Machine Learning Engineer Nanodegree
Capstone Project

The Background

Since I want to know how to apply ML in insurance, I searched the Competitions in the Kaggle with keyword "insurance", I find this topic _"Allstate Claims Severity".(https://www.kaggle.com/c/allstate-claims-severity)

The Problem Domain and Statement

My motivation is to use ML to reduce redundant tasks in the insurance field, maybe the customer can be provided with services more quickly. The Allstate can also benefit from this point.

This project wants to create an ML model to predict claims severity. They gave us the train data with more than 130 features to predict severity. Since we want to predict a continuous target variable (loss) with many categorical features, I define this problem as supervised learning and regression problem.

Datasets and Inputs

This project contains 2 csv files. They are:

1. train.csv and test.csv features:
   - id: the index of a training set data
   - cat1 to cat116: category variables(The company will not publish the customer's privacy information, so all column names are not provided.)
   - Cont1 to cont14: continuous variables
   - Loss(The target variable): the amount which the company pay for the customer.
2. In train.csv:
   a) 188318 rows
   b) 132 columns
3. In test.csv:
   a) 125546 rows
   b) 131 columns(the test.csv don't have the loss column)

https://www.kaggle.com/c/learnplatform-covid19-impact-on-digital-learning/data

Solution Statement

- Exploratory Data Analysis
  - We may use some plot method to see the features ,the correlation between different features.
- Preprocess Data

    - Since we have some categorical features, we need to convert them into numbers. So the model can use them.
    - There are many features in the train set, it may result in overfitting. So we may have to reduce the features by using PCA.
- Choose and Train Model
    - First we can use linear regression as the base model
    - We may also use XGBoost
    - To use Grid Search method test hyper parameters

Benchmark Model

We can see leaderboard in the Kaggle project page, the Top solution's score is 1109.70772 MAE.

Then we can use part of training data as testing data to see the linear regression model's accuracy as base. Then we compare next model to see how it works. We will choose a better model to run the test.csv, and upload the submission file to check the score.

Evaluation metrics

As the competition official evaluation is done by Kaggle using mean absolute error, we will use MAE as evaluation.

Project design

- Preprocess Data
    - Since we have some categorical features, we need to convert them into numbers. So the model can use them.
    - There are many features in the train set, it may result in overfitting. So we may have to reduce the features by using PCA.
- Choose and Train Model
    - First we can use linear regression as the base model
    - We may also use XGBoost
    - To use Grid Search method test hyper parameters
- Tools and Libraries used :
    - Python, Jupyter Notebook, pandas, seaborn, XGBoost, scilit learn.