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.

Problem Statement

Given the 116 categorical variables (labeled "cat") and 14 continuous variables (labeled "cont"), predict the continuous variable "loss". Categorical variables range anything from 2 unique factors to 326. Continuous variables range from 0.48 to 1. It seems specific labels were avoided to remove any sort of domain knowledge advantage.

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.

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.

Benchmark Model

The competition also includes a "All Zeros Benchmark" sample submission (entitled "sample_submission.csv"), which will produce a mean absolute error (MAE) of 3019.7148.

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:

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

The testing set has 125,546 observations.

Project Design

The plan I will implement is to use Python to perform exploratory data analysis (EDA) on the training set, including summary statistics and graphs made with matplotlib and seaborn. During this time, I will also create features that will provide more predictive power than just the variables alone. I will take the results of my analysis and begin building models using GridsearchCV, feature selection, and most likely a Boosted Tree Regressor, such as XGBoost.

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.