Machine Learning Engineer Nanodegree

Capstone Proposal

Maximilian Li

2016 - 10 - 24

Proposal: Predicting Insurance Claims

https://www.kaggle.com/c/allstate-claims-severity

Domain Background

The insurance company guarantees the policyholder a payout if certain damages occur, like a house fire or car crash. For this insurance the insurance holder pays a monthly rate.
To prevent abuse of the insurance policy, the insurance company has many levels of checks in between the insurance claim and the payout. This should prevent fraudulent claims like the policyholder laying fire to his own house or otherwise willfully damaging the insured property.

In the past companies have used their data to automate the approval process of incoming claims https://www.kaggle.com/c/bnp-paribas-cardif-claims-management

In this project, the severance of the claim (amount of money lost for the company) shall be predicted.

Problem Statement

The kaggle-competition provider *Allstate* is searching for an automated way to predict the amount of money claimed by their insuree. For this purposes a model using machine learning shall be developed.

Datasets and Inputs

The company Allstate provided the platform kaggle with an anonymized dataset.
This dataset consists of:
[Unique ID] [116 categorical features] [14 continuous features] [target feature: loss]
With 188.318 datapoints.

Additionally a testing set with 125.546 datapoints is provided
https://www.kaggle.com/c/allstate-claims-severity/data

During development, and for the review, flags will be available to restrict the dataset to 1800/1200 or 18000/12000 entries to speed up the process.

Solution Statement

Using the provided data a regression model shall be developed.

Benchmark Model

The regression model will compete against a benchmark model, which will always predict the average of all the training entries.

Evaluation Metrics

The model is evaluated on the mean absolute error (MAE) between the predicted loss and the actual loss on the test.csv dataset.
https://www.kaggle.com/c/allstate-claims-severity/details/evaluation

Project Design

Using the dataset, several plots of the data will be made, to obtain a better understanding for the data. Based on these plots, normalization and feature selection like PCA might be applied.

3 different Models, like a decision tree or SVM will be trained and tuned with crossvalidation techniques like sklearn's grid-search or train_test_split.

At the last stage, an analysis and performance comparison on the different models in regards to the dedicated testing set test.csv will be provided.