Grant T. Aguinaldo
2018-09-03
Capstone Proposal v2

Identifying the Key Features of a Customer Dataset that can be used to Predict Chrun

Domain Background (approx. 1-2 paragraphs)

We live in a subscription-based economy. The global consulting firm, McKinsey & Company noted that:

> *The subscription e-commerce market has grown by more than 100 percent a year over the past five years. The largest such retailers generated more than $2.6 billion in sales in 2016, up from a mere $57.0 million in 2011.*

However, while it seems that things are rosy for these businesses there is a stark reality: churn. Simply put, churn can be defined as the rate at which a company loses customers during a given time period. Indeed, despite record and rapid growth for these subscription-based businesses, churn can bring a subscription-based down, as quickly as it has grown, if companies do not deliver a superior product. Given the importance of the churn rate to the bottom line of any subscription-based business, it is key that business leaders not only track and monitor the churn rate but also identify the key factors that contribute to customer churn. By doing so, it then becomes possible to not only focus the business on customer acquisition efforts but more important than that, are the customer retention strategies that are needed for sustainable growth.

This project was inspired by previous work from the New York University where Deep Learning was used to predict customer churn. For this project, we will develop an approach for identifying the key indicators of a customer data set that most contribute to churn, and will also develop a predictive model, based on the features in the dataset that can predict customer churn. My motivation for undertaking this project is two-fold. First, the large shifts of traditional based business to subscription-based business models (like Adobe) as well as a large amount of activity of funding these businesses by venture captial firms. Second, is the notable acquisition of Whole Foods by Amazon in 2017 for which many believe was a move to strengthen Amazon's understanding of their current customer base via the data that Whole Foods have collected on their customer base. Finally, For this project, we will be using a dataset named "WA_Fn UseC_ Telco Customer Churn.csv "that was obtained from IBM's Watson's Analytics.

Problem Statement (approx. 1 paragraph)

In this problem, we will develop a machine learning model to predict customer churn from a given set of customer metadata (as given in the dataset). On a high level, the problem does have one potential, supervised learning, solution since there is a clear set of independent and dependent variables (discussed further in the following section). In addition, this problem is quantifiable in that all of the features of the data set are either continuous or categorical. In the case that features are categorical and non-numeric, we will employ the suitable methods to convert them into numerical values. The problem is also measurable since there are quantitative methods to assess or measure the quality of the model once developed. Finally, the project will be replicable since there are more than 7,000 customers in the dataset that have a labeled outcome regarding churn (Yes/No) and we use *random_states* within the model so that others can reproduce our findings.

Datasets and Inputs (approx. 2-3 paragraphs)

On a high level, the dataset that will be used includes customer data that can be used to not only predict the behavior needed to retain customers but once done so, you are able to take that insight and develop focused programs to retain those customers. Although this is a sample dataset, it is plausible that the data provided by IBM could come from a customer relationship management (CRM) database. As a whole, the dataset contains a total of 21 fields and 7,044 customers, or examples, that are a mixture of both continuous and categorical values. Of all of these fields (or features), the one that is most important to the project is the categorical field called "Churn." In the data, "Churn" is a binary feature that is either "yes" or "no," and for this project, will be the field that we will attempt to predict using a machine learning model. While there are a total of 21 total fields in this dataset, it is likely that not all of them are important when trying to predict churn. In other words, we suspect that some fields like customer ID are not relevant to the analysis, but fields like contract type to be very relevant to the analysis. For this analysis, we will also employ a type of dimensionality reduction technique (like PCA) to only find the more relevant features that strongly correlate to customer churn. As it relates to the distribution of observations within this dataset, preliminary estimates indicate that there are a total of 1,869 customers (approx. 26%) that have been lost by churn and 5,174 customers that have not. In order to develop the testing and training datasets needed for this project, we will use the stratified shuffle split method within the sk-learn library to split the data and maintain the class imbalance seen in the whole dataset (more can be found here as well).

Solution Statement (approx. 1 paragraph)

To solve this problem we will utilize a supervised learning model (i.e., logistic regression) to predict churn based off of a set of customer data. Specifically, we will use either all or some of the given features of the dataset to predict customer churn. The decision to use a supervised learning model to solve this problem is relevant since we are given labeled data (binary) for each entry that indicates of a customer was retained or not. The solution that we'll employ is not only quantifiable since all inputs to the model will be numerical but it will be measurable since we'll be able to use methods inside of sk-learn like the receiver operator characteristic (ROC) curves and the confusion matrix to analytically assess the performance and quality of the proposed solution.

Benchmark Model (approximately 1-2 paragraphs)

For this analysis, our benchmark model will be a standard logistic regression model as described in the sk-learn documentation where we will set the random_state value to 42. For this benchmark model, we will use all of the fields in the dataset except the churn field as inputs, and we will train the model to predict the churn value. As will be described in the project design section of this proposal, when developing the benchmark model we will use the train_test_split method with a random state to ensure that we minimize any over- or under-fitting of the model. With the benchmark model and the methods to us within the sk-learn library (n.b., cross-validation, learning and ROC curves), we'll be able to measure the performance of the benchmark model.

Evaluation Metrics (approx. 1-2 paragraphs)

For this analysis, the primary metric that we will use to determine the quality of both the baseline and solution the model is the area under the curve or more specifically, the area under the receiver operating characteristic curve (collectively "AUC"). This project involves building a binary classifier that will predict, given a set of customer data, if the customer will churn or not. That being said, we are looking to build a model that will have the highest AUC value (value can be between 0 and 1) out of all of the models that are tested. Within the context of this problem, a higher AUC value means that the model does not only capture all of the true positives but also minimizes all of the false negatives (i.e., we are seeking to build a high recall model). By using a high recall model, that is, a model that minimizes the number of false negatives, we are ensuring that all customers that are classified as being lost to churn have the opportunity to be reviewed by the analyst. In passing we define a false negative as being a situation where the model predicted that a customer was labeled as "no churn," but the customer was indeed lost to churn. Finally, as it relates to the utility of this model, I have taken the position that, it's okay for this model to predict that a customer did "churn" even though it wasn't labeled as such; however, it is not okay for our model to predict that a customer did not churn when in actuality, we did lose the customer due to churn.

Project Design (approx. 1 page)

This project will be broken down into four phases as described below.

Task 1: A cursory review of the data.

Under this task, we will perform a cursory review of the customer data set. This review will seek to understand the overall structure of the data. Some of the subtasks that will be completed during this task include: determining how balanced of a dataset are we working with (in terms of the Churn column), determining the data types of each field of the dataset, determining how complete is the dataset or if we need to impute missing data on any of the fields. During this task, we'll also understand each of the 21 fields provided in the dataset to further gain context about what the data is telling us, and how each field can be used to achieve the overall project goal of building a classifier to predict if a customer will churn or not.

Task 2: Data Preprocessing.

As part of this task, we will complete all of the data preprocessing, if needed, and subset the entire dataset to create the independent and response variables needed to train the baseline model. To accomplish this task, we will need to first complete any preprocessing that is needed to have a complete dataset based on our findings from Task 1. An example of the preprocessing that may be needed for this project can include, imputing missing data, converting text-based fields into numerical ones, or scaling the data to minimize the magnitudes of the cost columns as compared to the other features using the standard scaler. Again, if there are missing values that are identified during task 1, we will impute a relevant missing value like the average of the all of the missing values (at this stage of the project, we have not determined if there are values that are missing in the dataset, but will do so upon the start of this project). As it relates to detecting outliers, we will use the Tukey method of finding any outliers in the data. At this stage of the project, if any outliers are present in the dataset, they are likely to be in the Monthly Chart and/or TotalCharge columns. If outliers are found, they will most likely be removed from the dataset. Following the data preprocessing, we'll need to create two subsets of the dataset The first one will be of all the independent variables which will be of all of the fields in the dataset less the “Churn” and “customer_id” columns. The second subset will be of the response variable for each of the customers in the dataset. Finally, once we have each subset of data, we'll need to vectorize the dependent and independent variables using the one hot encoder method within the sk-learn library.

Task 3: Developing the baseline and solution models.

As part of this task, we will develop the baseline model by first using the train_test_split method on the dataset to create a testing and training dataset and then apply sk-learn's instantiate/fit/predict workflow to generate the model. For the baseline model, we'll use a logistic regression model with the standard defaults as given in the sk-learn library. Once the baseline model is developed, we will create the learning/validation curves and the ROC plot and determine the area under the curve. We will document the findings as it relates to the baseline model based on the shapes of the curves and the AUC value that we obtain from this task. Once the baseline model is developed, we will employ the grid search method to optimize the model's parameters and the value of the AUC value. If we are able to find a model with a higher AUC value than the initial run we will use the optimized version of the model as the baseline model.

Once we have a baseline model (based on logistic regression), we will proceed to selecting two other models (support vector machine (SVM) and XGBoost (XGB)) and proceed to applying the instantiate/fit/predict workflow to obtain the respective AUC values for each model. In the end, the model that that will be used as the solution model will be the one with the highest AUC value. Again, the solution to our problem will be the solution model with the highest AUC value.

Task 4: Determining the most important fields when attempting to predict customer churn.
As part of this task, we will seek to determine the most important fields that are needed to predict customer churn. To accomplish this task, we will use principal component analysis to determine the explained variance in not only the first two principal components but also determine the total number of principal components to obtain an explained variance of more than 90%. Next, we will apply the PCA transform of the first two PC create a biplot that will show us the relationship between the PC and the original features of the dataset. This will allow us to determine the most important features of the dataset needed to predict customer churn. Once we identify the fields that are most important to predicting customer churn, we will create subsets of the data that include these fields proceed to develop a predictive baseline and version of the model using methods that have been previously described. Performance of the model using the determined using the AUC values as previously described.

###