Capstone Project

Victor Nwaokocha

Machine Learning Engineer Nanodegree

March 4, 2018

Customer Churn Prediction.

1. Background.

Telecoms company spend a lot on acquiring new customers, the telecoms market is highly competitive at the same time customers are not exactly satisfied with the service provided by most of the telecoms company. Thus, customers leave service providers, when customers leave their current service provider and move to the next this is called churn.

Previous attempts or similar endeavors to predict churn include:-
[1] Using the deep unsupervised feature learning to predict customer churn for companies and it's not specific to telecoms. [2] Comparing machine learning techniques to predict customer churn which is somewhat related to this because it's based on the telecoms industry however, this does not apply to Nigerian telecoms.

2. Problem Statement.

The problem is simply being to predict the customers that are most likely to leave their current service providers. Classifying customers that will churn based on spending and level of consumption of certain services by the provider using a classification model, would help telecoms prepare and prevent customers from leaving.

Furthermore, this is a binary classification problem because it involves categorising the customers in the datasets and future customers, that would churn or won't churn.

3. Datasets and Inputs.

The datasets can be downloaded on the Kaggle competition page here:
https://www.kaggle.com/c/dsntelecomschurn2018/data

The following are features in the features in the data set:

- **Total Spend in Months 1 and 2 of 2017:** The aggregate spending of a customer in the months July and August 2017.
- **Total SMS Spend:** The aggregate spending on SMS by a customer income earned through the SMS benefit utilized by the subscriber.
- **Total Data Spend:** The aggregate amount spent on Data/Internet by a customer income earned through the SMS service used by the subscriber.
- **Total Data Consumption:** The aggregate data consumed by a customer in KiloBytes during the period under study.

- **Total Unique Calls:** The aggregate of unique calls made by a customer during the period under study.
- **Total Onnet spend:** The aggregate spending of a customer to make on-network calls (on the same network as the subscriber).
- **Total Off-net spend:** The aggregate spending of a customer on off-network calls (not the same network as the subscriber).
- **Customer Tenure in Months:** The time passed since the subscriber started using the services of the network provider and counted in months.
- **Network type subscription in Month 1:** The network type the customer is subscribed to in the first month which maybe 2G or 3G service.
- **Network type subscription in Month 2:** The network type the customer is subscribed to in the second month which maybe 2G or 3G service.
- **Total Call centre complaint calls:** Aggregate number of complaints made by the subscribers.
- **Most Loved Competitor network in in Month 1:** The customers most preferred competitor network provider in the first month.
- **Most Loved Competitor network in in Month 2:** The customers most preferred competitor network provider in the second month.
- **Churn Status:** The churn status of a customer, 1 means the customer has churned and 0 means no churn.

The data is balanced, the total number of customers that churn is 700 and 701 customers that didn't churn. In total there are 1400 rows in the train datasets.

4. Solution Statements.

The classification model can solve the problem. By cleaning the data, dropping irrelevant features and engineering new features like other spending made by the customer a classification algorithms such as *DecisionTreeClassifier* or *Support Vector Machines* can be trained using the refined datasets.

5. Benchmark Model.

Using Logistic Regression on the test sets the f-beta score is **59%** and the accuracy score is **67%** the logistics regression algorithm was trained with all the features in the datasets. Furthermore, based on the submissions on the leaderboard of the Kaggle competition the best score is **100%.**

The solution model developed can be compared to the benchmark model by downloading the test data, cleaning it, making predictions and finally submitting the predictions in the format of the sample predictions.

6. Evaluation Metrics.

These are metrics that would be used in measuring model accuracy:

Accuracy
This measures the number of correctly classifies predictions.

$$\frac{True\ Positive + True\ Negatives}{Dataset\ Size}$$

F-beta-Score

The f-beta score is the weighted average of the precision score and recall score. It ranges from 0 to 1, where 1 is the best(perfect precision and recall) and 0 is the worst value.

$$f1-score = 2 \times \frac{(Precision\ Score \times Recall\ Score)}{(Precision\ Score + Recall\ Score)}$$

7. Project Design.

Firstly, the data will be imported and explored. The exploration would be specifically checking for missing values in various columns, furthermore, all string based values would be converted to their numerical equivalent.

Also, checking for the correlations in the datasets and differentiating relevant features from the irrelevant features using RandomForest.

Furthermore, using the cross-validation to separate training sets from tests and grid search cv to find the best parameters for algorithms such as DecisionTreeClassifier. Comparing the performance of the different algorithms based on 1%, 10% and 100% sample size of the training features.

Lastly, making predictions with the models developed by the algorithms and testing the accuracy and f-beta scores of the models.

8. References.

[1] P. Spanoudes, T. Nguyen “Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors” available at https://arxiv.org/pdf/1703.03869.pdf

[2] T. Vafeiadisa, K. I. Diamantarasb , G. Sarigiannidisa, K. Ch. Chatzisavvasa ”A comparison of machine learning techniques for customer churn prediction” available at https://www.researchgate.net/profile/Konstantinos_Chatzisavvas/publication/273439405_A_Comparison_of_Machine_Learning_Techniques_for_Customer_Churn_Prediction/links/574edb0008aec50945bb3e95/A-Comparison-of-Machine-Learning-Techniques-for-Customer-Churn-Prediction.pdf