Machine Learning Engineer Nanodegree

Capstone Proposal

Scott Etheridge
October 20, 2018

Proposal

Domain Background

The Telecom industry has experienced greatly increased levels of competition over the last decade. While the competition is coming from the traditional Telecommunications and Cable industry players, it is also coming from companies in completely different industries. The ability to provide telecommunications service over the top (OTT) of the internet service has given companies in different industries the ability to become direct competitors. Google and Facebook have both implemented applications and services to complete in the traditional telecom markets. Likewise, the Wireless and Internet service markets have become commoditized services making it difficult to compete based on the quality of service/network.

The result has been that the telecommunications companies have competed based on price. It is never a good sign when an industry enters the stage of the product lifecycle where the only way to attract customers is based on price. This has led to the price wars that were observed from 2015 to 2018. Revenues have suffered greatly as a direct result of the price-based competition. In turn, the stock price of most of the telecommunications companies has remained stagnant during this period.

I have been working in the Telecommunications industry for the last thirteen years. I have a strong desire to help the Telecommunications companies thrive and improve their profitable - it is a matter of saving my job.

Problem Statement

While the Telecommunications companies are increasingly entering new markets to escape the declining wireline and wireless industries, they are also placing a laser focus on growing the revenues from their existing customer bases. They are placing a great emphasis on reducing the number of customers that are voluntarily terminating their contracts. The industry calls this "customer churn" – and the churn rate is defined as the percentage of the customer based that terminated their contract in the last month. The Telecommunications industry has a churn rate of ~2% (1).

The problem to be solved is to determine models that can be used to predict whether a customer is a candidate to terminate their contract. The goal will be to reduce the churn rate of the standard Telecommunications Company by being able to identify customers that are likely to churn. Then, the Telecommunications Company can target them with offers that will transform them into customers that are not likely to terminate their contract.

The problem will be set up as a classification problem. The labeled input data contains numerical and categorical data about the customer. Churn is the output (target) feature, which is a binary value of "Churn" or "Not Churn".

Datasets and Inputs

The dataset called "telco-customer-churn" will be used for this project. (https://www.kaggle.com/blastchar/telco-customer-churn)

The data contains the following information:

- 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

This dataset is from a telecommunications company and has been expressly created to support Churn analysis. Therefore, I do believe it is relevant to this project. Since it includes labeled variables as well as the target variable "churn", I believe it is a good candidate for supervised machine learning algorithms.

The dataset contains 7,043 rows (customers) with 21 labeled columns. The input data contains several continuous numerical features and several binary and multi-value categorical features about the customer. The labeled Churn feature will be the output (target), which is a binary value of "Churn" or "Not Churn".

Seventy-six percent of the customers in the base dataset are "Not Churn", while the remaining twenty-four percent of the customers are "Churn". Since the base dataset is imbalanced to the "Not Churn" value, the SMOTE sampling technique will be used to balance the dataset. Then, seventy percent of the dataset will be used for training and the remaining thirty percent for testing.

Solution Statement

The solution is to determine which of the supervised learning models most accurately predicts the binary classification of customer churn. Likewise, I will determine what variables contribute to customer churn, as well as their relative importance. The companies can then develop programs to understand each of the variables and resolve the ones that are the highest predictors of churn. For example, if it is determined that any customer that calls customer support three times in one month is a strong candidate for churn, the company can determine the root cause of the issues that propagated the customer support calls. If they can resolve the underlying issues, they can then reduce the potential customer churn. Likewise, if it is determined that customers with two-year

contracts are extremely low candidates for churn, the company can align their marketing and advertising to push two-year contracts.

I will use the following supervised learning models to predict whether each customer in the dataset is going to churn or not: Logistic Regression, Gaussian Naive Bayes, Support Vector Machine – Linear and RBF, LightGBMClassifier, and XGBoost Classifier.

Benchmark Model

Applying Machine Learning technics to customer data to identify contributors to customer churn is not a new phenomenon. It seems to be an industry standard to use the Logistical Regression model as the benchmark model (2). The Logistical Regression model will provide a benchmark f-score, precision and recall scores, as well as the ROC or area under the curve to be used for model comparison. Likewise, it will provide a benchmark set of features that contribute to the prediction of the target variable, as well as their respective weights.

Evaluation Metrics

The model will be evaluated based on the model's AUC or area under the curve associated with ROC curves. The AUC is a common evaluation metric for binary classification problems. Consider a plot of the true positive rate vs the false positive rate as the threshold value for classifying an item as 0 or is increased from 0 to 1: if the classifier is very good, the true positive rate will increase quickly and the area under the curve will be close to 1. If the classifier is no better than random guessing, the true positive rate will increase linearly with the false positive rate and the area under the curve will be around 0.5. One characteristic of the AUC is that it is independent of the fraction of the test population which is class 0 or class 1: this makes the AUC useful for evaluating the performance of classifiers on unbalanced data sets.

Project Design

The following outlines the steps that I plan to take to analyze the problem, so that I can propose a model that provides insight into the Churn issue.

1. Data Overview and Exploration
    a. Review the variables in the dataset
    b. Provide basic statistics on the data as they relate to churn
        i. Churn percentage for each of the variables

2. Data Preprocessing – manipulate the data (scaling, encoding, etc.)

3. Model Building
    a. Logistic Regression – Benchmark model
    b. Logistic Regression – SMOTE
    c. Gaussian Naive Bayes
    d. Support Vector Machine – Linear

    e. Support Vector Machine - RBF
    f. LightGBMClassifier
    g. XGBoost Classifier

4. Model Performance
    a. Model performance metrics
    b. Compare model metrics
    c. ROC - Curves for models

5. Conclusion - Provide a conclusion statement that declares the best model as well as a discussion of the insights gained.

References

1. "**Facebook is building a camera TV set-top box**." TechCrunch, 17 October 2018, https://techcrunch.com/2018/10/16/facebook-ripley-set-top-box/. Accessed 17 October 2018.
2. **"Google: Your new phone carrier." CNN Money, 1 January 2010.** https://money.cnn.com/2010/12/30/technology/google_wireless_carrier/index.htm**.** Accessed 14 October 2018.
3. "**Churn reduction in the telecom industry**." Database Marketing Institute, 15 October 2018, http://www.dbmarketing.com/telecom/churnreduction.html. Accessed 12 October 2018.
4. "**Managing Churn to Maximize Profits**." Harvard Business Review, 15 October 2018, https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&ved=2ahUKEwi_4Oi1iYneAhUHI6wKHSbbCgoQFjADegQIBBAC&url=https%3A%2F%2Fwww.hbs.edu%2Ffaculty%2FPublication%2520Files%2F14-020_3553a2f4-8c7b-44e6-9711-f75dd56f624e.pdf&usg=AOvVaw37Ykq-D-WIkcESKB5xoaYy . Accessed 12 October 2018.
5. AUC explanation - https://stats.stackexchange.com/questions/132777/what-does-auc-stand-for-and-what-is-it
6. "**CHURN IN THE TELECOM INDUSTRY – IDENTIFYING CUSTOMERS LIKELY TO CHURN AND HOW TO RETAIN THEM**." Database Marketing Institute, 17 February 2017, https://wp.nyu.edu/adityakapoor/2017/02/17/churn-in-the-telecom-industry-identifying-customers-likely-to-churn-and-how-to-retain-them/. Accessed 13 October 2018.
7. "**Reducing churn in telecom through advanced analytics**." McKinsey and Co., https://www.mckinsey.com/industries/telecommunications/our-insights/reducing-churn-in-telecom-through-advanced-analytics. Accessed 12 October 2018.
8. "**Predict Customer Churn – Logistic Regression, Decision Tree and Random Forest**." datascience, 20 November 2020, https://datascienceplus.com/predict-customer-churn-logistic-regression-decision-tree-and-random-forest/. Accessed 14 October 2018.
9. "**Kaggle Dataset."** https://www.kaggle.com/blastchar/telco-customer-churn