Analysis on a Dataset from Kaggle
by Aisha Aslam

Requirement:

A telco in island C is facing churn issue and would like to know following:

1. Why are customers churning?
2. What offers could be made to retain them?
3. Build a predictive model using tool of your choice to predict churners:
    1. Please share document high lighting how you have built model
    2. How you have measured accuracy
    3. What is lift of your model (along with validation methdology)

Understanding of Problem:

The dataset happens to represent users of a telecommunication company that subscribed to different calling plans, spent different times of days and amounts talking on their respective phones and either left or stayed with the company. The company is trying to figure out why customers are leaving and what would make them stay.

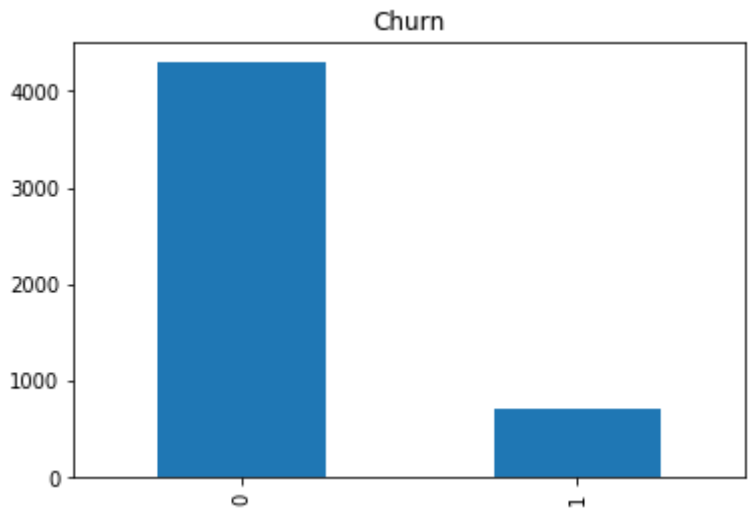
Understanding of Data:

The Dataset consists of a sample set of 5000 entries, with 21 attributes.
The 'Class' column is a boolean that represents churn (1) or retainment (0).
Structured Data.
Imbalanced Dataset (85%-15%).
Static (non-changing) Data.
Numerical Data.
No Missing or Null Entries.
2 Dimensional Data.
Predicted field is a Binary Categorical Feature
1 Dependent Variable: Class (Churn)
20 Independent Variables

Pulling Insights and Correlations from Data:

-Number of Unique users (by phone_number): 5000

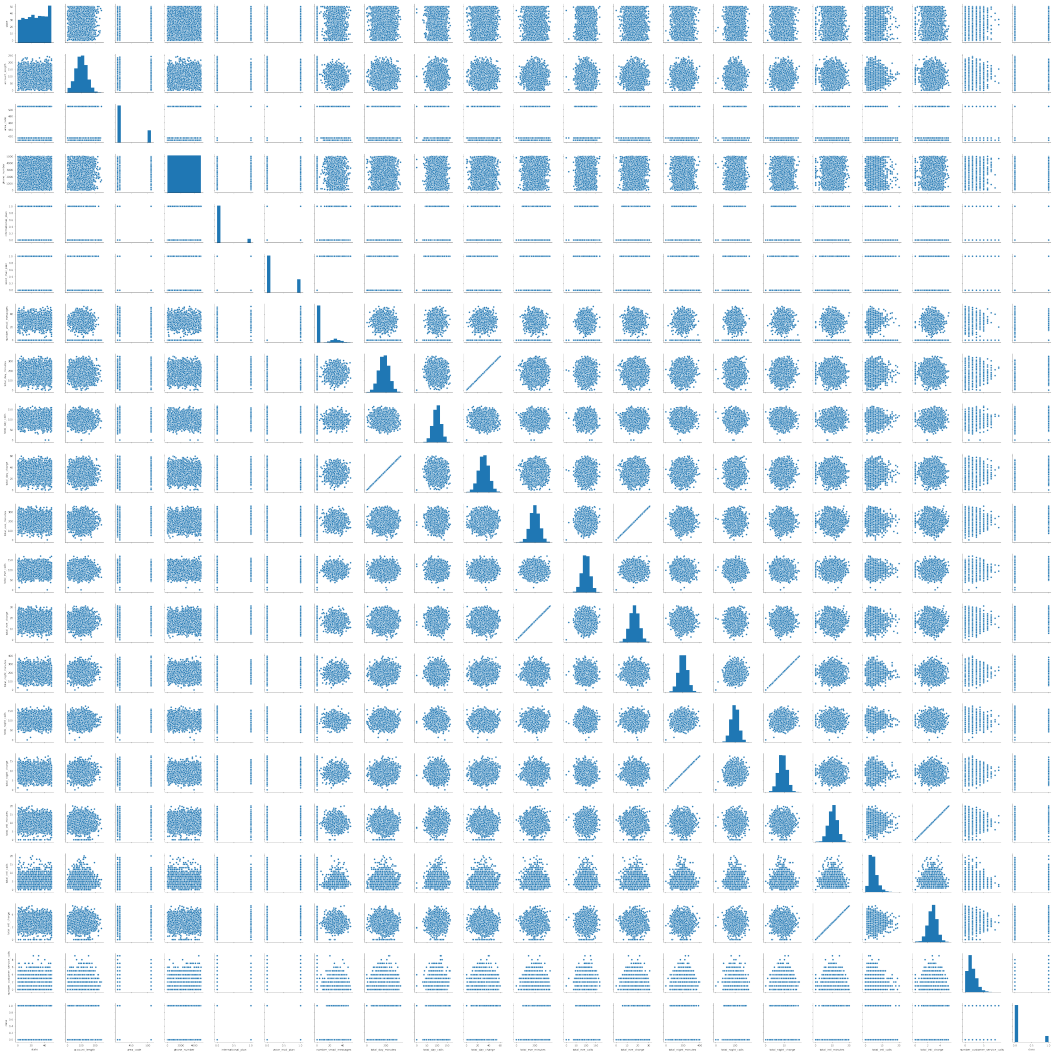
-Number of non-churned users = **4293** (85.9%)
-Number of churned users = **707** (14.1%)



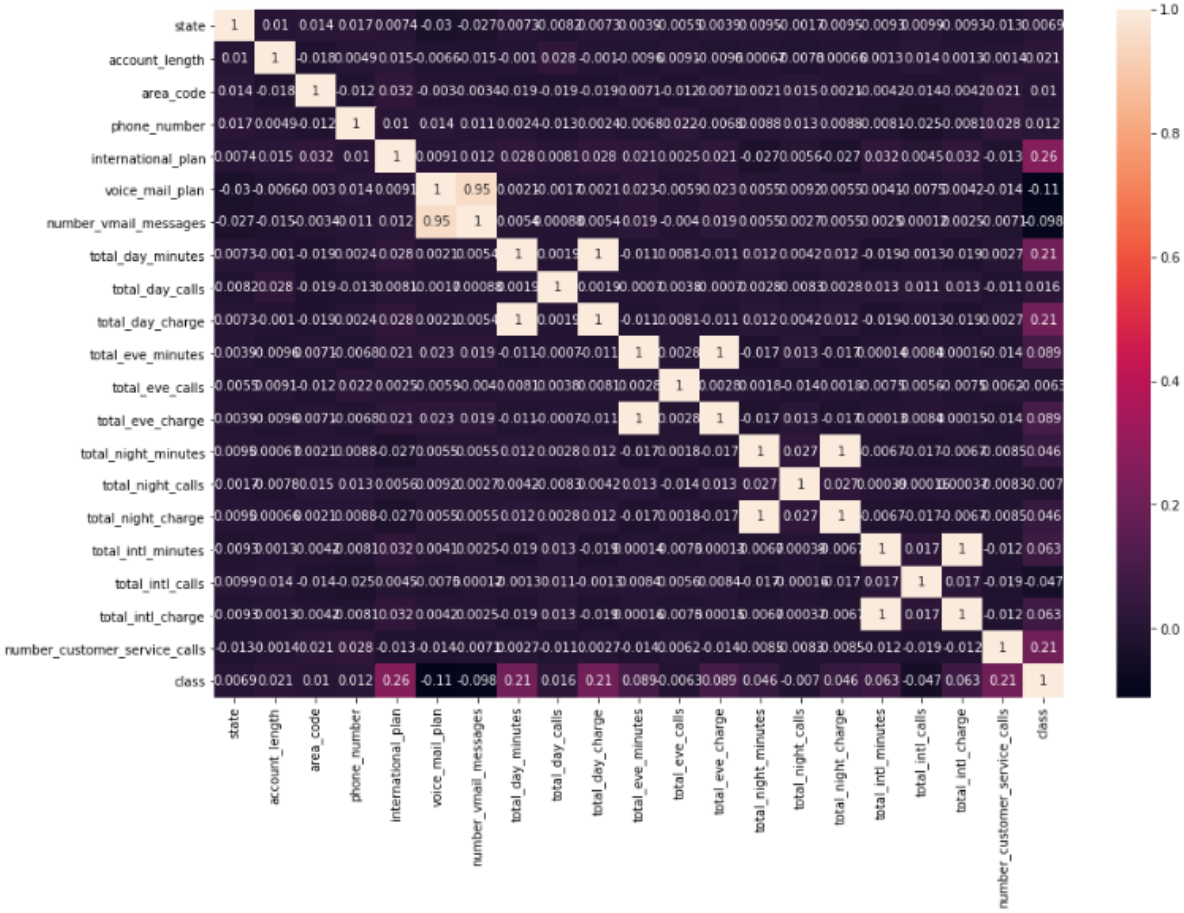
-The maximum, minimum and average (mean) of local calls:

Max Charges:	Min Charges:	Avg Charges:
Day: 59.8	Day: 0	Day: 30.6
Eve: 30.9	Eve: 0	Eve: 17.1
Night: 17.8	Night: 0	Night: 9.0
Intl: 5.4	Intl: 0	Intl: 2.7
Max Minutes:	**Min Minutes:**	**Avg Minutes:**
Day: 351.5	Day: 0	Day: 180.3
Eve: 363.7	Eve: 0	Eve: 200.6
Night: 395.0	Night: 0	Night: 200.4
Intl: 20.0	Intl: 0	Intl: 10.3
Max Calls	**Min Calls**	**Avg Calls**
Day: 165	Day: 0	Day: 100.0
Eve: 170	Eve: 0	Eve: 100.2
Night: 175	Night: 0	Night: 99.9
Intl: 20.0	Intl: 0	Intl: 4.4

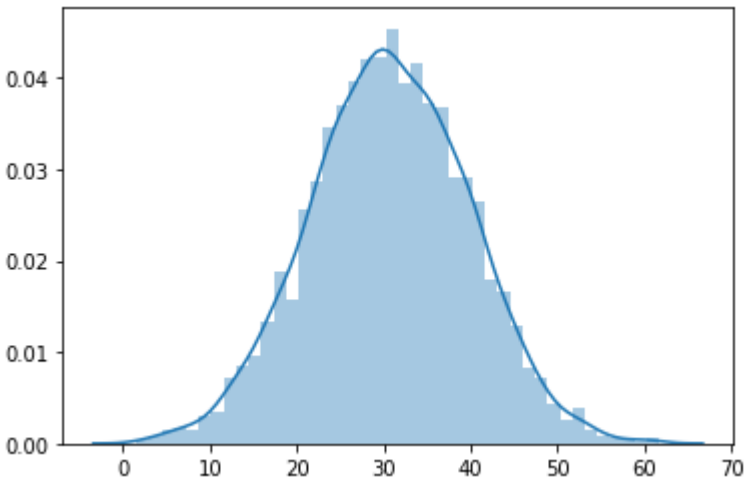
-Pairplot Analysis:



-Correlation Matrix of all Features (plotted using Seaborn library Python):



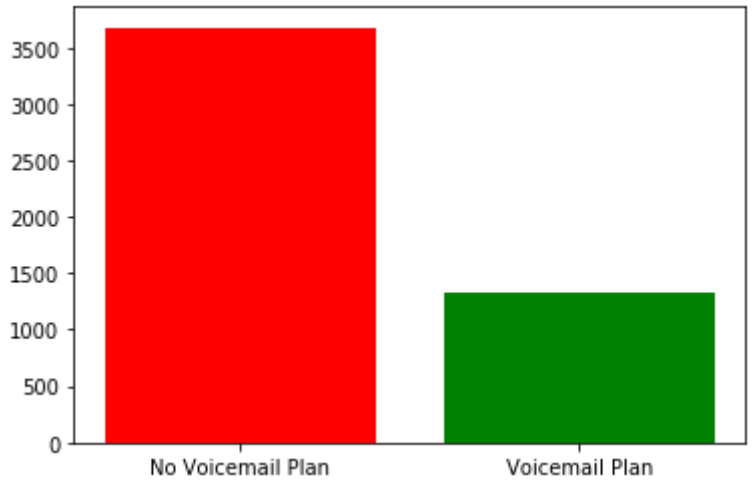
-Normal Distribution of International Charges per day (plotted using Seaborn library Python):



-Number of Users using Voicemail vs Non Voicemail Users:

Voicemail: 1323 (26.5%)

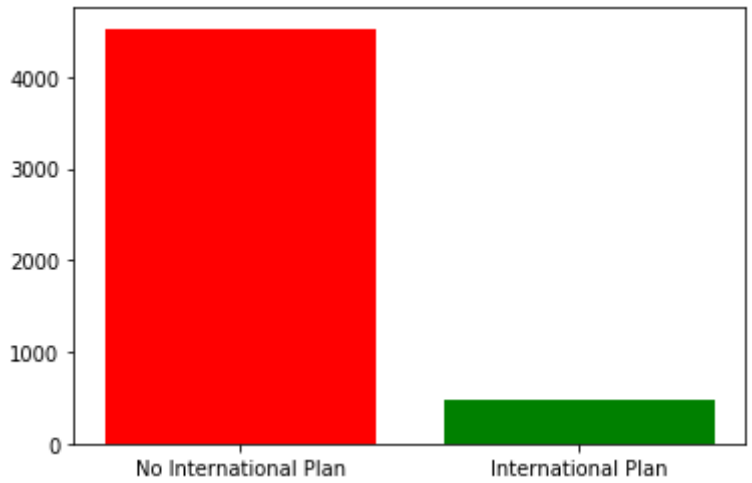
Non-Voicemail: 3677 (73.5%)



-Number of Users using International Plan vs Non International Plan Users:

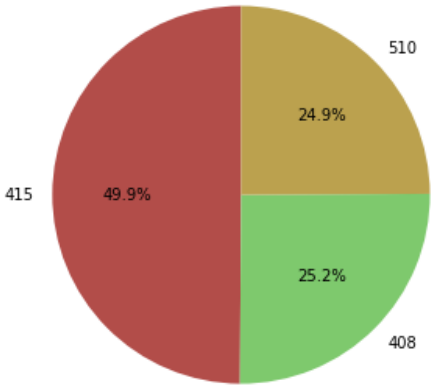
International Plan: 473 (9.5%)

Non-International Plan: 4527 (90.5%)



-Number of Users in an Area (by area_code):

Area Code	No. of Users
415	2495
408	1259
510	1246



Analysis and Observations from Data:

1. There are almost 14% churned users vs 86% users retained
2. The Churn Rate seems to have a higher correlation to Number of Customer Service Calls, International Plans, Total Day Minutes and Total Day Charges.
3. As seen through the pair-plots, most of the fields of charges, minutes and number of calls at three different times of the days are irrelevant to the prediction class if existing separately, causing more chances of overfitting. An indepth analysis of each time could help users with calling packages according to their most interacted with time of day
4. Normal distribution for international charges shows how much is spent on average by users as it is highly correlated to the churn class
5. There are less international plan users and voicemail plan users
6. The users can be divided into 3 regions by their area-code to see which area's users use the company's calling services most.

Selecting ML model:

Since we have historical data of customers, we can use it by training it to predict future results based upon existing records. Therefore a Supervised Learning model would be used in this case.

The remaining feature variables will be used to predict the target variable (class). Since it is a classification problem (churn vs non-churn) we will use a Classification algorithm. Also noting that there are only two choices of classes available, this is a binary classification problem.

The type and trends of data mean that models such as Logistic Regression, Decision Tree or Random Forest can be used.

Making Relevant Changes to Data (Preprocessing / Cleaning)

The dataset is a clean enough data set, with non NULL values, consistent types in each respected column, no special characters, mostly numerical, therefore data cleaning is limited to parameter tuning for optimized prediction models. Preprocessing for churn analysis included:

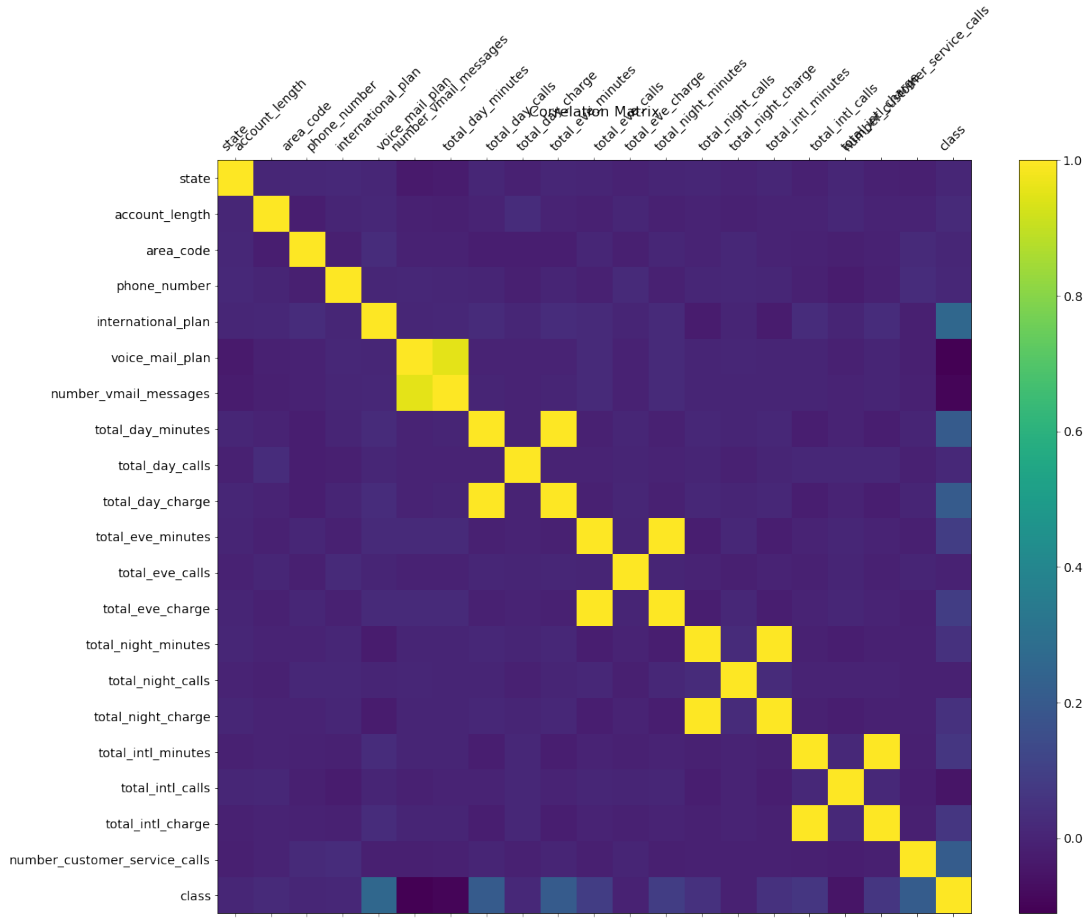
- removing columns to avoid Overfitting

- add new columns:
    - total_local_minutes = total_day_minutes + total_eve_minutes + total_night_minutes
    - total_local_calls = total_day_calls + total_eve_calls + total_night_calls
    - total_local_charge = total_day_charge + total_eve_charge + total_night_charge
    - average_charge_per_local_call = total_local_charge / total_local_calls
    - average_local_call_duration_in_minutes = total_local_minutes / total_local_calls
    - average_charge_per_intl_call = total_intl_charge / total_intl_calls
    - average_intl_call_duration_in_minutes = total_intl_minutes / total_intl_calls

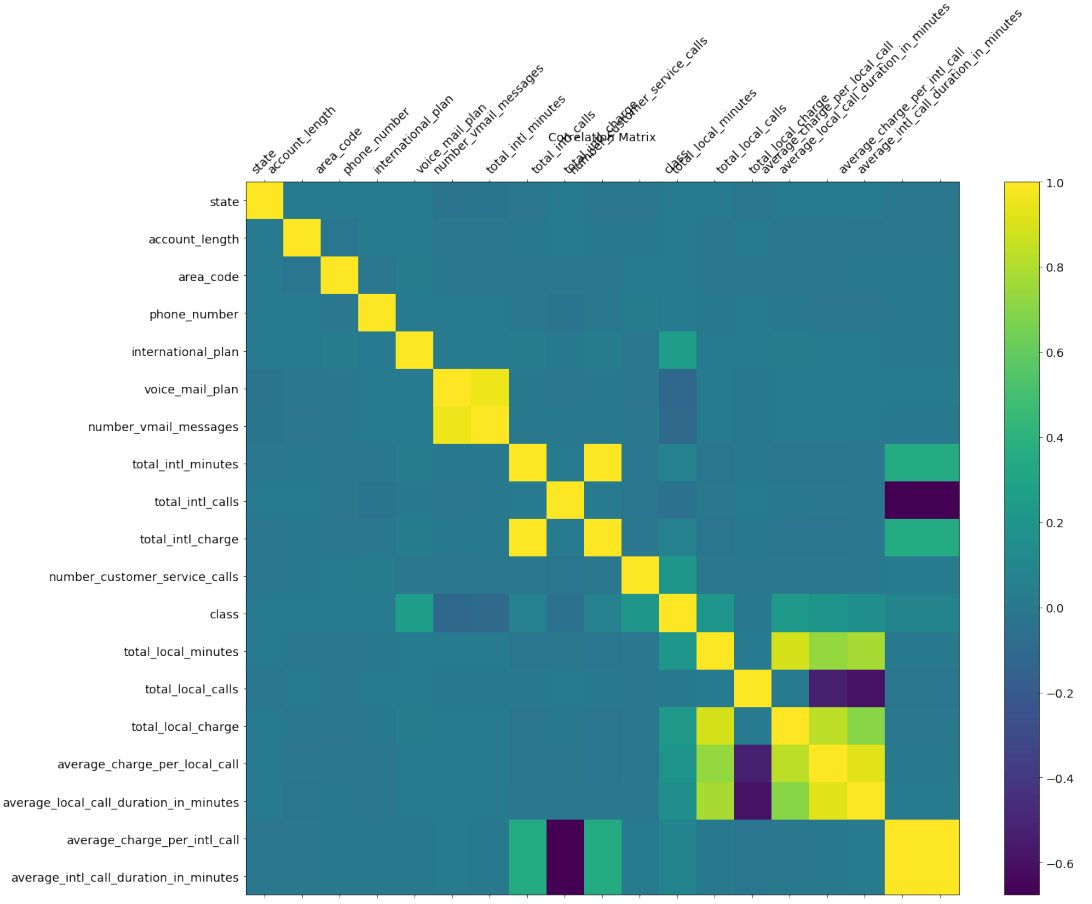
- removing outliers/anomalies

- splitting the data into training and testing. (80-20%)

Correlation of original columns:



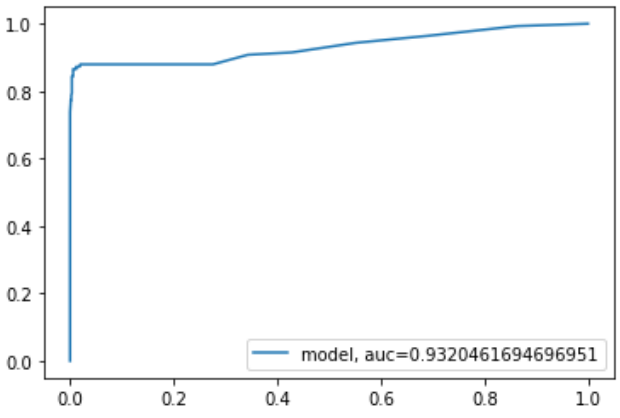
Correlation of new columns:

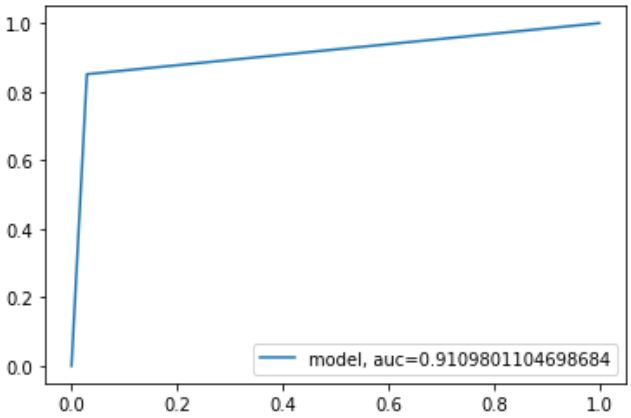


Prediction:

Using the Logistic Regression model, the accuracy came out to be: 0.861
While the Decision Tree and Random Forest Algorithms did much better with UAC-ROC scores of 0.91 and 0.93 respectively.

The ROC curve for the results looked like the following:





Comparing models:

X = 80 % , Y = 20 %

- Model 1 – Logistic Regression , trained on data X, evaluated on data Y.

  Accuracy score: 0.86

- Model 2 – Decision Tree , trained on data X, evaluated on data Y.

  Decision tree had 345 nodes with maximum depth 24.

  AUC-ROC score: 0.91

- Model 3 – Random Forest , trained on data X, evaluated on data Y.

  Average number of nodes was 394 with Average maximum depth 20

  AUC-ROC score: 0.93

We can clearly identify how the Binary Class Decision Tree performs better than the traditional Logistic Regression model with a marginal difference. Comparing the AUC-ROC value for both the ensemble models, we see that the prediction accuracy rate of the Decision Tree model is 2% lower than that of the Random Forest model that did fairly better than it. So with the Random Forest approach, we can identify the high risk churners more accurately.

Results and Limitations:

As seen by the prediction results, the model used to predict outcome of future churners would be Random Forest, until the data trend starts to change.

An optimized model would achieve better performance as the current model. For example more iterations could improve performance.

Depending on the number of features (columns) used in the model, the performance scores can be different.

Observed Top Features and their Importance (in descending order, highest to lowest):

	Feature	Importance
13	total_local_charge	0.263165
10	number_customer_service_calls	0.117336
4	international_plan	0.100967
11	total_local_minutes	0.078987
14	average_charge_per_local_call	0.053185

The **assumptions** made throughout the case study analysis include:

- each user has a different phone number, therefore all users are unique
- meaning of ‘state’ and ‘account_length’ are unknown, therefore assumed to be irrelevant
- that there is no other type of call being made except for those given in the dataset
- that all users started using at the same time because less usage , especially w.r.t minimum (0) values could mean that the user is new and therefore should not be made part of the analysis.

The **limitations** of the dataset can be listed as follows:

- Since no day/time range has been provided, it cannot be known how long have the users used the service.
- We do not know when the user started using the company’s calling services, or when exactly they churned, i.e. what stage in the product life-cycle they left?
- We do not know the country to tell how expensive the local calling rates are according to their currency value
- There is no information on which international countries were called, or separate international calling rates per country
- We do not know user demographics, age, gender, profession, education, income etc. to understand their need of the services being provided by the calling company
- There is no information about the period for which the data was collected to be able to tell if results are of old times or not, for example many changes occurred in the recent times with respect to technology improvements and third-party internet applications being launched that allow free calling services
- Many other parameters could be missing that could have affected the customer’s churn, like their usage of other services being offered by the company or their behaviors of each call.
- Data size too small for efficient prediction model. Larger dataset size could mean better training for future predictions.

Recommendations:

- The users can be retained through various calling packages according to their interests.

- Charges for local and international calls could be reduced.

- Users who use more at prefered times of days should be given packages accordingly.
- As seen by the max/min/avg data earlier, it can be noted that there are more number of calls and longer durations of calls at night, which could be because the calling charges are lower for night, so more night packages can be introduced by the company or the rates should be maintained for those times if it retains the most users.

- Customer Service callers need to be responded efficiently and followups taken from them so that they do not churn. More time and finances can be invested in training customer service representatives, calls recorded and areas of user complains noted for future improvement.

- Personalized recommendations according to type of user should be made. Users should be divided up into groups of those who make large number of small calls, users who make long calls, users who call internationally, and users who use voicemail plans. These users should be offered packages according to their needs after capturing their data and identifying their group.

- As addressed in the limitations, there should be more data captured for these users. With this it could be told that if the company is experiencing churn with users of low income demographics who are for example using more text message facilities than actual calls, it may be about creating a niche plan targeted to that segment to prevent the users from switching to the another provider.

- Identification of reasons of churned users could even help bring them back by creating promotional offers of using their services after leaving.

Conclusion:

All in all, the company should capture more relevant data and take more effective measures by the above analysis to keep their users from churning. The existing users' trend of usage should be compared to churned users so that it can be predicted when they would leave and how to keep them as well as new users from leaving.

Tools Used:

Languages: Python (libraries: pandas, numpy, seaborn, matplotlib, sklearn, scipy) , SQL
IDE: Jupyter Notebook
Database: MysSQL
BI tool: Power BI
OS: Linux + Windows