Machine Learning Engineer Nanodegree

Capstone Proposal

Amit Lathiya
April 12th 2019

Proposal

Domain Background

I have chosen project from Kaggle competition to predict severity of Insurance Claim ('Allstate Claims Severity'). Everyone in their lifespan sooner or later has to deal with filing Insurance Claim. It maybe due to several reasons like when devastated by serious car accident, property damage due to weather conditions, professional and personal liability, commercial loss etc. Allstate, a personal insurer in the United States is continually seeking ways to improve Claim service for the over 16 millions households they protect. Pushing papers through Insurance agents is the last place Allstate wants customers to spend time and mental energy. Allstate is currently developing automated methods of predicting the cost, and hence severity, of claims.

Knowing Claim severity helps insurer to align Claim assignments to appropriate groups and teams who have expertise and clearance authority to process such claims faster. Also Claim severity information is used by actuaries to access predicted future cost of claims which can be used in premium adjustments. This is the key strategy which drives the profitability of Insurance company. To remain highly competitive, Insurers must provide competitive pricing model for premiums to customers and knowing claim severity is critical information towards this strategy.

Current challenge most insurer faces is knowing the severity of claim unless very late in process of claim settlement which can take months or even years based on nature and extend of loss. There are several factors internal and external involved to access severity of Claim.

Having more than 12+ years of experience in Insurance Claims IT enterprise product development and delivery I find this project very attractive and challenging. It helps me understand how Insurers can effectively use machine learning to solve challenging problems which can not only expedite Claim processing service but also drive profitability through prize optimization.

References for predictive modeling on Claim Severity:
https://riskandinsurance.com/georgia-pacific/
https://www.investopedia.com/terms/a/average-severity.asp
https://www.casact.org/pubs/forum/05spforum/05spf215.pdf

Problem Statement

Claim severity can be predicted from claim cost that is loss amount (total claim settlement amount). Predicted Loss amount can be approximate function of Claim features. Knowing this value can help insurer categorize claim severity as illustrated in below example table.

Loss range	**Claim Severity**
Loss >= 50,000	High
50,000 > Loss >= $5,000	Medium
5,000 > Loss > 0	Low

Insurers usually have years of prior claim data which can be leveraged to predict cost of claim. We can apply several machine learning algorithms on claim features to predict loss. Given the claim features and loss, this is case of supervised linear regression classification. Model can be trained on claim features to predict loss. Loss error will be difference between predicted and actual cost. Model training goal will be to minimize this loss error function and make final predictions on test data where cost of claim is not known.

Datasets and Inputs

Dataset is provided by Allstate in this Kaggle project. Each row in dataset represents insurance claims. We have to predict the value of loss on test data. Variable prefaced with 'cat' are categorical while those prefaced with 'cont' are continuous. Below are further details on these datasets.

- train.csv – Dataset for model training consisting of 188,318 claims. Loss value is given for each claim.
- test.csv – Dataset for model testing consisting of 125,546 claims. Loss value need to predict for all these claims.

Both train and test datasets have 14 continuous and 118 categorical features. Allstate has not given further details on what these features are, but these could be claim attributes such as age, body part injured, comorbid conditions and facility location as well as text mining the intake notes to determine the likelihood of an adverse development which can be key factors in determining loss value.
Source: https://www.kaggle.com/c/allstate-claims-severity/data

Solution Statement

Given the problem statement and dataset, its case of supervise linear regression in machine learning. After performing feature analysis and engineering (scaling, transforming etc), we can apply several ML algorithms from sklearn library to train model. Goal of training will be to minimize mean absolute error which is difference between actual loss and predicted loss. Dataset can be split into train and validation sets to evaluate performance metric while training. Also, we can run training through deep NN to compare metrics from all trained models to select best model in the last step.

Benchmark Model

We can run training through sklearn simple Linear Regression model without much of feature engineering to establish Benchmark Model. MAE (mean absolute error) score from this training can be compared against fully optimized model score which are trained on engineered features. In benchmarking process, we can ignore very high or low predictions to get median estimates on MAE.

Evaluation Metrics

Since its case of Supv linear regression we can use MAE as metric to quantify performance of both benchmark and solution model. Mean absolute error (MAE) is a measure of difference between two continuous variables Actual and Predicted Loss value. The Mean Absolute Error is given by:

$$\mathbf{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

where yi = actual loss, xi = predicted loss, n = number of rows in samples.

Project Design

Worflow for solution towards solving this problem given below:

Data Analysis

First of first we need to analyze data. Starting with shape of dataset we need to determine type of features (cat vs cont). Analyze statistics on ordinal (cont) features. Check for mean, median and std for all continuous features and targets. Analyze categorical features reviewing number of categorical values involved.

Feature Relevance

Analyze relevance of features by determining correlation between features. For cont features we can use pandas.corr() to understand correlation between cont variables. Correlation will help us to understand similar features which we can reduce later for simplifying training process. We can use seaborn heatmap or pandas scatter matrix to visualize correlation.

Feature Engineering

- One Hot encoding – Before training on categorical features, we need to one hot encode them so that we can feed numerical values to ML model.
- Scaling – Observe any data skewness in cont variable. If there is too much skewness, then normalize skewness through log transformation. Visualize normalization through plot.

Dimensionality Reduction

Training model with large number of features can get very complicated and lead to overfitting. To get good generalization and simplify training process we have to reduce the dimensionality of input

features. This can be done through Principal Component Analysis optimizing number of components selection based off explained variance ratio.

Cross Validation

Train data is split into train and valid dataset. Model prediction is done on valid data (inferencing). We can reserve 10% of data for validation and remaining 90% for training. Applying kfold on partition can alleviate problem of overfitting as each partition is evaluated in training process.

Model Application

Once the data analysis and feature engineering are performed on data, its ready to be feed to various ML algorithm for training. As number of categorical features are more than continuous features we can approach Tree based learning optimization applying bagging and boosting techniques. Some of the ML techniques from sklearn and keras library for given regression problem to evaluate MAE are listed below.

- Linear Regressor for benchmarking
- XGBRegressor
- Decision Tree Regressor
- Deep NN using Keras

Training process is further optimized for each of the model application through hyperparameter tuning to get best MAE.

Conclusion

MAE evaluated from all model application is plotted and best model is chosen (with least MAE) for test data loss value prediction.