Allstate_Claims

February 18, 2018

1 Machine Learning Engineer Nanodegree

1.1 Allstate Insurance Capstone Proposal

Jose Manuel Garcia
February 18, 2018

Allstate Claim Severity Predictive Model

1.1.1 Domain Background

Allstate is on the largest personal insurance companies in the United states and protect millions of people from potentials damages. When car accidents happen or when other unexpected disasters occur it is a huge relief knowing you are cover by some type of insurance. Recent my girlfriend was involved in a car accident that almost totaled her car. Although she escaped with no bodily injuries her car had thousands of dollars in damages. Luckily, she has full coverage insurance and was able to get all repairs paid for and a rental car without paying anything out of pocket. This event peaked my interest in the insurance business which lead me to this project. Although for this particular instance the insurance claim process was easy and straight forward it is not always the case. This is why Allstate is currently developing automated methods of predicting the cost, and hence severity, of claims. This project will provide insight into better ways to predict claims severity.

1.1.2 Problem Statement

Allstate is attempting to improve its customer experience by automating methods of predicting the severity of the claim. The severity of a claim is based on my different parameters that are unique to every claim. An insurance adjuster takes a look at each claim and all of its parameters and determines how much will be paid out. In some cases the claim is quickly resolved but there is some instances in which it takes a long time to resolve the claim. For this project a train set of data will be used to train a machine learning algorithm to be able to predict the cost of a claim based on the given parameters of the claim.

1.1.3 Datasets and Inputs

For this project Allstate provides a test.csv and train.csv.

The train.csv contains the following: * claim id * cat 1 to cat 116 - this data is category based and contains either single letters or two letters * cont1 to cont14 - this data is a continues set of

numbers that are not negative * loss - this is the ammount that the insurance company has to payout and this is also the target variable * 188,318 rows * 132 columns

The test.csv file containt the same information as train.csv except for the loss because it needs to be predicted. * 89,023 rows * 131 columns

https://www.kaggle.com/c/allstate-claims-severity/data

1.1.4 Solution Statement

The objective is to find the relationship between the 130 features and the value of loss. The first step that needs to be taken is to clean and restructure all of the data. Doing so will help improve the predictive power of the learning algorithm. This is because most algorithms can be sensitive to the distribution of values in the data and can't lead to poor performance if the data is not properly normalized. Normalizing also ensures that each feature is treated equally when applying supervised learners. In this data set, there is also features that are represented by letters. Typically learning algorithms require inputs to be numbers and cannot be training using letters. A solution to this is using one-hot encoding scheme that creates dummy variables for each possible category of non-numeric variables.

The learner that we will be using for this project will be the AdaBoostClassifier because it is great for problems like this in which overfitting can be an issue. The AdaBoostClassifier parameters will be tuned using GridSearchCV and the mean squared error will be used as a scorer. Then the features that provide the most predictive power to the model will be determined. Reducing the number of features to the most important ones reduces the complexity of the model which is what we want. This will be done by using the feature_importance_ attribute.

1.1.5 Benchmark Model

Since this a Kaggle competition hosted by Allstate Insurance the benchmark is set by the best Kaggle score on the leadership boards which is currently 1109.70772 mean absolute error. The goal is to get as close as possible score but as of now I am hopping to score in the top 25%.

1.1.6 Evaluation Metrics

Kaggle will be evaluating this project using mean absolute error and the lower the mean absolute error the better. Of course, there are many other ways to evaluate this model but for the Kaggle competition this is the way the project entries are evaluated.

1.1.7 Project Design

A big part of machine learning is acquiring the data to build your model and making sure it is clean and reliable. For this project Allstate has already done most of this work by providing all of the data for this Kaggle competition. The next step will be to process the data by scaling, normalizing and converting non-numeric data into numbers. This will greatly improve the predictive power of the AdaBoostClassifier. The data will then be shuffled and slip in to training and testing data. The model and now be tuned using GridSearchCV and the scorer that will be used is mean absolute error. Next the most relevant features will be found using the feature_importance attribute. Once the most relevant features are known they will be trained using the model found using GridSearchCV its performance will be compared to the model that used all of the features.

1.1.8 References

1. Allstate Claims Severity, Kaggle
    - https://www.kaggle.com/c/allstate-claims-severity
2. AdaBoosting Classifier
    - https://en.wikipedia.org/wiki/AdaBoost
    - http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
    - https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/
3. Allstate Insurance
    - https://en.wikipedia.org/wiki/Allstate