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 churn is defined by losing customers or clients.

Which as in overall companies like mention here telephone companies pay numerous amount of money to explore the possible reasons and causes to customers churning from the company as for voluntary causes such as customer decides to leave or go for another service provider or involuntary cause which can be due to natural causes that prevent the customer from continuing his/her subscriptions, as in overall spending capital over how to keep customers satisfied and committing to the company is often less costly than exploring new ways to get new customers to replace old ones.

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

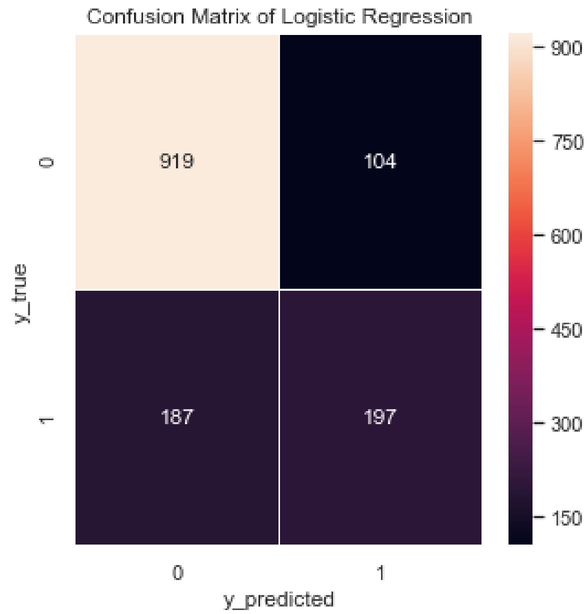
In this Case of the Project we'll be dealing with :

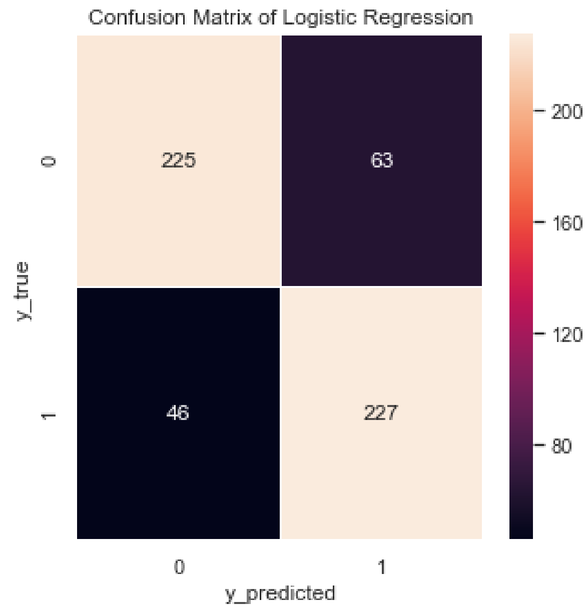
- **Exploring the Dataset**
- **Manipulation of data to Tenure Groups for later classification**
- **Use of Info provided to create Statistical Analysis of the state of Each Customer**
- **Understand the Cause of Customer Churning from the overall Subscribed Services & Other Info**
- **Visualizing the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


Confusion Matrix of Logistic Regression
919
104
187
197
y_true
y_predicted
0
1
900
750
600
450
300
150


Confusion Matrix of Logistic Regression
225
63
46
227
y_true
y_predicted
0
1
200
160
120
80

Evaluation Metrics

Recall and precision:

Recall and precision: Precision-Recall is a useful measure of success of prediction when the classes are imbalanced whereas, In information retrieval, precision is a measure of result relevancy that helps define correlation proportions , while recall is a measure of how many truly relevant results are returned moreover a model with high recall but low precision returns many results, which most of its predicted labels are incorrect results when compared to the training labels. In more contrast way a model with high precision but low recall is just the opposite, returning very few results of labels, but most of its predicted labels are correct of the comparison of the training labels. A mostly perfect model where with high Precision & Recall would return as much results as possible with mostly all of them are correct compared to the Labels. [here](https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c) is 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 as It takes into consideration both the Recall and Precision metrics that makes it check the credibility of both metrics results.

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

Precision: high precision means that an algorithm returned mostly more relevant results to the labels than irrelevant

Recall: high recall means that an algorithm has returned almost of the relevant results possible.

Confusion Matrix for calculating Logistic Regression Scores:

Confusion Matrix for calculating Logistic Regression Scores: A confusion matrix is a table that is used to define the performance of a model or a classifier on a data that is mostly known. While it can somehow confuse readers it can be easy to understand them, where it has main aspects which divide into True Positives, True Negative, False Positive, False Negative. Quick easy article [here](https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/).

F-1 Score: is a measure of a test's accuracy, as it takes into consideration both precision and the recall scores of a test to determine the accuracy whereas, P is the amount of correct positives divided by amount of all positives returned by the classifier or model, moreover R is amount of relevant results divided over the amount of all relevant results.

Project Design