Machine Learning Engineer Nanodegree

Capstone Proposal

Mohammed Ashraf Farouq Ali
June 8th, 2019

Domain Background

Telephone Companies play vital Role in our daily base lives, with numerous of thousands of services we encounter each day and hour

Context

Customer attrition, also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers.

Telephone service companies, Internet service providers, pay TV companies, insurance firms, and alarm monitoring services, often use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.

Companies usually make a distinction between voluntary churn and involuntary churn. Voluntary churn occurs due to a decision by the customer to switch to another company or service provider, involuntary churn occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a

distant location. In most applications, involuntary reasons for churn are excluded from the analytical models. Analysts tend to concentrate on voluntary churn, because it typically occurs due to factors of the company-customer relationship which companies control, such as how billing interactions are handled or how after-sales help is provided.

predictive analytics use churn prediction models that predict customer churn by assessing their propensity of risk to churn. Since these models generate a small prioritized list of potential defectors, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn.

What we'll be Discussing through this project is the various type of customers and how the process of their subscription went through out and how that affects the future state of customers on the long term and understand analysis of the problem.

Acknowledgements

The dataset has been collected from an **[IBM Sample Data Sets]**

- *Check more about the dataset here.*
- *Related research paper here.*

Problem Statement

In this Case of the Project we'll be dealing with :

- **Exploring the Dataset**
- **Manipulation of data to Tenure Groups for later classification**
- **Use of Info provided to create Statistical Analysis of the state of Each Customer**
- **Understand the Cause of Customer Churning from the overall Subscribed Services & Other Info**

- **Visualizing the descriptive statistics of the whole Dataset**
- **Preprocess Data & Remove Un-Necessary Data**
- **Remove Un-Correlated Data**
- **Shuffle & Split Data**
- **Train & Test Both SVM & Log Reg Models**
- **Resample, Shuffle & Split Data then retrain**
- **Measure & Compare Final Scores and Improvements**

DataSet & Input:

As for our input we'll be using the Churn variables to compare results, and as well we'll split data and shuffle it for training purposes after analysis of what features are more relevant to our process.

The data set includes information about:

- **Customers who left within the last month – the column is called Churn**
- **Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies**
- **Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges**
- **Demographic info about customers – gender, age range, and if they have partners and dependents**
- **As well it consists of 7044 * 21 Cells of Info Ranging From Customer ID's to their Churn State**

Solution Statement

In this Case of the Project we'll be dealing with :

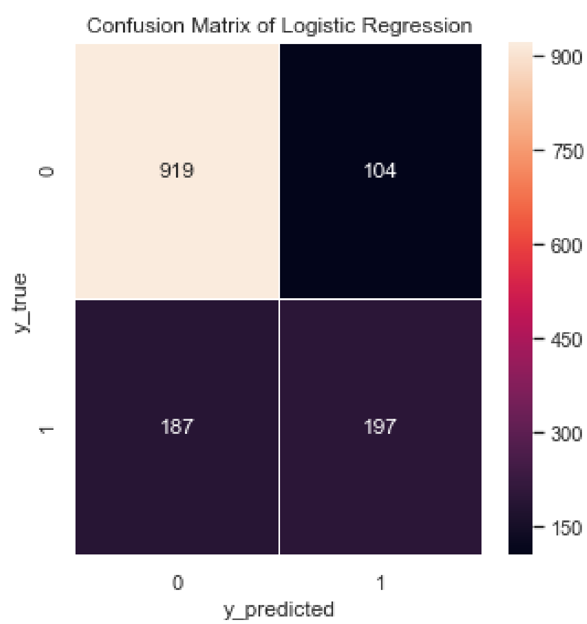
- **Exploring the Dataset**
- **Manipulation of data to Tenure Groups for later classification**
- **Use of Info provided to create Statistical Analysis of the state of Each Customer**
- **Understand the Cause of Customer Churning from the overall Subscribed Services & Other Info**
- **Visualizing the descriptive statistics of the whole Dataset**
- **Preprocess Data & Remove Un-Necessary Data**

- **Remove Un-Correlated Data**
- **Shuffle & Split Data**
- **Train & Test Both SVM & Log Reg Models**
- **Resample, Shuffle & Split Data then retrain**
- **Measure & Compare Final Scores and Improvements**

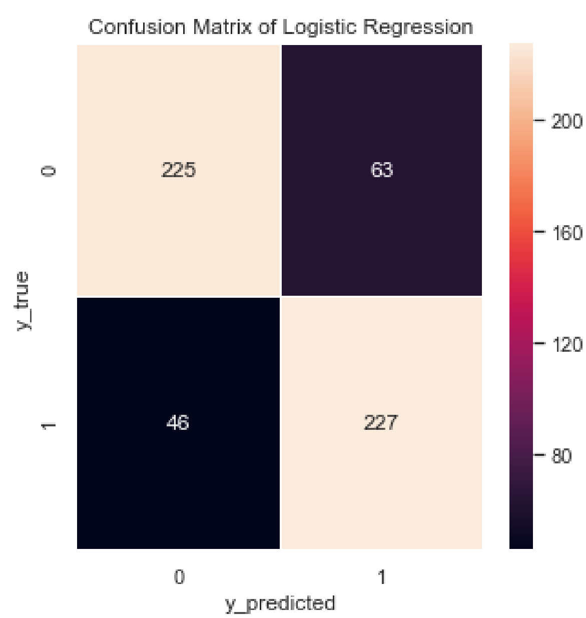
Benchmark Model

In This case we'll be using SVM(Support Vector Machine) & Logistic Regression Algorithms for the Classification and we'll be calculating Scores to test for Improvement before and after Sampling & Refinement, as they are to be the best models for benchmarking and training for this case here, we'll be calculating the precision , recall, accuracy & F-beta score for SVM, and Accuracy , F-1, Precision, Recall Score for Logistic Regression that will calculated based on the ***Confusion Matrix*** *Scores.*

Before Sampling



After Sampling



Evaluation Metrics

Recall and precision: Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. An ideal system with high precision and high recall will return many results, with all results labeled correctly. And here a very good article about Recall and precision score.

F-Beta: F-Beta score is a way of measuring a certain accuracy for a model. It takes into consideration both the Recall and Precision metrics. If you don't know what those are, it's highly recommended to check the previous post about Recall & Precision.

--A quick definition of recall and precision, in a non-mathematical way:

Precision: high precision means that an algorithm returned substantially more relevant results than irrelevant ones **Recall:** high recall means that an algorithm returned most of the relevant results.

Good article and info here.

Confusion Matrix for calculating Logistic Regression Scores: A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. Quick easy article here.

F-1 Score: is a measure of a test's accuracy, as it considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. More Info here.

Project Design

- **1. Data & DataSet Import**
    - **1.1 Importing Dataset from Kaggle If Using Google Colab**
        - **1.1.1 Install Kaggle Packages**
        - **1.1.2 Importing Kaggle Api & Creating Kaggle Directory**
        - **1.1.3 Download Dataset**
        - **1.1.4 Unzip Dataset Package**
        - **1.1.5 Function For Running Plotly Graphs on Google Colab**
    - **1.2 Import required Libraries**
- **2. Data Manipulation**
    - **2.1 Creating Cluster groups of Tenures**
    - **2.2 Load Dataset after Manipulation**
- **3. Analysis**
    - **3.1 Data Exploration**
        - **3.1.1 Load DataSet**
        - **3.1.2 Checking for Missing Data**
        - **3.1.3 Data Format Description**
        - **3.1.4 Data Descriptive Analysis**
        - **3.1.5 Data Unique Values Calculation**
- **4. Methodology**
    - **4.1 Data Preprocessing**
        - **4.1.1 Data Normalization**
        - **4.1.2 Loading Data After Normalization**
        - **4.1.3 Moving Churn Column to Last for Correlation Matrix Readness Ease**