**DATA REPORT**

**BUSINESS UNDERSTANDING**

A vast majority of companies suffer from customer churn - the loss of customers to competition. This can be incredibly damaging to a company's profitability as they might have spent hundreds/thousands of dollars to acquire new customers. When a customer leaves, the company is not only losing out on potential revenue but also the considerable resources spent in acquiring the customers are lost. Telecommunication companies, Syriatel in this case, are no exception.

SyriaTel is a client looking to combat customer churn. The company provided client data in order to better understand whether a consumer would stop doing business with the company. My goal is to create a machine learning model that can help predict whether a customer will churn given certain data about their usage. I will use the CRISP-DM methodology to analyze the dataset and prepare it for machine learning algorithms. Selected features will be used to give business insight and efficiently direct customer retention efforts towards at-risk customers.

In particular, I am going to answer the question, Is a particular target going to churn or not?.

**DATA UNDERSTANDING**

I shall be using the SyriaTel Customer Churn dataset downloaded from Kaggle. This is a public dataset contained in a CSV file that details customer usage patterns and also includes a column delineating whether the customer has churned or not. The nature of the 'churn' column lends the dataset towards a binary classification problem, where a machine learning model can be constructed and trained on the data to predict whether a customer will churn or not given their usage patterns. The 'churn' column will be used as our target column in this binary classification problem. The dataset contains 3333 rows and 21 columns giving details about each datapoint.

The features include;

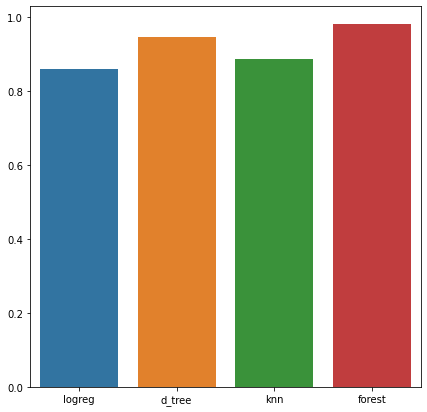
* state - State where the customer resides
* account length - period the customer has been subscribed to the service
* area code - area code of the customer
* phone number - clients phone number
* international plan - indicates if the customer has an international plan or not
* voice mail plan - indicates if the client has a voice mail plan or not
* number vmail messages - number of voice mail messages
* total day minutes, calls, charge - total minutes, calls and charge of the customer in the day
* total eve minutes, calls, charge - total minutes, calls and charge of the customer in the evening
* total night minutes, calls, charge - total minutes, calls and charge of the customer in the night
* total intl minutes, calls, charge - total international minutes, calls and charge of the customer.
* customer service calls - number of calls made to the customer service
* churn - our target

**DATA PREPARATION**

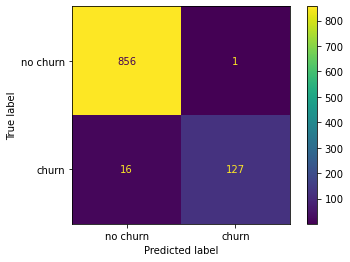
A common rule of thumb is that 80% of the project is data preparation. This phase, which is often referred to as “data munging”, prepares the final data set(s) for modeling. It has five tasks:

* Select data: Determine which data sets will be used and document reasons for inclusion/exclusion.
* Clean data: Often this is the lengthiest task. Without it, you’ll likely fall victim to garbage-in, garbage-out. A common practice during this task is to correct, impute, or remove erroneous values.
* Construct data: Derive new attributes that will be helpful. For example, derive someone’s body mass index from height and weight fields.
* Integrate data: Create new data sets by combining data from multiple sources.
* Format data: Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.

**MODELLING**



**EVALUATION**



**CONCLUSION**

Understanding customer behavior can help different departments know where to focus their attention. Decisions about where to place the most effective and proactive sales force, where to place new advertising, or what types of products to develop or discontinue required that particular decision or investment You can do it with confidence. Previously identified behaviors may be thwarted if negative and encouraged if necessary.

This dataset showed different consumer trends and showed some possibilities. These two areas of opportunity addressed common retail pitfalls which include expensive products (International calls are the most expensive) and Day calls are the most common categories available. They have directly impacted most of the consumer's costs. These and future trends can be used to prepare for future trading.

**RECOMMENDATIONS**

**Question 1**

From our distribution, we can see that day calls generate the highest revenue at more than 100,000 and international calls generate the least revenue at less than 10,000.

Recommendation: developing plans for international, day, evening and night calls.

Since total day minutes make up the majority of the minutes consumed and therefore relay to the largest percentage of the total cost. SyriaTel can create and market advertisements for consuming minutes during low peak business hours or consider changing their day pricing (adjusting evening, night and international costs to compensate).

**Question 2**

From our top ten highest revenue states, West Virginia leads with over 6000 with the rest generating revenue between 5000 and 4000.

Recommendation: reducing call rates in other states and running a campaign to advertise the reduction.

**Question 3**

Our data shows us that area code 415 is our highest grosser with revenue of around 100,000 followed by area code 408 then 510.

Recommendation: reducing call rates in other states and running a campaign to advertise the reduction.

**NEXT STEPS**

In the future, this analysis could be improved by adding additional data as it becomes available. Some of the area of interest that arose from these analysis that might be worth exploring further include;

* What makes a state perform better than its counterparts?
* What are some of the reasons why certain area codes spend more?

Additionally, it would be useful to have more current data, as the dataset used to train and test our model is from 2012.