Kaggle: Allstate Claim Severity Capstone Project

Domain Background

The domain background of this project is the casualty insurance field. Allstate is partnering with Kaggle to free up their customers' time by automating the car insurance claims process. They want to find ways to predict the cost of car insurance claims they receive using data science. This is a project I am especially interested in because Kaggle is a very reputable source for companies looking for data scientists. Furthermore, this competition is a recruitment opportunity and I will be potentially considered for hire if I do well in this competition.

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

Problem Statement

Given categorical and continuous variables, use a machine learning model to regress on the continuous variable "loss".

Datasets and Inputs

The datasets to be used in this competition are "train.csv" and "test.csv". The model is to be trained off of the training set and then evaluated on Kaggle using the testing set. There are 188,318 observations in the training set and 125,546 observations in the testing set. There are 116 categorical variables (labeled "cat") and 14 continuous variables (labeled "cont"). Categorical variables range anything from 2 unique factors to 326. Continuous variables range from 0.48 to 1. No missing values are present in the datasets. It seems specific labels were avoided to remove any sort of domain knowledge advantage.

Solution Statement

The solution that we are looking for is the model that most reduces the evaluation metric for the predicted "loss" values. To do this, we are going to use regression supervised learning techniques. As

single model submissions were only able to achieve a testing score of 1108, I will be attempting to mix multiple models through “ensembling” or using the results of multiple models to improve the prediction on the final model.

Benchmark Model

The current high-scoring public kernel from Misfyre has a cross-validation MAE of 1130 (and a testing score of 1108). I will be basing my models off of Misfyre’s script and hope to improve upon it.

Link: https://www.kaggle.com/misfyre/allstate-claims-severity/encoding-feature-comb-modkzs-1108-72665/code

Evaluation Metrics

Results are to be evaluated off of the Mean Absolute Error (MAE) between the predicted loss and the actual loss. The formula for MAE is a follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \hat{x}_i\right|$$

MAE makes sense for this problem because unlike its close counterpart root mean squared error (RMSE), larger errors from the predictions will not be penalized as harshly. In other words, MAE is looking for how far the predictions are from the median while RMSE is looking for how far the predictions are from the mean.

Project Design

The plan I will implement is the following:

- Use Python to perform exploratory data analysis (EDA) on the training set including:
  - Summary statistics
  - Graphs made with matplotlib and seaborn.
- Create features that will provide more predictive power than just the variables alone including:

    - Transformations of the data
    - Interactions between variables
    - Binary forms of certain variables
    - Clusters created by unsupervised learning methods
- Begin testing different models like:
    - Random Forest
    - AdaBoost
    - XGBoost
- Enhance the model by:
    - GridSearchCV / Random Parameter Search + CV
    - Ensembles

Presentation

The presentation will be a pdf with the following sections:

- Intro: a short description of the goals of the capstone project
- Summary Analysis: a cursory glance at and a 5-number-summary of the data
- EDA: an in-depth analysis of the data, including distribution/interaction analysis, ANOVA/ correlation/significance tests. Features that are created during this process will be highlighted here.
- Model development: a description of the steps I took to prepare, train, and tune my model
- Results: the final prediction MAE achieved by my model.