Timothy Lee
Machine Learning Engineer Nanodegree Candidate
April 5, 2017

Capstone Proposal:
Predicting Red Hat Business Value

Domain Background

Per the Harvard Business Review, "deep insights from customer big data should enable highly skilled employees to be more creative and free up time to connect with customers in new ways that add value." Thus, there's a large demand for data science in identifying likely customers for a company based on their internet activity. This is already happening all around us, with the clearest example being recommendations for certain products based on what we click in Google search. The side panels are filled with advertisements that coincide with links we've clicked in the past; these are called targeted advertisements. This is an example of using one's past activity/experiences to drive future purchases. Prior to this new field of data mining, companies had to broadcast advertisements to an impractically large audience and hope that their message reached the likely buyers for their products. This is expensive and there's no guarantee that it would work. Instead, companies can now hire data scientists and engineers to efficiently focus their ad campaigns and intentionally target their products towards the best possible customers.

Article: https://hbr.org/2016/08/using-data-to-strengthen-your-connections-to-customers

Problem Statement

The problem being investigated is how to be more efficient in targeting ads towards likely customers; more specifically, how to use a user's internet activities to predict whether they would buy a product or service. Most companies have data regarding the people who browse their products online, but do not yet have the technology to properly analyze and derive meaning from it. This is a problem because there is a plethora of insight about marketing and customers to be gained from that data.
In the words of the actual Kaggle competition that will be used for this Capstone, participants are tasked with the problem of creating "a classification algorithm that accurately identifies which customers have the most potential business value for Red Hat based on their characteristics and activities. With an improved prediction model in place, Red Hat will be able to more efficiently prioritize resources to generate more business and better serve their customers."

Datasets and Inputs

Link to dataset: https://www.kaggle.com/c/predicting-red-hat-business-value

This data is from a past Kaggle competition sponsored by a company called Red Hat, a software and services company that uses a subscription model.
The following descriptions are information from the Kaggle competition page.
They provided two files: a people file and an activity file. The people file contains each unique person and their characteristics. The activity file has all the unique activities that a person has performed over time. There is a Boolean business value outcome column in the activity file that indicates whether that specific activity by the user resulted in the user fulfilling Red Hat's desired outcome.

The activities are labelled as a certain 'activity_id' but the meaning of the IDs are hidden to us. However, Red Hat does indicate activities as Type 1 through Type 7. Type 1 activities have more information (i.e. 9 characteristics) whereas Type 2-7 activities only have 1 characteristic.

There are 498,687 activities in the test activity file (act_test.csv) and 2,197,291 activities in train activity file (act_train.csv) for 189,118 people in the people file (people.csv).

Solution Statement The proposed solution is to use supervised learning on the given Red Hat data to develop an algorithm that can accurately predict whether a given person and her activities will have high potential business value. Supervised learning uses training data to predict a target variable. The training data for our context is the data provided by Red Hat and the target variable in the data is 'outcome'. With this properly trained algorithm, Red Hat should be able to accurately identify whether a customer has potential business value.

Because there is a large amount of data that is well structured but diverse, I will be using XGBoost gradient boosted trees as the supervised learning classification algorithm to develop a solution for this problem. This is because there are up to 9 characteristics for each activity which indicates that a tree classifier is ideal for classification. In addition, the activities are very diverse (i.e. Type 1 vs. Type 2) so boosting is necessary to reduce error.

Benchmark Model
To objectively validate my results, I will be using a Random Forest model as a benchmark model against my XGBoost model. This is a less robust and simple algorithm that will show that the gradient boosted trees are either more or less effective for this problem.

Evaluation Metrics
Because our problem involves binary classification and the dataset's outcomes are a balanced mix of positive and negative results, the Area Under the ROC Curve (sklearn.roc_auc_score) will be used as the evaluation metric to compare to the benchmark Random Forest model.

ROC curves are great for binary classification visualizations. They plot the False positive rate against the true positive rate for every possible classification threshold, which is the value at which you separate a positive and negative classification. A classifier that separates classes well will have a larger area under the ROC curve.

Project Design Outline
The basic outline for this project will be the following:

1. Load data from separate files into python application
2. Prepare and transform the data
    a. Combine datasets using people_id key
    b. Transform feature set using One Hot encoding
    c. Accommodate for missing data
    d. Load prepared data into a DMatrix for XGBoost
3. Use XGBoost to train the algorithm with the data
    a. Use rounds/early stopping to adequately boost model
4. Test and re-evaluate the model until boosting no longer significantly improves predictions