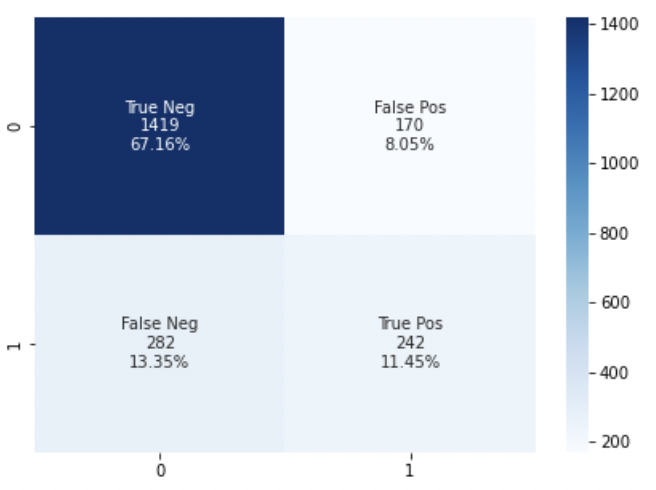
and PaperlessBilling\_Yes have a great significance with Positive influence (in the given respective order) on the target variable ’Churn.’ Similarly, ContractTwoYear, InternetService\_No, ContractOneYear, OnlineSecurity\_Yes, Techsupport\_Yes, Payment\_Method\_Mailed check, Payment\_Method\_Credit card, and tenure also have high significance with Negative influence. The model has an accuracy of 80% and an RMSE (Root Mean Square Error) value of 0.44.  
The confusion matrix plotted for the model is as shown in the figure. The model has 159 False Positives and 262 False Negatives. Here, efforts must be focused on reducing the False Positives.

The model has a precision of 0.66 and a recall of .54 as shown in the figure.

The ROC (Receiver Operating Characteristics) of the given model is shown in the figure and has a significant AUC (Area Under the Curve) of 84.68%

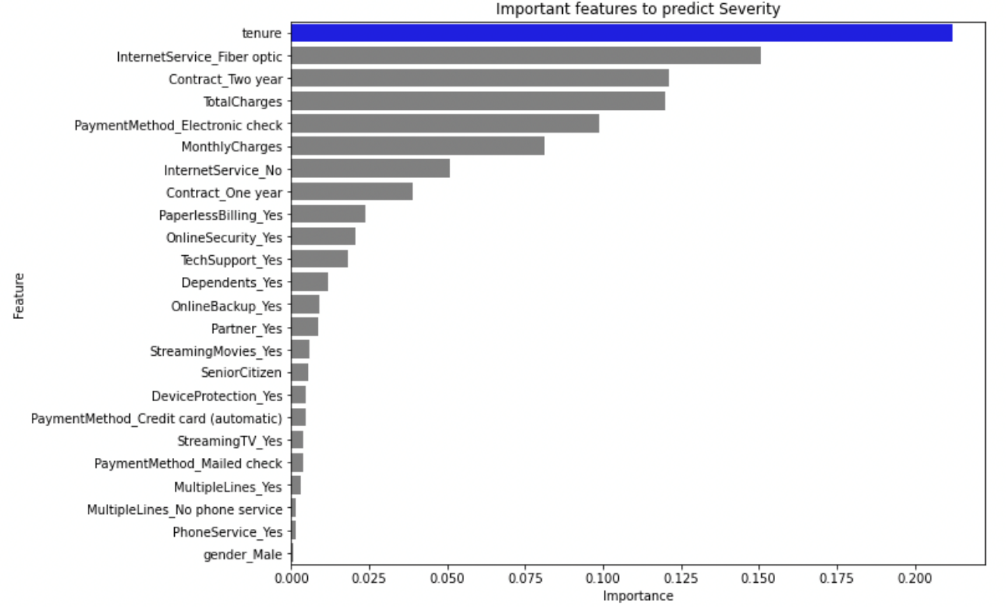
***Decision Tree Model***

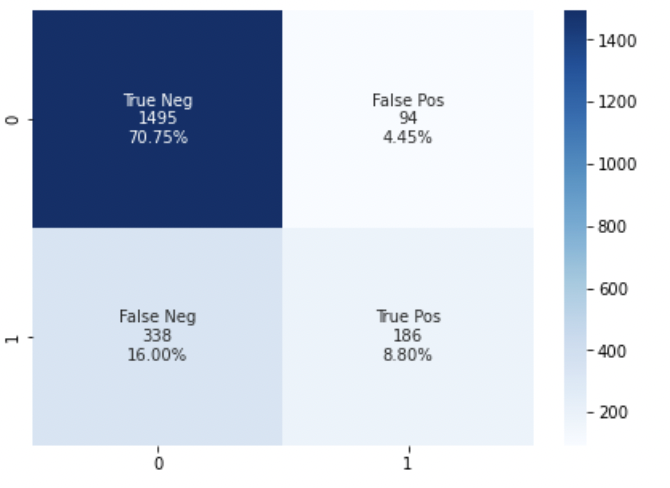
The crucial features as determined by the Decision Tree model are shown in the adjacent figure. Tenure is observed to be the most significant feature. The model has an accuracy of 78.6% and has a precision of 0.59.

The confusion matrix of the Decision Tree Model is as shown in the plot and has a False Positive of 170 and a False Negative of 282.

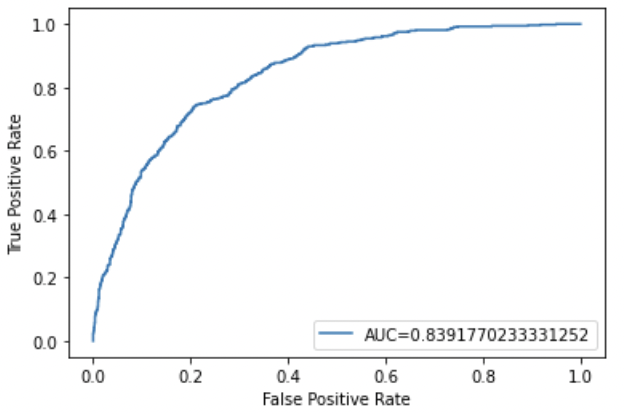
##### The ROC of the model is as shown in the plot and has an AUC of 81.25%.

***Random Forest Model***

The crucial features as determined by the Random Forest model are shown in the adjacent figure. Tenure is observed to be the most significant feature.

The model has an accuracy of 79.5% and has a precision of 0.66.

The confusion matrix of the Decision Tree Model is as shown in the plot and has a False Positive of 94 and a False Negative of 338.

The ROC of the model is as shown in the plot and has an AUC of 83.91%.

**Conclusion**

Other telecommunication services have been replaced by internet-based services. Consumers frequently explore all available options due to the need for high-speed internet services from both residential and commercial customers, as well as the emergence of numerous businesses to offer these services at lower costs. As a result, customer retention in this industry is a monumental undertaking and churn rates are extremely erratic.

The following are the key lessons learned via analysis and dashboards for clients to reduce churn rates: 1. Make new plans available to clients with partners and dependents 2. Improve the online backup and security services while concentrating on clients who utilize Internet services, particularly the fiber optic service (Combo offers could prove to be useful) 3. Develop aggressive monthly pricing strategies to outperform competition and keep long-term clients.

The insights have been validated by the data models created. In addition to this the payment methods should also be upgraded for better customer experience. Also, it is observed that the Random Forest Model is the best in predicting the Churn while the Logistic Regression Model is better in determining the features importance or the significance of the features.

**References**

An Emphasis on the Minimization of False Negatives/False Positives in Binary Classification. (2021). Retrieved 20 June 2022, from [https://medium.com/@Sanskriti.Singh/an-emphasis-on-](mailto:https://medium.com/@Sanskriti.Singh/an-emphasis-on-) the-minimization-of-false-negatives-false-positives-in-binary-classification-9c22f3f9f73

Better Heatmaps and Correlation Matrix Plots in Python. (2021). Retrieved 20 June 2022, from <https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-> 41445d0f2bec

Mandrekar, J. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal Of Thoracic Oncology*, *5*(9), 1315-1316. doi: 10.1097/jto.0b013e3181ec173d

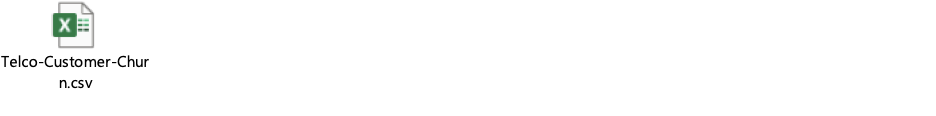
Płoński, P. (2020). Random Forest Feature Importance Computed in 3 Ways with Python. Retrieved 20 June 2022, from <https://mljar.com/blog/feature-importance-in-random-forest/>

Random Forest in Python. (2018). Retrieved 20 June 2022, from <https://towardsdatascience.com/random> -forest-in-python-24d0893d51c0

sklearn.metrics.RocCurveDisplay. (2022). Retrieved 20 June 2022, from <https://scikit-> learn.org/stable/modules/generated/sklearn

*Telco customer churn*. (2020, May [Www.Community.ibm.Com](http://Www.Community.ibm.Com).<https://community.ibm.com/accelerators/catalog/content/Telco-customer-churn>

Telco customer churn Dataset



Understanding Confusion Matrix. (2021). Retrieved 20 June 2022, from <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>

Zach, V. (2022). How to Fix: numpy.linalg.LinAlgError: Singular matrix - Statology. Retrieved 20 June 2022, from <https://www.statology.org/python-numpy-linalg-singular-matrix/>

**Appendix**

* Tableau Dashboard

Graphical user interface, application, Teams

Description automatically generated