Capstone Report

Problem Statement:

Develop an automated method of predicting the cost, and hence severity, of claims. Model performance is evaluated on the mean absolute error (MAE) between the predicted loss and the actual loss.

Benefits

Main audiences: Allstate Leadership and Claim Department

Since models developed, once deployed, will be going on autopilot. Accurate loss predictions are crucial as they directly impact the profit line of the company. Automated models will reduce the number of manual claims that need to be calculated by the claim department, which will lead to time save on labor hours and minimize human mistakes on calculations.

However, the claim department will still need to monitor and spot checks output as control. Inaccurate predictions will results in significant workloads for reworking and should be avoided by all mean possible.

Indirect beneficiary: Allstate Clients

- The whole idea of automated methods is to improve the speed and accuracy of claims payment to over 16 million households being covered by Allstate policy. Client's satisfaction with receiving fair/accurate reimbursement is the ultimate success metric for the models.

Dataset Description

Obtained from Kaggle, data are in form of 2 dataset

- train: includes 116 category variables and 14 continuous variables. Target variable is column ‘loss.’ Size: (188318, 132)
- test: same as train set, without ‘loss’ column. Size: (125546, 131)
- Data Dictionary: not available as Allstate did not provide definitions on the attributes;

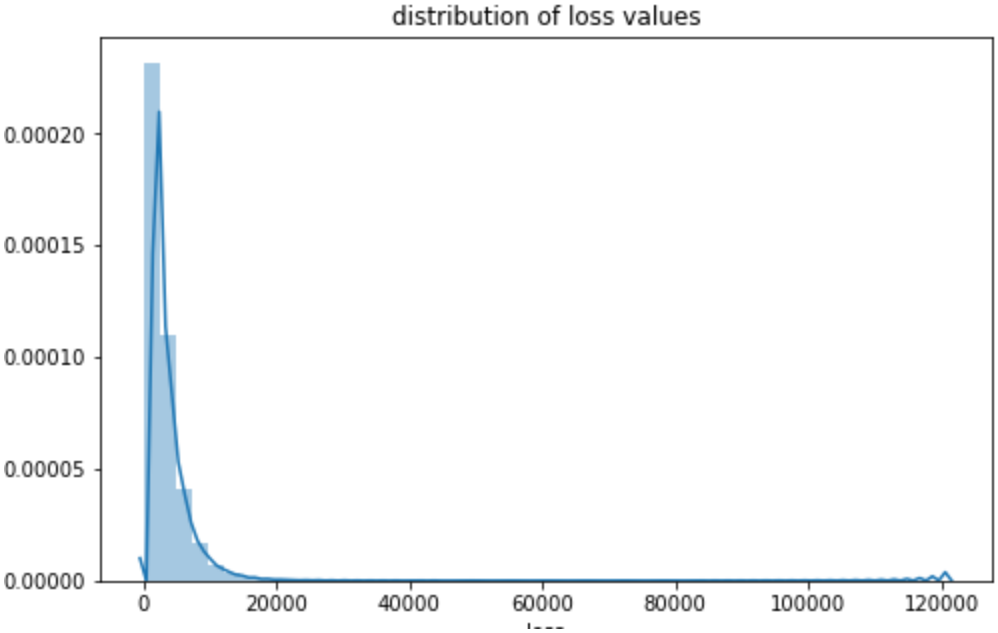
Cleaning and Wrangling

- no missing data. Categories valuables have alphabetical values (i.e., A, B, HK, etc.); continuous variables have values ranging from 0-1. Target variable ('loss') has absolute dollar values

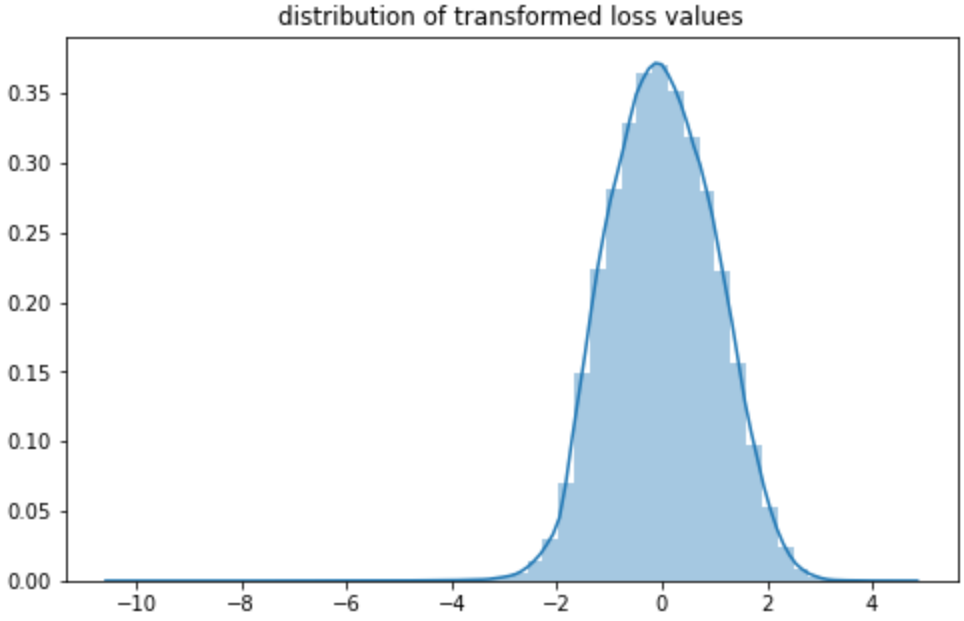
Exploratory Analysis

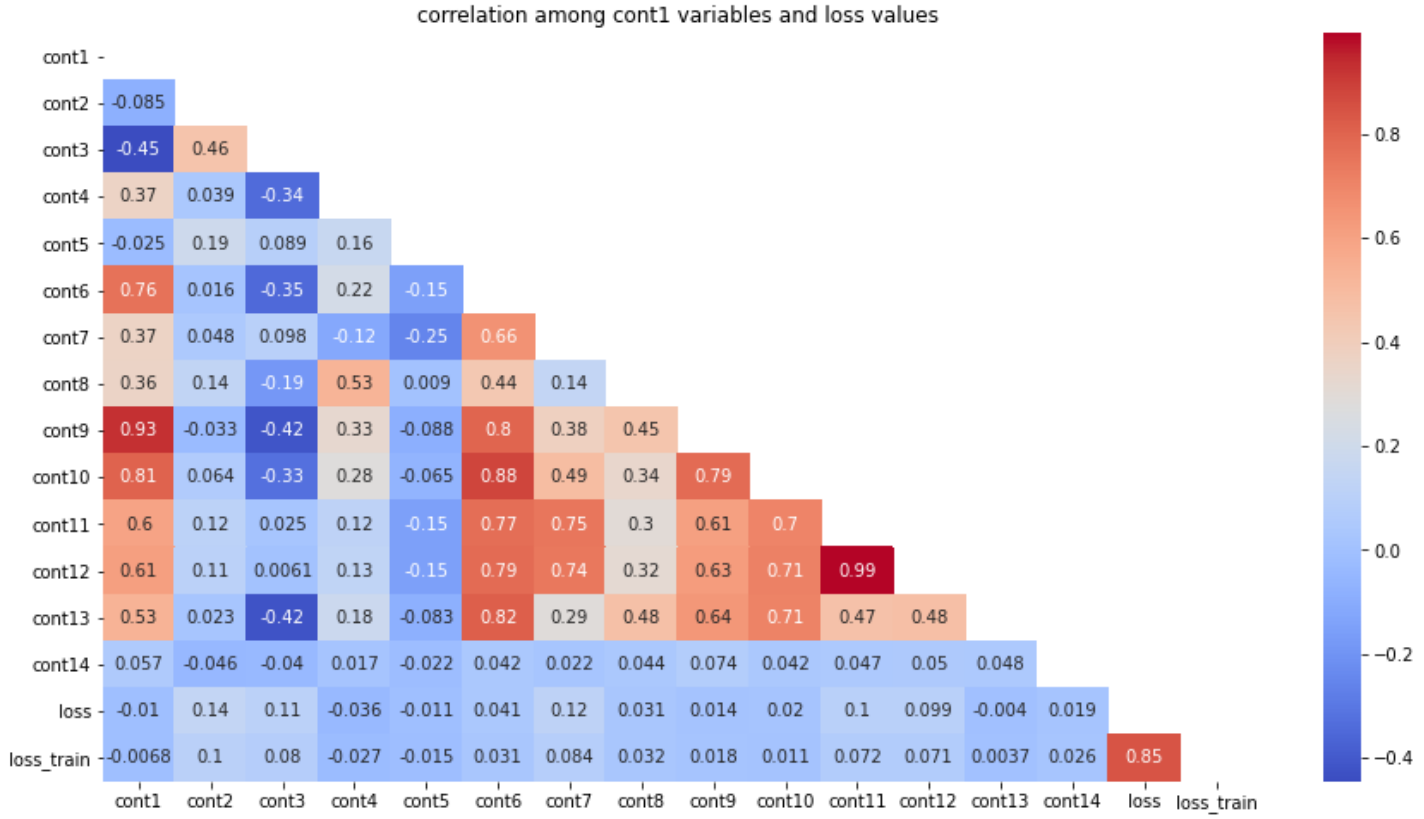
The goal is to discover features that may have predictive power to target variables (i.e.loss), Also, remove features that have high collinearity

Findings:



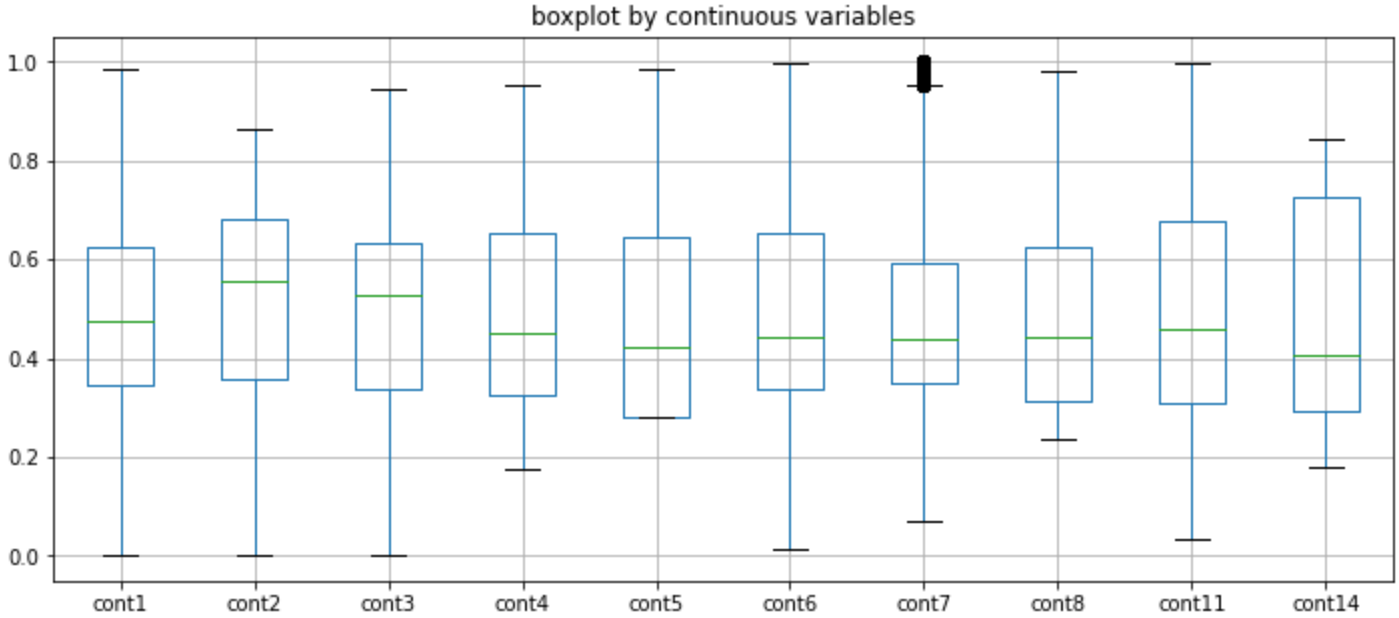
- Target variable 'loss' has distribution skewed to the right, apply Power Transformation during for model fitting
- Post transformation data look like below:





Features collinearity (> 0.8)

- cont1 and cont9, cont10
- cont6 and cont9, cont10, cont13
- cont 11 and cont12
- Drop features cont9, cont10, cont12, and cont13



- Variations exist across all variables. Cont 4, 5, 8, and 14 have highest minimal values. Cont 14 has the lowest mean

Categorical variables:

- There could be colinearity among the categorical and numeric variables. Unfortunately, we don't know the definition of the variables hence we cannot draw direct inference on that.
- It is also difficult to analyze 116 features one by one. Instead, we will use recursive feature elimination (RFE) to narrow down variables based on feature importance

Summary

- Target variable will require power transformation during model fitting
- Drop features cont9, cont10, cont12, and cont13 due to collinearity to other existing variables.
- Categorical features will be selected using RFE techniques

Machine Learning

Strategy

1. Identify the best machine learning algorithm based on continuous variables to obtain baseline performance
2. Add categorical variables and apply RFE to filter low power features
3. Fine tune models through hyperparameters tuning

Part 1: Training only with continuous variables

In this part, we want to get a quick read on performance by type. Early indication suggested that boosting models performance the best. Tree based models and linear models have equivalent performance but tree based models take a longer time to train. We will add categorical variables in Part 2 to see if performance ranking will change.

Part 2: Add categorical variables

- Applied one-hot encoding to all categorical variables, a total of 1033 features are formed
- Applied LassoCV, followed by RFE using Lightgbm and Xgboost estimators, narrowed down to 123
- Applied RFECV using Lightbgm as estimators, get feature numbers down to 112

With these newly defined features set, we reran all models in Part 1 and added Catboost. Results on the next page:

Model Performance

(Train data with CV =5, no tuning)

Model	MAE	Time
Dummy	1809	1s
Linear Reg	1278	3s
Ridge	1278	3s
SGD	1278	3s
Extra Tree	1249	10 mins
Random Forecast	1219	23 mins
Ada	1653	1 min
Lightgbm	1161	1 min
Xgboost	1217	3 mins
Catboost	1157	19 mins

Observations: Overall, the model performance ranking stays the same with the addition of categorical variables. Boosting perform the best. Tree based models now have obvious performance over linear models, but the training time is too long, and we won't pursue further.

Catboost also has a long processing time but it has the best performance, we will keep that to see how much more performance we can get from it. Lightgbm and Catboost remain solid models.

Part 3 - Hyperparameters tuning

In the last part, we will perform grid-search on SGD Regression, Lightgbm, Xgboost, and Catboost on parameters such as learning rate, n_estimators, and max_depth.

Below are the **Kaggle** test results:

Model	MAE	Time
Dummy	1783	1s
SGD Regression	1266	3.6s
Lightgbm	1129	12s
Xgboost	1149	44s
Catboost	1122	38s

Recommendation/Next Steps

Catboost has the best MAE among all models, followed closely by Lightbgm.

While the evaluation metric is MAE, clients should be aware it is the average of above 120k insurance claims. MAE understated the impact from outliers to customer satisfaction, particularly for claims payments that are significantly underestimated from the true reimbursement that clients should get. On the other hand, there are low-value claims with high loss predictions. That will result in unnecessary final burdens to Allstate.

The model can be further improved if the definitions of the categorical and continuous variables are given. It will help practitioners to perform more precise features selection and target the outlier issues with a better context.

From a high-level exploratory analysis on the outliers, over predictions tend to happen more often at the low-value claims, while under predictions happen toward claims $15K or more in values.

We recommend Allstate to perform audits for claims that have high predicted values (threshold to be determined by Allstate management based on tolerance level). Make it is easy for customers to report inaccurate claims to analyze the miscalculated factors and adjust models accordingly.