**Allstate Claim Severity**

**Abasiekeme Attang**

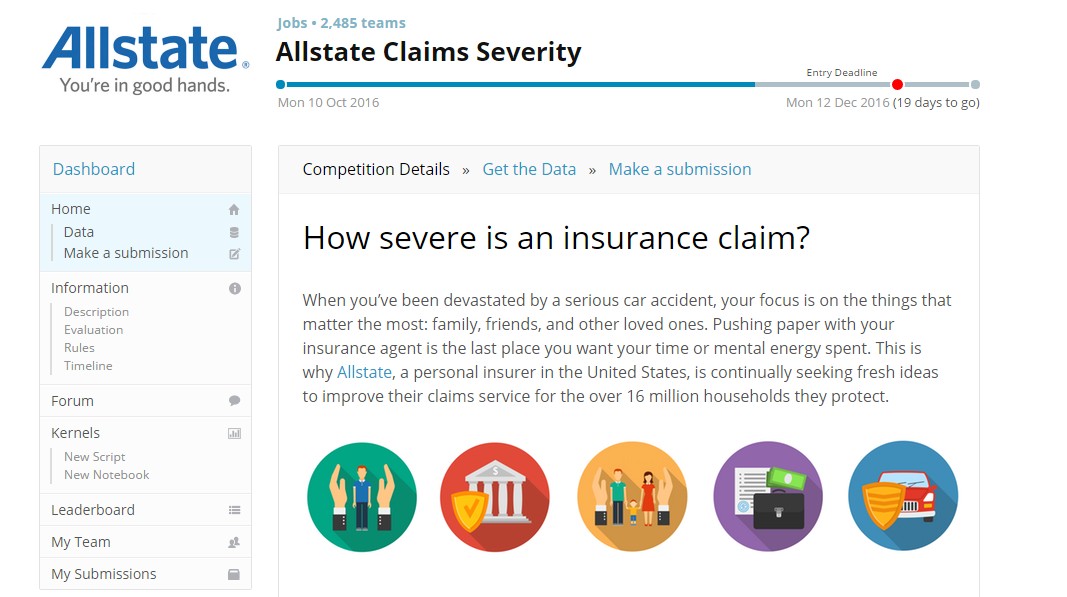


Table of Contents

**Executive Summary** ........................................................................................................................3

I. Introduction............................................................................................................................3

II. Methodology ........................................................................................................................3

III. Conclusion ..........................................................................................................................3

**Exploratory Analysis**.......................................................................................................................4

I. Target variable - Loss.............................................................................................................4

II. Categorical Variables ............................................................................................................5

III. Continuous Variables ..........................................................................................................5

**Dimensionality Reduction** ..............................................................................................................7

I. Feature Selection ...................................................................................................................7

II. Principal Component Analysis .............................................................................................8

**Data Modeling** .................................................................................................................................9

I. Linear Regression ..................................................................................................................9

II. Ridge Regression & Lasso Regression.................................................................................9

III. Elastic Net Regression and SGD.......................................................................................11

IV. Random Forest Regression................................................................................................11

**Conclusion** .....................................................................................................................................13

**Executive Summary**

**I. Introduction**

The Allstate Claim Severity is a data science challenge on Kaggle 1 . The goal of the challenge is to predict the cost of the insurance claims given a set of anonymous variables. A training set with 188 thousand records and a testing set with 125 thousand records were provided and the competitors were asked to generate predictions on the testing data based on their models built on the training data. Mean absolute error is used to evaluate model performance.

**II. Methodology**

An exploratory analysis was thoroughly done to understand the data as well as pre-process it for data modeling. Feature selection and principal component analysis were conducted to reduce the dimensionality after one hot encoding the categorical variables. Then K-means clustering was applied to increase the prediction power. In the end, different models (linear, ridge, lasso, elastic net, sgd and random forest regression) were built, tuned and compared.

**III. Conclusion**

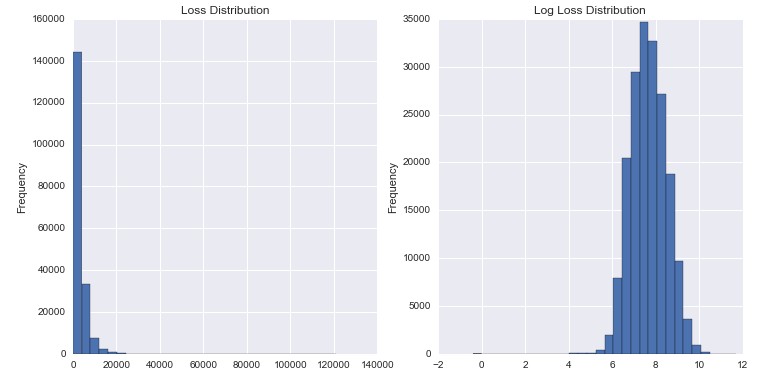
After feature selection, the main noises of the original data was removed and hence my models performed better. The difference between using principal component analysis and feature selection was not statistically significant. Besides, K means clustering didn’t help my model obtain higher accuracy. For model performance, linear, ridge, lasso, elastic net and sgd had almost same accuracy, which was around 1270; while random forest was more powerful

and its best Mean Absolute Error was 1220.

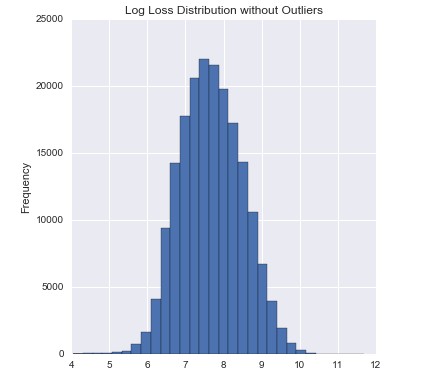
1 https:/[/www.kaggle.com/c/allstate-claims-severity](http://www.kaggle.com/c/allstate-claims-severity)

**Exploratory Analysis**

**I. Target variable - Loss**



As we can see from the plots above, the original distribution of los is significantly skewed to the right. And it is still a little skewed after the log transformation is applied. So I found and removed 17 outliers and the distribution is close to normal.



**II. Categorical Variables**

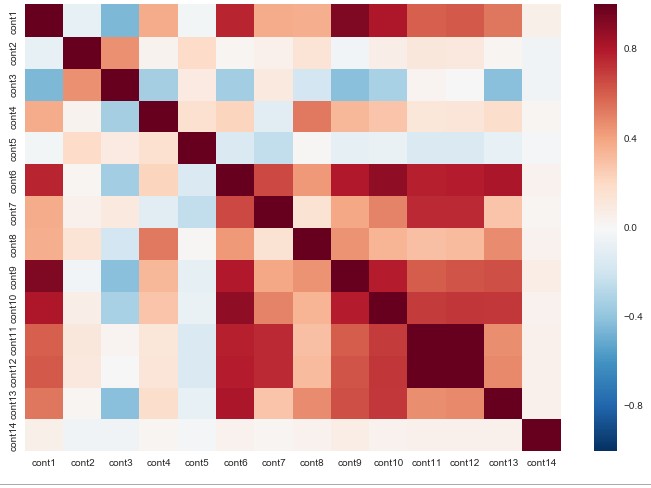
|  |  |
| --- | --- |
| binary | 72 |
| 2 – 10 categories | 10 |
| 10 – 20 categories | 11 |
| more than 20 categories | 6 |
| **Total** | **116** |

There are 116 categorical variables in this dataset. 72 are binary, 27 have 2 to 10 categories,

11 have 10 to 20 categories and 6 have more than 20 categories. This will be an issue for my analysis as those categorical variables will generate more than 1000 dummy variables after applying one hot encoding. Dimensionality reduction is needed here and it will be addressed later.

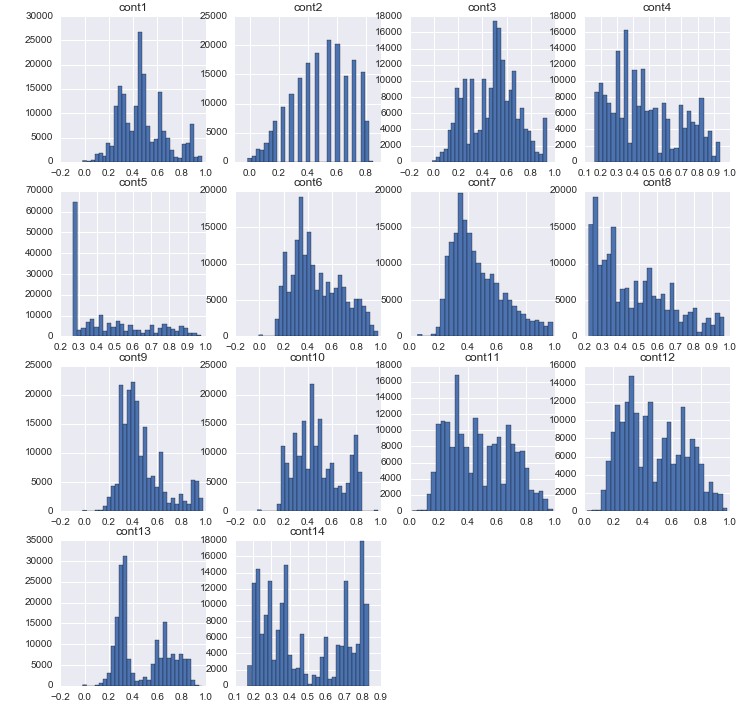
**III. Continuous Variables**

I have 14 continuous variables in total. I plotted the correlation matrix as below.



From the plot above, there are some strong correlations between those numerical variables

Since I am dealing with a regression problem, multicollinearity might be a problem. But I will use feature selection or principal component analysis to take care of it.

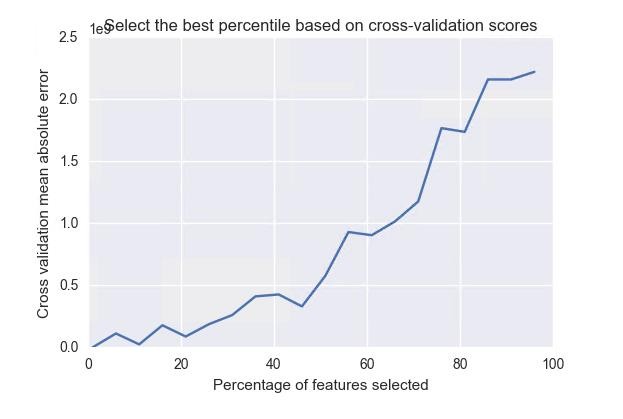


I also checked the distributions of those numerical variables, most of them are not normal as we can see from the above plot. So I normalized them using sklearn.preprocessing.normalize.

**Dimensionality Reduction**

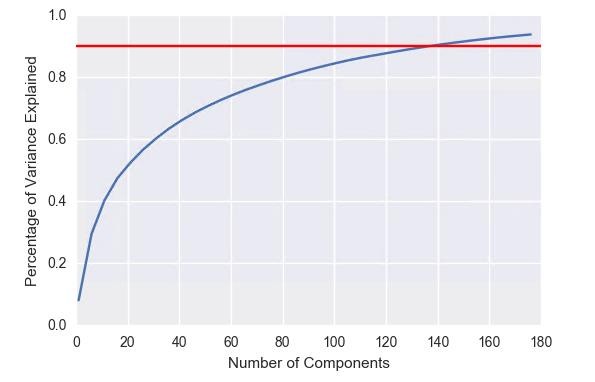
I tried two approaches to reduce the dimensionality of the data. One is feature selection using linear regression and the other is principal component analysis. My goal is to compare the model performance of each method and select a better one to be used in my final model.

**I. Feature Selection**



Using linear regression, I cross-validated the model performance for different percentage of features selected. In the plot, the mean absolute error increases linearly and this is because some added features are noises. In this case, the best percentile is 11% and 127 features are chosen after feature selection.

**II. Principal Component Analysis**



As the data is very sparse, I needed 130 components to capture more than 90% percent of the variances.

**Data Modeling**

**I. Linear Regression**

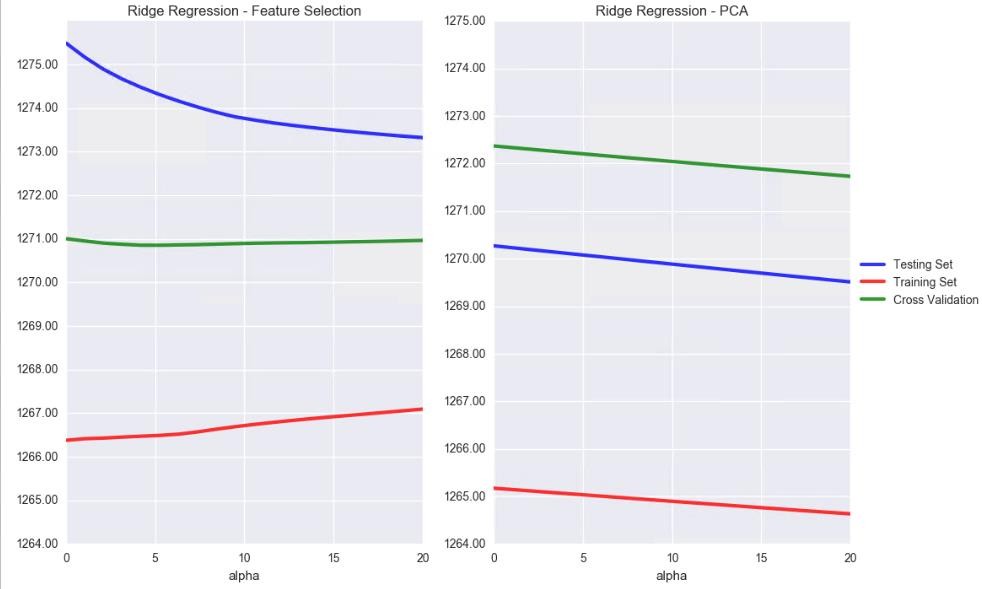
|  |  |  |
| --- | --- | --- |
| **Method** | **Training MAE** | **Testing MAE** |
| Model with selected features | 1287.224 | 1286.9260 |
| **Model with PCA** | **1278.276** | **1309.7990** |

In linearly regression, modeling on pca components gave up a smaller Mean Absolute

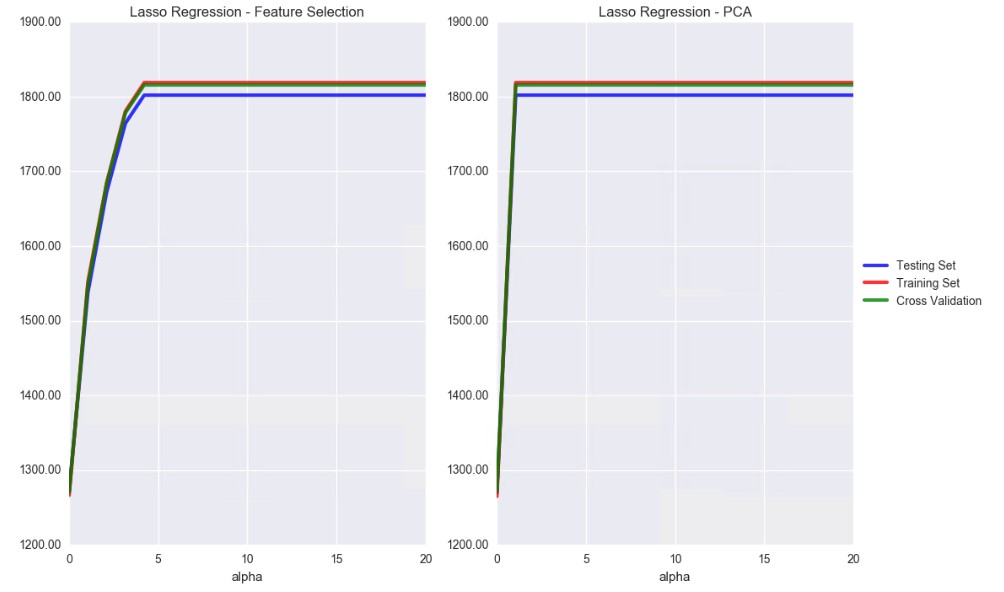
Error on both training set and testing set.

**II. Ridge Regression & Lasso Regression**

**Ridge Regression**



**Lasso Regression**



According to the two model performance plot above, we can see that no matter using selected features or pca, the highest accuracy of the ridge and lasso model is about the same. However, the performance of the ridge model is very stable when alaph is tuned; while that of the lasso model changes significantly.

Linear, ridge and lasso regression are those of simplest regression models that they don’t have many parameters to tune with. That’s why their performances are roughly on the same level. In the following section, I tried some more complicated models to see if I could decrease the Mean Absolute Error.

**III. Elastic Net Regression and SGD**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Method** | **Training MAE** | **Testing MAE** |
| SGD | Feature Selection | 1274.03 | 1277.18 |
| PCA | 1272.89 | 1279.95 |
| Elastic Net Regression | Feature Selection | 1273.11 | 1270.02 |
| PCA | 1269.72 | 1272.14 |

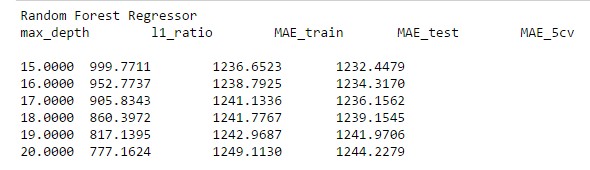
For elastic net and stochastic gradient descent regression, I used sklearn.gridsearchCV to find the best combination of parameters. Unfortunately, both elastic net and SGD were not able to increase accuracy. This is partly due to that the two models are built on top of ridge, lasso regressions. Their performance won’t vary a lot.

**IV. Random Forest Regression**

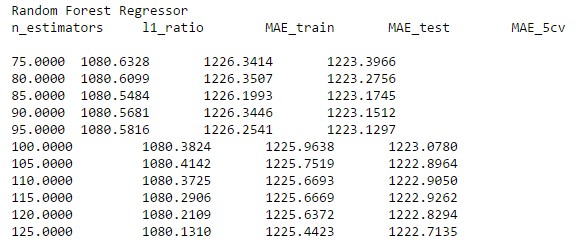
For random forest regression, I manually inspected the parameters instead of gridsearch. Luckily, the power of ensemble did decrease the Mean Absolute Error. In general, random forest is a stable and powerful model that can achieve relatively higher accuracy.

**Maximum Features:**

**Maximum Depth:**



**N Estimators:**



In the end, the best parameter combination and the corresponding performance was recorded in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Method** | **Training MAE** | **Testing MAE** | **CV MAE** |
| Random  Forest | Feature Selection | 1080.71 | 1274.32 | 1220.15 |
| PCA | 1090.35 | 1230.12 | 1222.34 |

**Conclusion**

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Model** | **Best MAE** | **Mean MAE** |
| **Feature Selection** | Linear | 1275.02 | 1263.92 |
| Ridge | 1270.85 |
| Lasso | 1270.35 |
| Elastic Net | 1270.02 |
| SGD | 1277.18 |
| Random Forest | 1220.15 |
| **PCA** | Linear | 1270.26 | 1264.72 |
| Ridge | 1271.73 |
| Lasso | 1271.97 |
| Elastic Net | 1272.14 |
| SGD | 1279.95 |
| Random Forest | 1222.34 |

As we can conclude, the difference between modeling with 130 principal components and

127 selected features is not statistically significant. Different models obtained approximately same accuracy for each dimensionality reduction method. In the end, the random forest model outperformed all the other models and achieved the lowest mean absolute error.

I submitted my predictions of the random forest model on kaggle. The MAE on the public leaderboard is 1217.52, which beats the random forest benchmark.

