Proposal

October 15, 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.

There are multiple machine learning competitions of this domain on Kaggle. Here are some examples.

- Give Me Some Credit. Build models to predict the probability of default, to determine whether or not a loan should be granted. This is a classification problem.
- Allstate Claim Prediction Challenge. An earlier competition held by Allstate. Predict bodily injury liability insurance claim payments based on the characteristics of the insured's vehicle. This is a regression problem.

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. Each record contains both cateorical and continuous features, and the target is the numerical cost of the claim which is continuous. Thus this is clearly an regression program.

In this project, I am going to predict claim cost using different machine learning models (mainly Gradient boosting and Neural Network).

0.5 Datasets and Inputs

The datasets are provided by Allstate on Kaggle. It's open to public here.

The training set has 188318 records and 132 columns, including unique ID, 116 categorical features, 14 continuous features, and target. The target is a continuous variable which indicates the loss of each claim. 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. 5-fold cross validation will be implemented of each model. That is, for each fold, 80% of the data is used as training set, the rest 20% is used as validation set.

0.6 Solution Statement

Data pre-processing is done first, including checking missing values, data transformation, etc. Then feature engineering, including check the interaction of features, two-way interactions, and some three-way interactions. Multiple supervised 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. In this project, I choose glmnet as benchmark model. Compared to benchmark, we could know how much increase these new models got. Both the benchmark model and new models will take 5-folds cross validation.

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 evalution 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. Check the distribution of continuous and categorical variables, check the correction between continutous variables.
- Feature engineering. Create features interactions, select the good ones with the original features to train models. Also, we may transform the target, or even customize the cost function.
- Split the data into training set and validation set (cross validation). Building machine learning models, tuning parameters.
  - The model includes regression, glmnet, gradient boosting tree, neural network, etc.
- Compare and pick the best model for each method. Build two layer stacking model. These single models consititute the first layor of the stacking model. The second layer chooses either Gradient boosing or Keras nerual network.