Suicide Rate Prediction Around the World

Ye Jin, Yuren Dong

I. Introduction

Suicide has become one of the leading cause of death around the world. According to the National Institute of Mental Health, “suicide was the second leading cause of death among individuals between the ages of 10 and 34, and the fourth leading cause of death among individuals between the ages of 35 and 54” (2020). What make it worse, suicide rate has always been increasing over the past 30 years. Luckily, society started to put more attention on mental health issues and social scientists have been investigating the cause and solution of this problem. With the help of machine learning, data scientists have been able to use Decision Tree Analysis to find the major predictor of suicidal thoughts in Korean adults (Bae, 2019). The study showed that financial difficulties and depression are the top 2 factors related to suicidal ideation within Korean adults. In this project, we performed the analysis on suicide rates over 30 years from 1985 to 2015 across 100 countries with different demographic and socioeconomic features recorded. We applied linear regression, decision tree, and random forest modeling on our preprocessed data. Comparing the results from these three different approaches, we hope to retrieve more information behind the suicide data and predict future suicide rates by investigating the driving factor of this phenomina.

II. Methods

Data Wrangling

In this project, we used Python to analyze our data. After retrieving the data from Kaggle dataset “Suicide Rates Overview 1985 to 2016” (2018), we first checked the nan values for each feature. Realizing that 70% of the data are missing the “HDI value” column, we considered it an incomplete feature and decided to drop it. Then we replaced country names with their corresponding longitude latitude to better represent each country, while we also got this data from another kaggle dataset named “Counties geographic coordinates”. In order to successfully perform the regression models, we transformed the categorical data in our samples to numerical data. Specifically, we used

“LabelEncoder” to transfer features “year” and “age” and used “OneHotEncoder” to transfer the feature “sex”. After checking the the data distribution, we realized that the standard deviation for suicide number, population, gdp, and gdp per capita are significantly high. Thus, we used Z Score to normalize these features.

Exploratory Data Analysis

We performed exploratory data analysis to explore the characteristics of our data sets. Firstly, we separated the female and male samples into two groups and calculate the average of suicide number per 100k population over past 30 years. We are using the suicide number per population here as female and male might have different size of population, which the suicide number might be dependent on. At the same time, we also separate our data based on their age groups and calculate the average of suicide number per 100k population over past 30 years.

Regression Modeling

To begin with regression modeling, we set the 'suicides/100k pop' as our dependent variable and demographic features including 'longitude', 'latitude', 'year', 'age', 'female', 'male','population','gdp', and 'gdp/p' as our independent variable. Then we used “train_test_split” to set our test data and train data or both dependent and independent variables with test_size=.2.

1. **Linear Regression Model**

   In this step, we fit our normalized data into the Linear Regression Model and use correlation and root mean squared error to evaluate our prediction. First, we fit our trained data into the linear regression model using the “LinearRegression” function. Then, to test our model, we compare the prediction result from the linear regression model with the actual data through visualization. Motivated by “thrashman2324” from Kaggle (2019), we evaluated our model using root mean squared error between the prediction and actual values. At the same time, motivated by “nilskaizen” from Kaggle (2020) , we also interpreted our results by finding the weight coefficients of the model. Specifically, we used “coef_” to summarize and analyze our weights. Additionally, based on our interpretation from exploratory data analysis, we also run separate models for male and female using the same set of analysis.

2. **Decision Tree Model and Random Forest Model**

   We first fit the Decision Tree model and Random Forest model with the testing and training dataset generated using “DecisionTreeRegressor” and “RandomForestRegressor. Then we recorded feature importance for analysis in the next section. In addition, we plotted prediction vs testing suicide rate and computed the root mean squared error of the prediction

III. Results

Exploratory Data Analysis

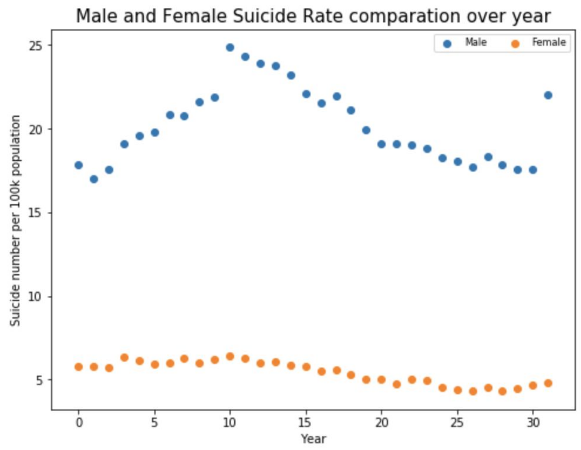
In this step, we separated our data in terms of sex and age and plot out the average suicide rate per 100k population with respect to years.

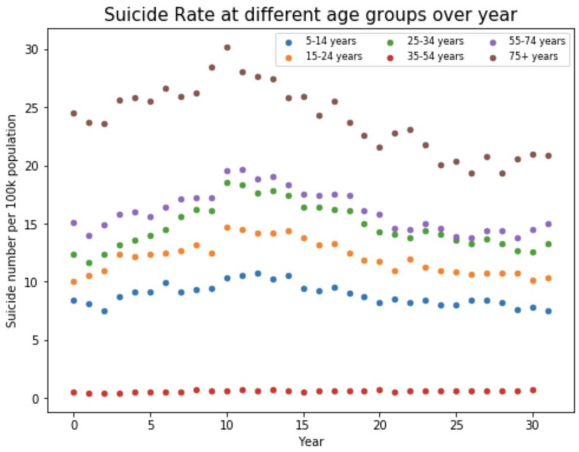
In *figure 1*, we can clearly identify that male population around the world has significantly higher number of suicide rate for the past 30 years (1985-2015).

In *figure 2*, we can realize that people over 75 years old surprisingly have the highest suicide rate. On the contrast, people in their middle age around 35-54 years old have really low suicide rate while the rest of the age group are in the middle.

Figure 1: Male and Female Suicide Rate comparation over year (left)

Figure 2: Suicide Rate at different age groups over year (right)





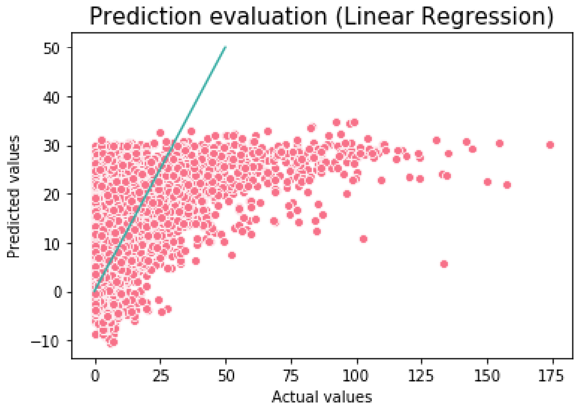
Model Performance Comparison

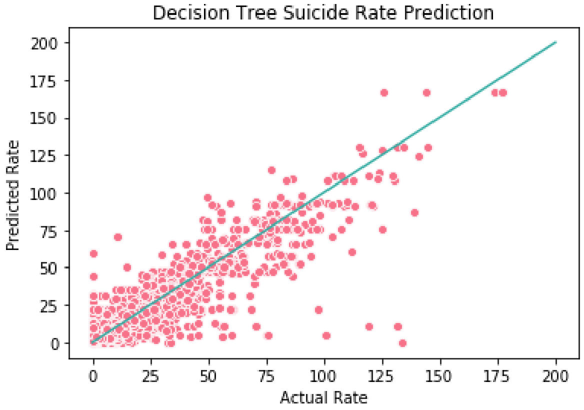
Overall, the linear regression model didn't work as well as the decision tree model and the Random Forest model. For root mean square error, we got 16.4295 in linear regression model, 7.7163 in decision tree model, and 6.9084 in random forest model. At the same time, following the tutorial of Haithem Ben Rhouma from Kaggle, we plotted the values predicted by model against the actual value to evaluate the performance of our model (2020). Overall, linear regression model shows a low correlation between the two variables where some of the variables are correlated while the others don't. Decision Tree model and Random Forest model were able to capture more information and perform a better prediction, as we can see that more data points fall around the 50-50 line which indicates the perfect prediction.

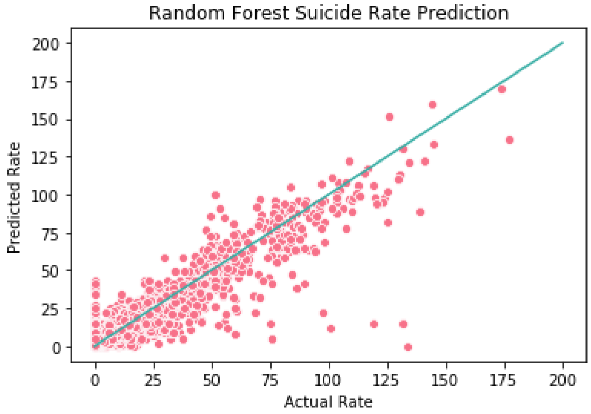
Figure 3: Linear Regression Suicide Rate Prediction Evaluation (left)

Figure 4: Decision Tree Suicide Rate Prediction Evaluation (right)

Figure 5: Random Forest Suicide Rate Prediction Evaluation (right)







Feature weights

We took a look at the weights output from different models.

In the linear regression model, we retrieved the coefficients in regression model and concluded that sex and age are good predictors of suicide rates as they possess high weights in the linear model, which validates our result from the exploratory data analysis part.

In the decision tree model and random forest model, we directly extract the feature importance data from the regressor, and have them listed as following: we notice that age, latitude,longitude, and sex are the deciding factors for determining the suicide rate.

Features	longitude	latitude	year	age	female	male	population	gdp	gdp/p
Linear Regression	0.0237	0.0780	-0.0523	1.3398	-4.9164	4.9164	0.0738	0.3967	-0.8597
Decision Tree	0.1571	0.2319	0.0394	0.2654	0	0.1719	0.0769	0.0298	0.0277
Random forest	0.1529	0.2240	0.0373	0.2561	0.0678	0.1024	0.0881	0.0348	0.0367

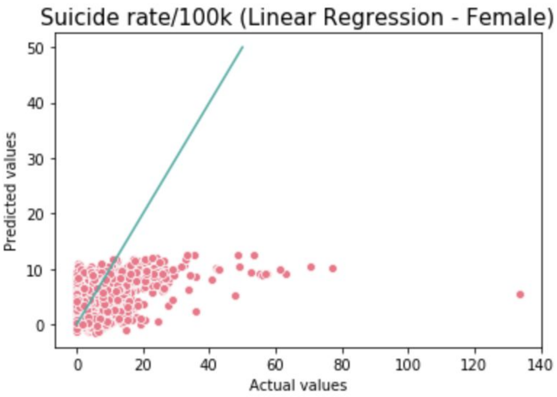
Linear regression

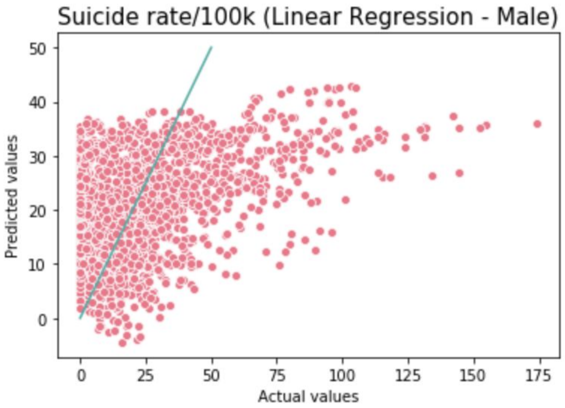
Female and male seem to be the leading factor of our linear regression model, which doesn’t align with the leading weight we got from the decision tree model. In this case, we want to test the importance of sex to the overall model and check if it is an actual predictor of suicide rates. Based on these information, we decided to seperate our data by sex, and run our model on data with female only, male only, and withouth sex characteristics respectively. When we separate our data by sex groups, linear regression model seems to perform better for female compared to male where we got a root mean square error of **7.3978** for female and **22.4722** for male. We suspected that this might be caused by the fact that male population overall has more suicide cases and more diverse data background. When we drop the sex characteristics from the train data, we got a root mean square error of **18.2647**, which is worsen compared to the model with sex characteristics. Thus, sex is a true leading predictor in linear regression model.

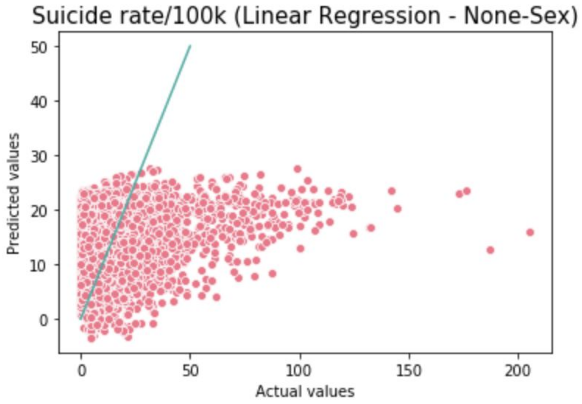
Figure 6: Female Suicide rate/100k: Predicted vs Actual (Left)

Figure 7: Male Suicide rate/100k: Predicted vs Actual (Right)

Figure 8: Suicide rate/100k without sex characteristics: Predicted vs Actual (bottom)







Based on the weights we got from these three models, we can realize that without the consideration of sex, age is always the leading predictor of suicide rate while gdp is the second leading predictor. However, gdp per population surprisingly is negatively related in these models.

Features	**longitude**	**latitude**	**year**	**age**	**population**	**gdp**	**gdp/p**
Weights_ Female	0.0214	0.0429	-0.0857	0.8859	0.1355	0.2228	-0.024
Weights_ Male	0.0460	0.1928	-0.0519	3.4528	0.3871	0.6852	-2.4841
Weights_ non-sex	0.0356	0.1204	-0.0705	2.1332	0.0752	0.6162	-1.2558

Decision Tree and Random Forest

Since the analysis of suicide rate and association is already performed in the linear regression part, it will be skipped for this section.

We're especially interested in age, latitude, longitude, and gdp.

Age is the factor of largest feature importance for these 2 models, thus we want to validate its significance through testing a model without the feature. The root mean squared error increased from 7.573 to 11.396 for Decision Tree, and 6.886 to 10.205 for Random Forest. We have the distribution of change plotted in *figure 9* and *figure 10.*

Figure 9: Decision Tree: with age on the left, without age on the right

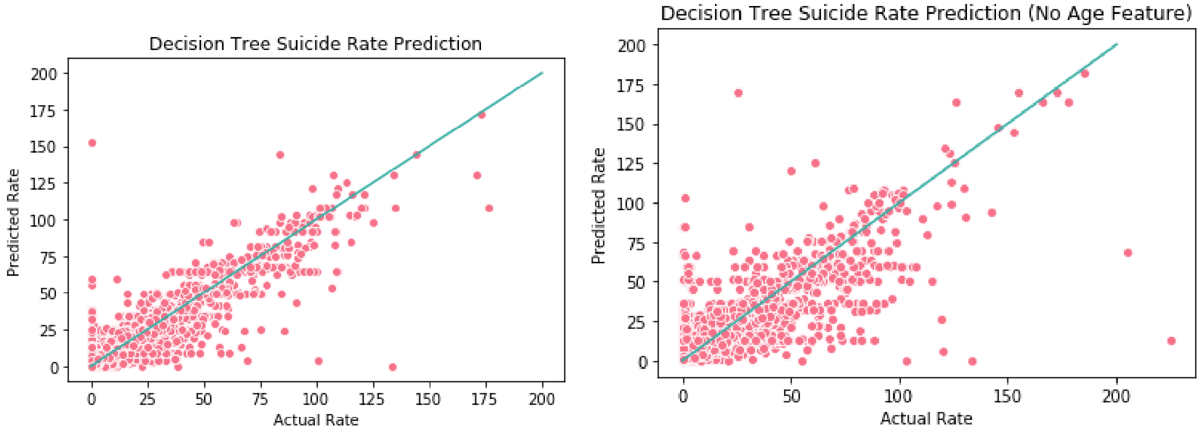
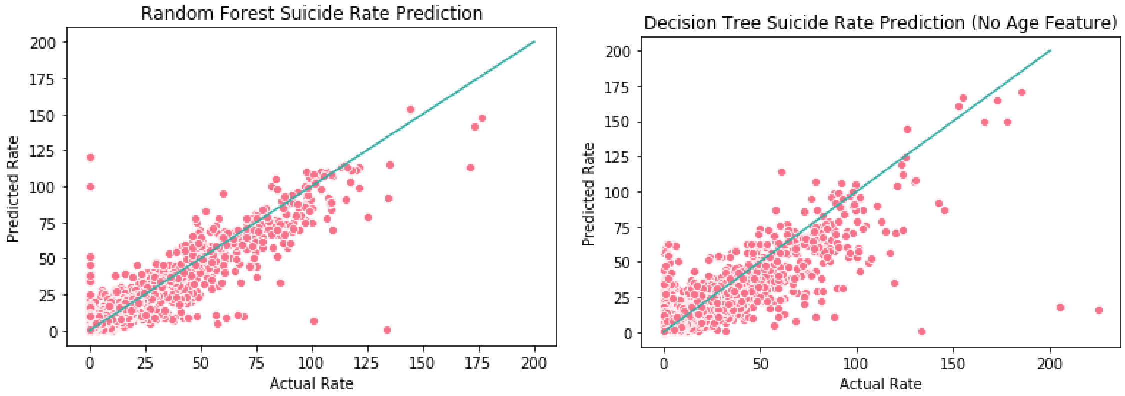


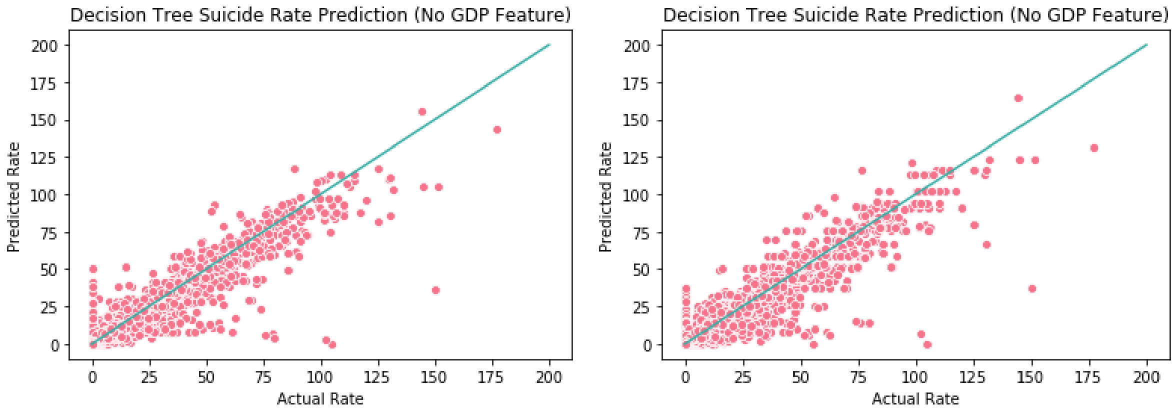
Figure 10: Random Forest: with age on the left, without age on the right



We see that age does have a large impact on our prediction for suicide rate.

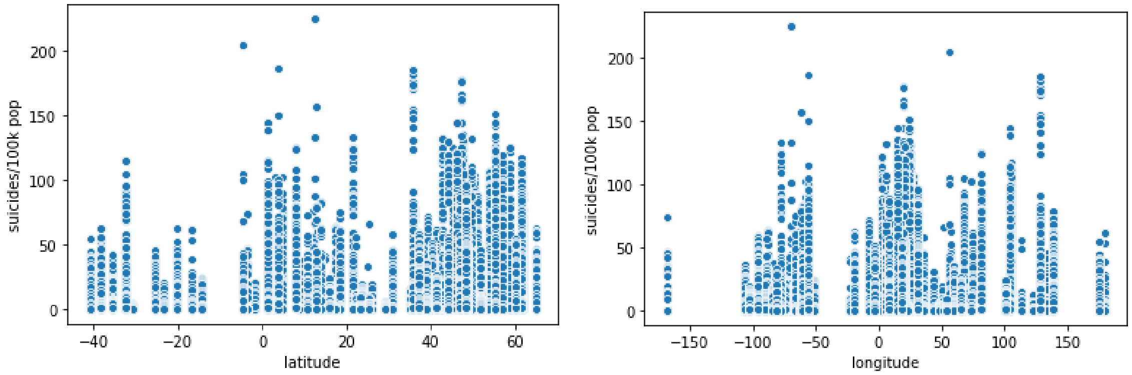
The next factor we looked at is gdp, as it was stated by previous study that finance difficulty would have a large impact on suicide ideation. However, from Figure 11, we notice that gdp's presence in calculation doesn't have much impact.

Figure 11: Decision Tree Prediction on the left, Random Forest Prediction on the right

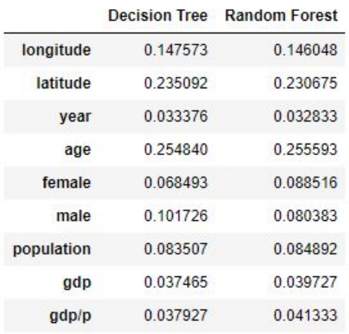


In addition, latitude and longitude were chosen as a representation of country code, and we didn't expect much inference of suicide rate from geometric coordinate. In Figure 12, we plotted the longitude and latitude of our samples against suicide rate. There seems to be some pattern hidden behind the plots, which remains to be researched on.

Figure 12: latitude (left) and longitude (right) vs suicide rate



To remove the large training/ testing bias on these models, we performed 5- fold cross validation, and received feature importance similar to what we've presented in Feature Weights section



	Decision Tree	Random Forest
longitude	0.147573	0.146048
latitude	0.235092	0.230675
year	0.033376	0.032833
age	0.254840	0.255593
female	0.068493	0.088516
male	0.101726	0.080383
population	0.083507	0.084892
gdp	0.037465	0.039727
gdp/p	0.037927	0.041333

IV. Discussion

In this project, we directly analyzed the coefficients of our regression models. However, from Stefan Haufe's paper "On the interpretation of weight vectors of linear models in multivariate neuroimaging", we realized that for linear regression, there are more appropriate interpretation methods for feature importance compared to what we were doing, which might lead to more precise interpretation of results (2014).

From the analysis in the result section, we notice that some of the determining factors identified through feature importance comparison were not validated as significant, e.g. the gdp feature. This contradicts with the previous social science study, and leads to the proposition that this dataset might not be sufficient enough to establish inference from gdp to suicide rate. Other social-economical data would be required to perform further research.

The hidden pattern behind the latitude and longitude graph would also be an interesting direction to research on. There are various possible explanations for the patterns. Factors like weather, humidity, continental geometry might be associated with these geometric coordinates. Overall, Decision Tree and Random Forest Models produced more precise predictions of our data compared to Linear Regression Model. Sex and Age are the overall leading predictors in these models.

V. References

Rusty. (2018, December 01). Suicide Rates Overview 1985 to 2016. Retrieved December 17, 2020, from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

Cohen, E. (2017, September 02). Counties geographic coordinates. Retrieved December 17, 2020, from https://www.kaggle.com/eidanch/counties-geographic-coordinates

Bae, S. (2019). The prediction model of suicidal thoughts in Korean adults using Decision Tree Analysis: A nationwide cross-sectional study. *Plos One, 14*(10).

doi:10.1371/journal.pone.0223220

Suicide. (n.d.). Retrieved December 17, 2020, from https://www.nimh.nih.gov/health/statistics/suicide.shtml

Suicide Rates Overview 1985 to 2016. (2018, December 01). Retrieved December 16, 2020, from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

Suicide Rates Prediction: EDA + Regression! (2020, July 02). Retrieved December 17, 2020, from https://www.kaggle.com/haithembenrhouma/suicide-rates-prediction-eda-regression

Predicting Suicide Rates with Linear Regression. (2019, February 11). Retrieved December 17, 2020, from https://www.kaggle.com/thrashman2324/predicting-suicide-rates-with-linear-regression

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Suicide Rates Machine Learning and Prediction. (2020, December 07). Retrieved December 17, 2020, from https://www.kaggle.com/nilskaizen/suicide-rates-machine-learning-and-prediction

VI. Contributions

Ye is responsible for the introduction, method, and results sections of this report. Yuren is responsible for the results and discussion sections of this report. Specifically, Ye mainly focused on the exploratory data analysis and Linear Regression part of the project while Yuren mainly focused on the data processing, Decision Tree and Random Forest modeling of the project.