Samuel Burden
August 15, 2017
Capstone Proposal

Product Classification Using Machine Learning

Domain Background

The ability to correctly classify data is a desire and need for any successful business. Unclassified data makes it extremely difficult for a business to glean very much useful information about their customers, competitors, employees, etc. The task of classifying data by hand though is an extremely large task that proves too costly and time consuming. This is where the power of machine learning is able to help both small and large businesses gain better insights into their data.

Product classification is one data classification problem that most companies face, especially with the proliferation of e-commerce where companies can sell millions of different products. The ability to categorize the products is necessary for financial reporting as well as making it easier for customers to find the products they need. If a customer cannot find a product because it was misclassified this is potential lost revenue for the company. Machine learning makes it possible to look at the various features or attributes of a product and make a prediction about the appropriate product category. According to a recent MIT Technology Review survey it is becoming imperative for companies to embrace machine learning to stay competitive and among companies already using machine learning more than 40% use it for classification of data[1].

Problem Statement

The Otto Group is a large e-commerce company that is looking to properly classify their products into nine different categories. This problem has been take from a Kaggle competition. The problem is how to take data from a subset of their product data and build a system for classifying all their products correctly. Machine learning classification algorithms such as Logistic Regression, Decision Trees, and Support Vector Machines along with deep learning techniques such as neural networks are all potential solutions to this problem. One way to measure the success of the solution is to simply measure the accuracy on a set of testing data.

Datasets and Inputs

The Otto Group has provided some training data of over 60,000 products they currently sell. Each product has 93 features that represent a relevant count for that feature. The feature names have all been obfuscated to protect their competitive advantage, therefore the names of

[1] "Machine Learning: The New Proving Ground for Competitive Advantage", https://s3.amazonaws.com/files.technologyreview.com/whitepapers/MITTR_GoogleforWork_Survey.pdf

the features are simply feat_1, feat_2, …, feat_93. Each of the products also has a target field that contains the product's correct category.

Solution Statement

A solution to the Otto Group's product classification problem is to use a variety of classification algorithms and evaluate which produces the best model based on accuracy of correctly classifying the products. A number of algorithms will be evaluated including XGBoost, AdaBoost, and a neural network. Two metrics that will be used for evaluating the solution will be the overall accuracy of the classification based on the provided data and measuring the log loss of each model. Log loss will be further defined below.

Benchmark Model

Since this was a Kaggle competition we have two different benchmark models and a metric for each benchmark has been provided which can be used to compare potential solutions. The metric provided in the Kaggle competition is the log loss of the model. The first benchmark is essentially just random guessing of the product category and has a log loss of 2.19722. The second benchmark is more sophisticated and uses the ensemble learning method of a Random Forest and achieved a log loss of 1.50241.

Evaluation Metrics

Since this was a Kaggle competition, we have a clearly defined metric that was used to evaluate the various submissions. The log loss metric is one standard metric used in machine learning to help determine how well a model is performing. Generally, the lower the log loss the better the model. Below is the mathematical representation of log loss when multiple classes are being used in classification.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

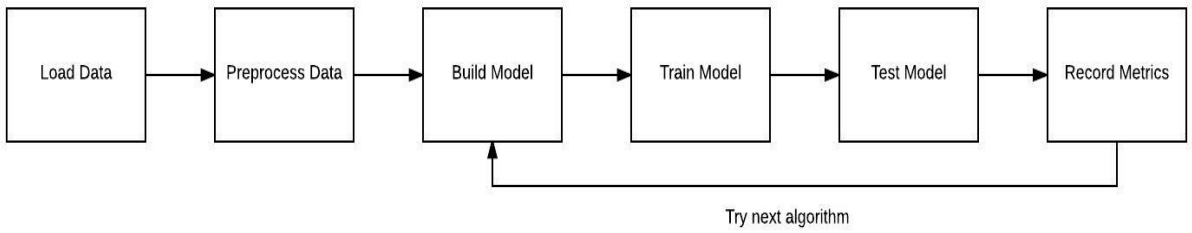
N = the number of products
M = the number of product categories, in this case 9
y_{ij} = 1 if product i is in category j
p_{ij} = the predicted probability that product i belongs to category j

Project Design

There will be three different machine learning techniques used to build models that will attempt to classify the products of Otto Group. The algorithms being used are listed below:

- XGBoost
- AdaBoost
- Neural Network

The below diagram shows the overall workflow that will be used to find the best algorithm:



Here are the steps in the above workflow:

1. Load the data from the training data provided by Otto Group.
2. Preprocess data by making training, validation, and test splits. The target category for a product also needs to be separated from the features. Also, any data normalization needs to be performed.
3. Build the model for one of the algorithms being evaluated.
4. Train the model on the training and validation data splits.
5. Test the model on the test data split.
6. Record the log loss/error and accuracy metrics.
7. Go back to step 3 to build the model for the next algorithm.

Once all the metrics have been recorded we can compare the algorithms to each other as well as to the bookmark metrics to determine which model does the best with classifying the products.