An ML Approach for Analyzing Customer Churn

Context:

Prediction of Customer Behavior and Customer Retention has always been a challenge for any business. It costs more to acquire new customers than it does to retain existing customers. This business Metric helps to understand the reason behind the churn and to take effective initiatives to deal with the churn percentage

Customer churn is the percentage of customers that stopped using the company's product or service during a certain time frame. It costs more to acquire new customers than it does to retain existing customers.

Customer churn impacts heavily on any business. There are other business metrics involved in the customer churn. A high customer churn rate indicates that a large percentage of your customers no longer want to purchase your products or services for various reasons, which can be a sign that your business is lacking in certain departments.

In this analysis, we dealing with data provided by a Teleco Company in California Q2 2022 on Customer Churn. The purpose of this study is to analyze the provided data, try to find some meaningful interpretations and focus on finding a relevant ML model for predictive purpose.

Objective:

The goal of this study focuses on – "Why customers churn? How can improve customer retention?".

We will try to find out what factors may contribute to the Churn of the customers and provide supporting analysis to build an action plan to reduce the churn.

Our scope to find solution for the following:

- After the subscription for the service, at which time point the Customer tend to leave most?
- Do the services provided by the company like Monthly Bills, Internet Speed, Competitor's Offers contribute to Churn?
- Do Customer's Demography Contribute to churn? (e.g. Location, Family Size, etc.).

About the Dataset:

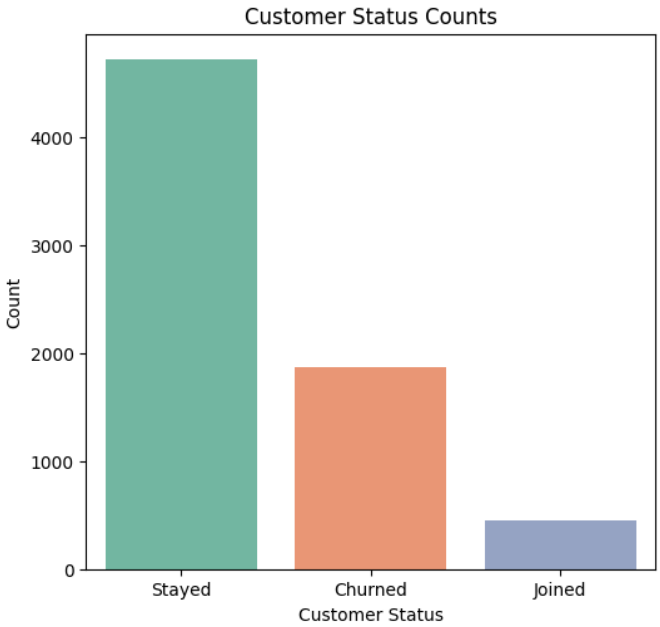
The source of the data is here. As mentioned in this Kaggle website – this dataset contains 3 tables, in CSV format:

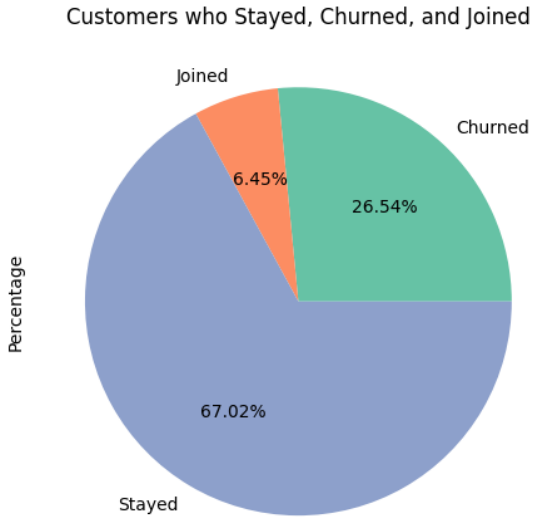
- The customers Churn table (7043 rows, 38 columns) contains information on all 7,043 customers from a Telecommunications Company in California Q2 2022.
- Each record represents one customer, and contains details about their demographics, location, tenure, subscription services, status for the quarter (joined, stayed, or churned), and more!
- There is another file with the Data Dictionary and Zip code. The zip code population table contains complementary information on the estimated populations for the California zip codes in the customer churn table.

Our interest was focused on the 'Customer Churn Table' dataset. The 'Customer Status – Joined, Stayed or Churned' column is our event of interest.

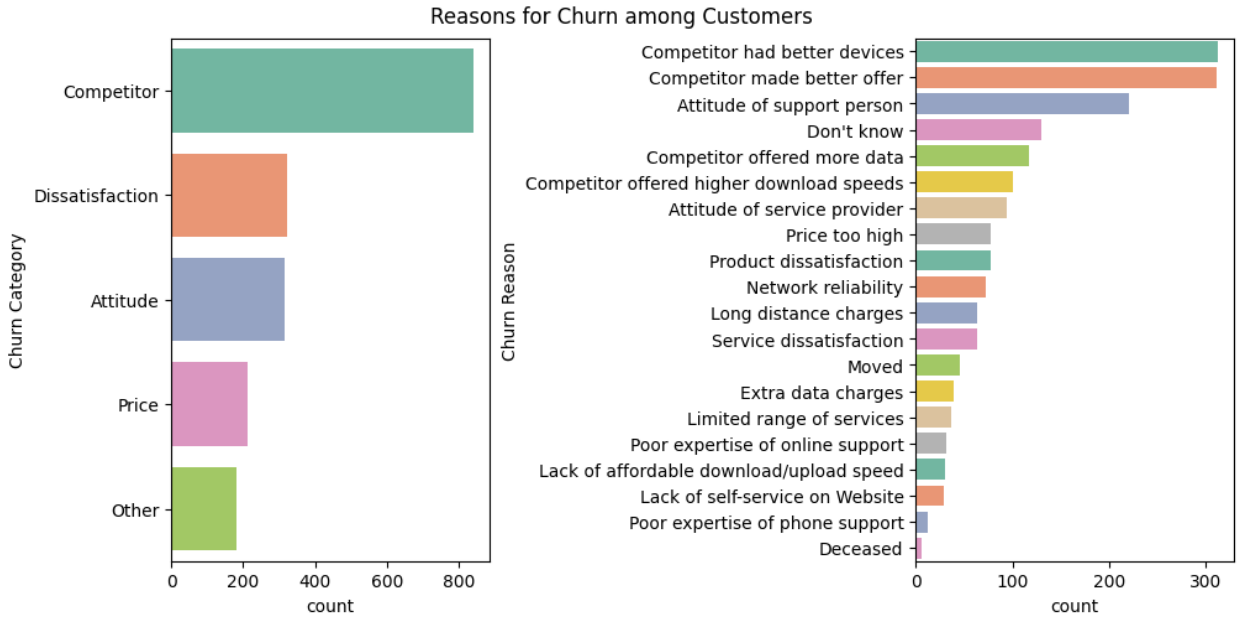
Exploring the Reason for Churn:

The exploration of the Churn reason among churned customers revealed some interesting findings. At the end of Q2 2022, 67% of the customers stayed with the service provider and 26.54% of the Customer Churned (a customer churn rate 25% is considered high).





While leaving, they were asked the reason. The ranking of the reasons was as below:

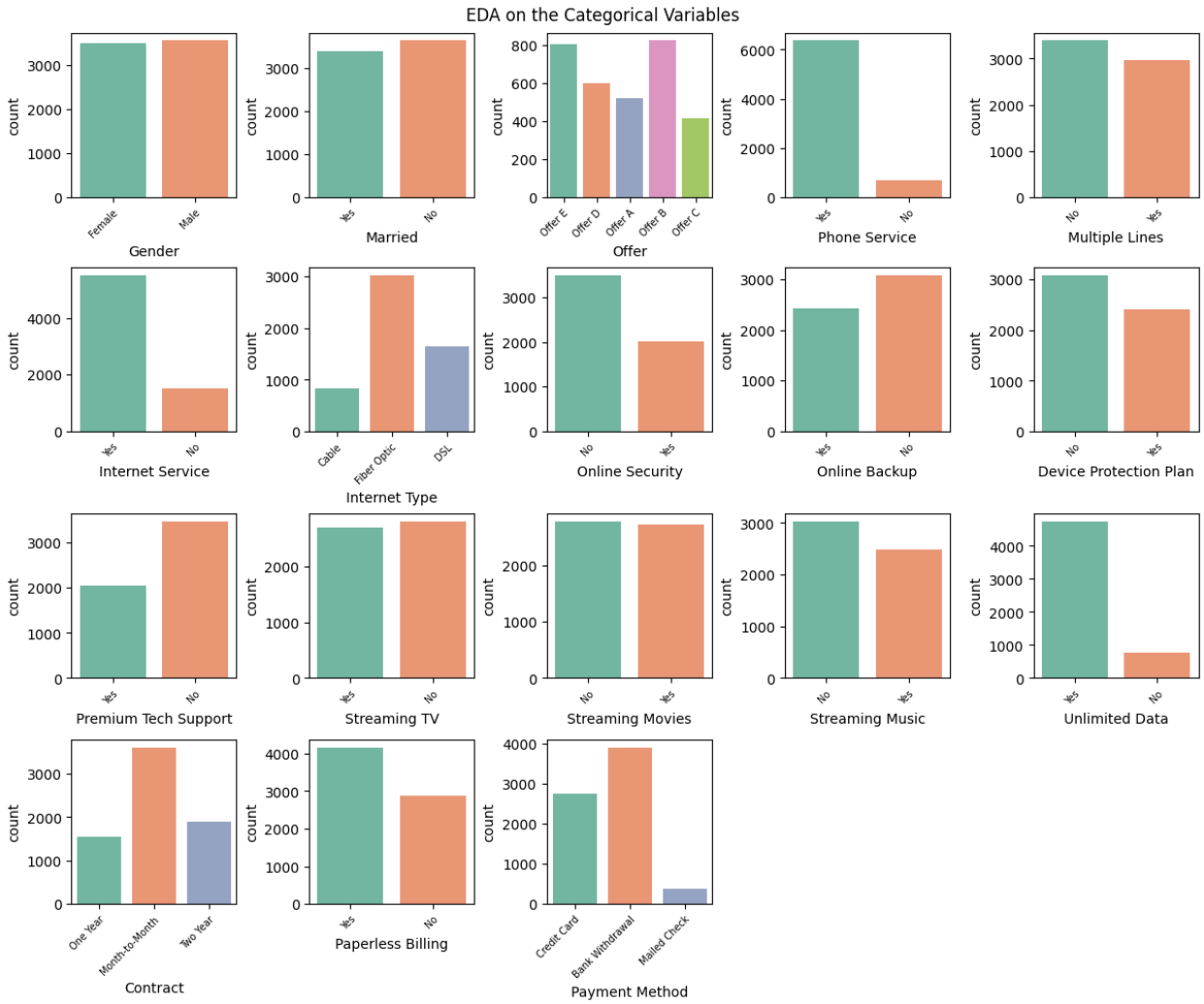


The majority of the Churned customers noted the competitor had better offers and services. And the second major reason cause of the churn was the attitude of customer representative as their reason to churn.

From the response, we observe that 26% of our customers churned, so this is an Imbalanced – Class Classification problem.

Exploring Categorical Variables:

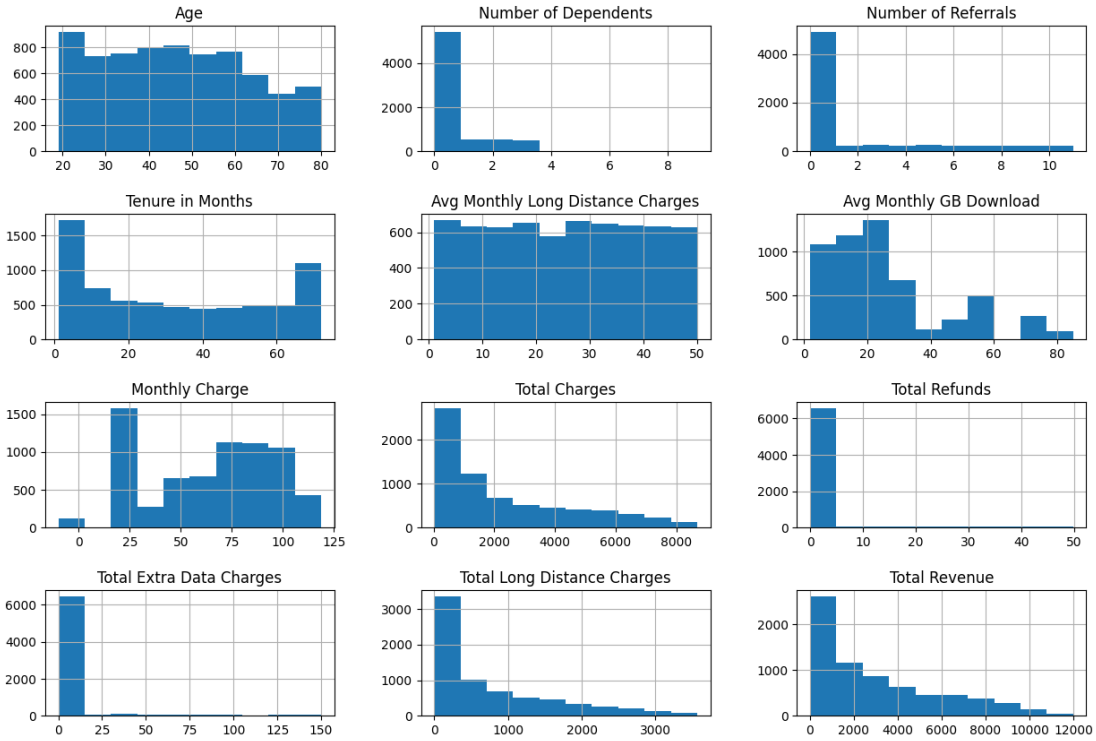
From the EDA of the categorical variables, we can get a basic overview from the chart below:



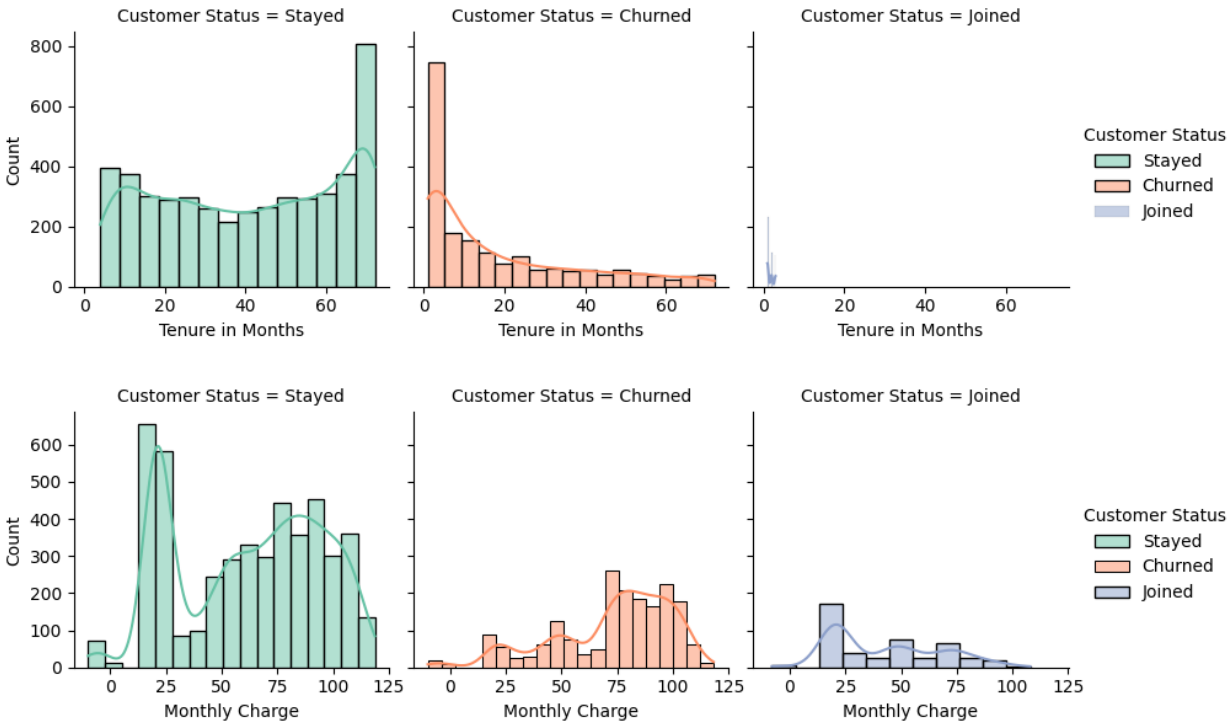
We can see those variables like Phone Services, Unlimited Data, Payment Method etc. are highly imbalanced. Also, responses of some variables such as Multiple Lines, Internet Services etc. are dependent on the response of 'Phone Service' and 'Internet Service' features.

Contribution of Numerical Variables:

Below is the histogram of the numerical Variables of the Dataset

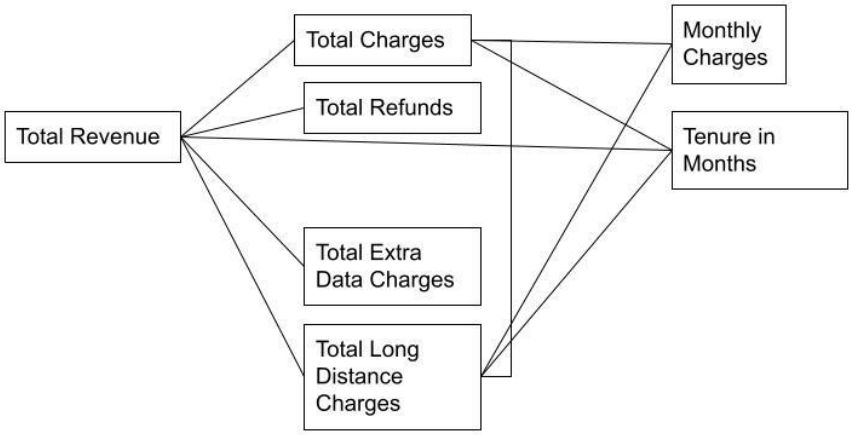


From the histogram of variables like – 'Tenure in Months', 'Monthly Charges' etc., the churn behavior seems significantly different.



There are also some interesting findings from the correlation matrix:

Here we observe that Tenure in Months, Total Charges, Total Long-Distance Charges, Total Revenue, Monthly Charges are significantly correlated. So, this will introduce multicollinearity in Logistic Regression Models. From the data dictionary, the relationship among the variables can be shown as –



To avoid multicollinearity, we will keep the 'Monthly Charges' and 'Tenure in Months' columns and omit the rest in the data preprocessing step.

Data Preprocessing:

1. **Missing Values:**

The data frame has a good chunk of missing values. It turned out that some variables were related to the response to another feature and left blank. Those missing cells had to be replaced with some corresponding values. The data dictionary file was very useful for this part for replacing the missing values.

2. **Drop Unnecessary Columns:**

For our ML based Analysis, there were some unnecessary columns (i.e. "CustomerID", "Zip Code", "Latitude", "Longitude", "Churn Category", "Churn Reason", etc.) which I had to drop.

After further investigation, several other variables seemed to be related with each other (High Multicollinearity). I dropped those corresponding variables ("Total Revenue", "Total Charges", "Total Long Distance Charges") and proceeded with independent variables in the model.

3. **Converting String to Float:**

While processing the data, there were some variables I had to convert from string type to float and integer.

4. **Dealing with Categorical Features:**

There were 20 categorical features in the dataset where some of them had multiple categories. Hot-Encoding or dummy encoding would have increased the dimensionality of the data frame. So, I used 'LabelBinarizer' from scikitlearn to encode the categorical variables. This helped to keep control on the high dimensionality problem by keeping the important features and disposing the redundant.

5. **Dealing with Numerical Values:**

The numerical variables were not Normally distributed (Bell Curve), so I normalized the numerical variables (MinMax Scaler from Scikitlearn Library).

6. **Dealing with the Response Feature (Churn Status):**

The churn status had 3 categories – Stayed, Joined, and Churned. To avoid ambiguity, I used the 'pandas.dummies' method for encoding and took the 'Churn Status – Churned' binary response column as my response variable.

7. **Upsample the Response Variable:**

Since this is an Imbalanced Class Classification problem, I used the SMOTE method to upsample the dataset before splitting the train and test data.

8. **Train and Test Split:**

The training and test split of the Dataset was 75%-25% after upsampling the dataset.

9. **Designing Pipeline with Scikitlearn:**

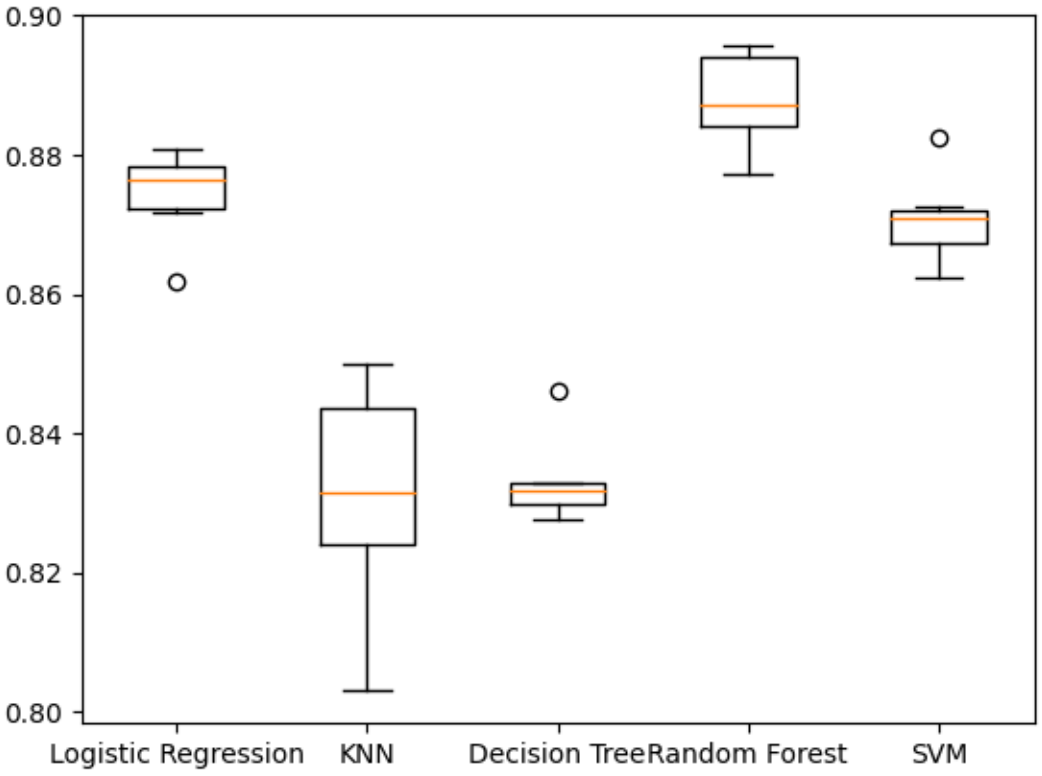
LabelBinarize and LabelEncoder fit and transform signatures not compatible with Pipeline. This needed to be manually, so pipeline for the data preprocessing couldn't be the implemented.

Modelling:

As our classifiers, we considered 5 different Algorithms:

- Logistic Regression
- KNN
- Decision Tree
- Random Forest
- SVM

Below is the box plot of the distribution of the k-fold cross validation score for different classifiers:



Our parameters of consideration on Model Performance would be-

- Accuracy
- F1 Score (for Precision and Recall)

The random forest and logistic regression models performed well in terms of accuracy (RF 89% accurate and Logistic 88%).

But we would also emphasize on the F1 score since our count of false positives and false negatives. For business policy planning and design, it's the management's call to decide on which performance metrics should be prioritized (Precision or Recall); in other word which segment of customers should we target to churn. This will help to contribute towards developing effective customer retention policies for company.

Summary:

There might be varieties of classifiers that might classify Churn accurately, but the challenge I faced was while data preprocessing. Our classifiers (Random Forest and Logistic Regression) performed well for classification, but I felt there is still scope for defining pipelines and finding important features. The 'Labelbinarizer' (Scikitlearn library) helped to reduce the dimensionality, but it didn't comply well for pipeline designing and important feature finding.

In general, Churn Prediction is a challenge for all kinds of business. It contributes a lot in telecommunication business in terms of planning, budget, marketing, campaign, service etc. The goal of this initiative is to reach out to the customers to prevent leaving the company. The performance of the classifier will always be a trade-off between precision and recall. If done with proper interpretation, this might lead to a helpful Churn prediction model.

The findings from the analysis regarding Customer Churn due to Attitude of Support/ Service person, Competitor's Offers, Contract type etc are the helpful findings, but using these findings for predictive modeling needs more deeper attention from data and software engineering perspective.