Proposal

October 12, 2017

0.1 Udacity Machine Learning Nanodegree

0.2 Allstate Insurance Claims Prediction – Capstone Proposal

Xiang Cao 10-10-2017

0.3 Domain Background

As a important domain of Machine Learning, supervised learning has a wide variety of applications in different areas including computer vision, natural language processing, recommendation system, credit risk prediction, etc.

The essence of Machine Learning is to learn hidden pattern underneath data, data could be unstructured like image, language, or tabular data like stock trading data, customer behavior, etc., or a mixture of them. According to the pattern / knowledge it learnt, we can customize our product and service to people more accurately and efficiently.

In finance/insurance area, there are large amount of user data, with which we can customize finance/insurance product to increase company revenue and improve user experience. Machine Learning provides promising methods to do this.

0.4 Problem Statement

Allstate held a competition on Kaggle, which aims to develop automated method to predict insurance claims cost, and hence severity. The data includes the historical claims cost and customer information for each claim. In this project, I am going to predict claim cost using different machine learning models.

0.5 Datasets and Inputs

The datasets are provided by Allstate on Kaggle. The training set has 188318 records and 132 columns, including unique ID, 116 categorical features, 14 continuous features, and target. All the features are anonymous. The testing set has 125546 rows and columns except target. In this project I split training set into training and validation to train and evalution models.

0.6 Solution Statement

Multiple models are implemented in this project. Including regression, gradient boosting tree, neural network, etc.

0.7 Benchmark Model

For each model with default or empirical parameter, we would get benchmark for each of them. We tweak the models by parameter tuning and ensemble. Compared to benchmark, we could know how much increase those new models got.

0.8 Evaluation Metrics

The cost need be predict is continuous variable. For a regression problem, we usually use Mean Square Error(MSE) or Mean Absolute Error(MAE) as metric. According to the competition evaluation criteria, MAE is used. We are going to build models to minimize MAE.

0.9 Project Design

The project mainly has the workflow: - Expolorary data analysis. - Feature engineering. - Split the data into training set and validation set (cross validation). Building machine learning models, tuning parameters. - The model includes regression, glmnet, random forest, gradient boosting tree, neural network, etc. - Compare and pick the best model for each method. Ensemble/Stack best single models.