Machine Learning Nanodegree
Capstone Proposal

Naveen Balasubramanian

February 04, 2018

1. Domain Background

Insurance has been one of the hot business which sails on disruptions both from Mother Nature and Government Regulations. A very important factor that establishes a new Insurance firm in the market is Customer Service. The situations under which an Insurance firm interacts with the customer are very acute.

Insurance companies have lot of data about the customer claims and claim releases. The data includes all the conditions under which a claim is made and how much claim is valid to be released. This data is really precious that insurance companies are looking for numerous ways to achieve customer satisfaction and hence to gain profit. This is where technology comes into existence to bridge the gap between a business and its customers.

Machine Learning is one promising field which can deal with data in huge quantity and can provide insightful analytics on the data. Using Supervised Learning technique of Machine Learning we can determine an algorithm which can take up a labelled set of data as input and produce the predicted class label as output. The algorithm will learn from all the scenarios on the labelled data so that it could even able to correctly predict the class label or output y for unseen scenarios. This forms the basis for supervised learning. There are different machine learning models build to predict the insurance claims made around the world. One such model is available in the below link.

http://cs229.stanford.edu/proj2017/final-reports/5242012.pdf

In this project let us see how this machine learning technique can be used to automate the trivial insurance claim process making it more faster and hence better customer satisfaction.

a. Motivation

My personal motivation in taking up this project is due to my interest and experience in the banking and insurance sector. I have been working in this industry for the past few years and have acquired a diverse knowledge from different projects I have come across. Machine Learning is a new field I'm focusing on and there are pioneers in my domain who had already tried and succeeded using Machine Learning and AI in banking. But as insurance is a critical field on numbers and money, there is still a huge space where Machine Learning has not penetrated due to trust issues. So, as my personal goal I wanted to develop machine learning models which can automate or speed up the most redundant human intervened process and prove that it is accurate in development and prediction.

2. Problem Statement

AllState Corporation is the largest publicly held casualty and personal lines property insurer in United States. They prime moto is customer satisfaction and they strive hard to serve their customers in a hassle-free way. During some devastated accident situations, filing for claims would be the last thing that any customer would think of doing. AllState is looking for ways to predict the claim cost and hence the severity of the claim. By this way, they can serve their customers more quickly and also with less paper works. The problem statement here is to create an algorithm using Machine Learning techniques which can easily and accurately predict the claims severity for each claim from the available historic data.

Here in our case, the historic data collected during claims from customers will act as the labelled dataset which contains the various details on the situation under which the claim is made. The claim lose amount associated with each labelled data will act at the training output. Supervised learning algorithm is the best fit here which takes the labelled dataset and try predicting the claim lose amount. At one point, we can just use the algorithm to validate the claims posted by the customers with their claim amount and approve it instantaneously if the claim is in the right limit.

3. Datasets and Inputs

Kaggle Dataset Link: https://www.kaggle.com/c/allstate-claims-severity/data

There are 3 files provided – train.csv, test.csv and sample_submission.csv. As this is a Kaggle competition, the sample_submission file is provided just to let the participants know on which format the output should be submitted. In our case, we can ignore this file for now. Let us now see in detail about the other two files. Both train.csv and test.csv, has the below fields.

id	in-built identity column provided in the dataset.
cat1 – cat116	Categorical data. There are no column names and there are 116 different categorical features.
cont1 – cont14	Continuous data. There are no column names and there are 14 different continuous features.

There is a total of 131 features in each of train and test files. The train.csv alone contains the extra labelled class column called loss which is the claims severity value that has been paid to the customer for the different conditions of categorical and continuous data. AllState didn't provide the exact feature names of any column due to data security and protection. Using this train.csv, we will be developing our model and testing it for accuracy with the data in test.csv. To be precise, there are 188,318 rows of data in the training dataset and 125,546 rows of data in the testing dataset. This is very much needed in order to choose the right model for our project.

4. Solution Statement

As explained earlier, the dataset provided by AllState has 130 features and a class variable called loss. Our intention is to come up with an algorithm which can able to predict the loss value given the 130 features. Supervised Learning is a well-suited technique to tackle this problem. But there are certain road blocks which needs to be taken care in order to avoid creating a biased/overfitted algorithm. The first road block is the large number of categorical variables. The different alphabets which forms those categories has to be converted to number so that they can used in model development.

The second road block is the high-dimensional space. We need to reduce the number of features to a small set which can able to accurately predict the loss value. Hence as a preprocessing step, PCA or NMF has to be done over the dataset in order to reduce the dimensionality. Once we confirm the final features, the next task is to test few models like lightGBM/XGBoost to check how they perform in predicting the loss value.

I would like to pick linear regression, lightGBM and XGBoost for this problem as these models have proven high accuracy rate and faster execution time. The linear regression is just picked as a base model to get an idea on how the dataset works and to find the base MSE. I will be using the cross-validation and the grid search technique in order to fine-tune the models.

5. Benchmark Model

As this dataset is from a Kaggle competition, there are lot of benchmark models created by the competitors. The highest model score for this particular competition is 1109.70772 mean absolute error. As we are considering mean absolute error, the lower the value the better is the score. We can start with the linear regression to find the base scores and once we confirm our best model, we can run it on the test dataset provided by Kaggle. We can then submit the project in Kaggle to check the score. My personal goal on this project is to achieve a score not greater than 1125 error on the Kaggle private leaderboard.

6. Evaluation Metrics

The project is evaluated on Mean Absolute Error. As we explained before, the lower the value the better is the model evaluation score. For this project, Kaggle also uses the same metric to evaluate the model and hence we will also concur with the same.

7. Project Design

Data Preprocessing

- Check for missing and null values and remove row if needed.
- Convert all the categorical column non-numeric data into numeric value for using them in feature evaluation.

Dimensionality and Correlation Analysis

- Apply PCA/NMF for dimensionality reduction, so that we will have a good subset of features that can predict the class value.
- Perform the correlation analysis to find the best correlation among different features.

Model Implementation

- Implement the Linear Regression to check the base model performance on the dataset hence the base MSE score.
- Next to implement and train the lightGBM and XGBoost models.
- The models will be optimized using cross-validation and grid search.

8. References

Supervised Learning Technique: https://en.wikipedia.org/wiki/Supervised_learning
Dimensionality Reduction: https://en.wikipedia.org/wiki/Dimensionality_reduction
AllState Details: https://www.allstate.com/about.aspx
Project Details: https://www.kaggle.com/c/allstate-claims-severity/
Benchmark Models: https://www.kaggle.com/c/allstate-claims-severity/leaderboard