# Project Report

**Ming Chen**

After performing a series of Data Exploration and Analysis, Feature Selection, and Model Tuning and Selection (more details can be found on my GitHub), I would suggest the following two models:

## Model

### Model 1: Linear Regression Model

The final Linear Regression model has **28 features**. The feature names and coefficients are shown as below. The model is saved in the '**results/model1.sav**' and can be loaded to make predictions on new data.

| intercept | x1 | x15 | x19 | x2 | x25 | x29 | x31 | x42 | x5 | x59 | x62 | x72 | x74 | x76 | x84 | x87 | x98 | x99 | x28 | x78 | x3 | x55 | x7 | x53 | x14 | x69 | x79 | x80 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00037062 | 60.85 | 15.9 | 108.7 | 93.07 | 49.2 | 123.5 | 67.66 | 95.05 | 91 | 69.23 | 12.27 | 84.83 | 77.71 | 40.11 | 69.01 | 45.23 | 93.98 | 49.93 | 76.68 | 26.61 | 25.82 | -27.3 | 21.62 | 18.82 | 29 | -31.6 | 44.84 | 22.34 |

### Model 2: Gradient Boosting Regression Model

The final Gradient Boosting Regression model consists of 100 decision trees. Feature importance can be found in the analysis report. The model is saved in the '**results/model2.sav**' and can be loaded to make predictions on new data.

## Model Comparison

- **Performance:** Model 1 has better performance than Model 2. Model 1 performed well on both training and testing datasets. (Model 1 train error = 0.997, test error = 0.997; Model 2 train error = 0.977, test error = 0.924).
- **Interpretability and complexity:** Model 1 is a linear model, simpler and more interpretable than the Model 2, which is a ensembled tree-based model.
- **Improvement:** it's possible to improve the performance of Model 1 by adding more features to the model. In this project, I only used the minimum number of features that achieves an R2 score larger than 0.995.
- **Training time:** it was much faster to train Model 1 than Model 2 (see the analysis report).

## Summary and Conclusion

**I would suggest Model 1 - the Linear Regression Model**. The linear regression model has the highest performance among other models. It is simple and easy to train and interpret. It is possible to further improve the performance by increasing the model complexity - adding more features and interaction effects. Model 1 should perform well on the "test.csv" data because features from "unseen" data has very similar distribution as the "train.csv" data. Compared with Model 1, Model 2 has a lot of hyper-parameters to tune and gradient boosting machines are easy to get overfitted. When training the linear model, I didn't create new features, add polynomial terms and feature interaction effects to the model. There are several reasons: 1) With the currently available features, the model already performed very well. 2) I didn't find obvious non-linear relationship between explanatory variables and the target variables. The exploratory analysis showed that most explanatory variables have a linear relationship with the target variable. 3) The tree-based model Model 2 considers feature interaction effects by nature, but still has lower performance.