London Fire Risk Analysis

1. Background and Introduction

The London Fire Brigade (LFB) has contracted Spot-on Safety Analytics (SSA) to provide Data Science services in modernizing its fire readiness protocols. This organization services a dense metropolitan area with the highest rate of emergency requests in the country and is one of the largest firefighting and rescue service in the world. SSA's team has been task with the development predictive capabilities to target high risk population centers, and London Ministry officials desire that the SSA prototype solution can serve as a foundational solution which can be expanded for other city services as part of a long-term Analytics modernization strategy.

The LFB has provided the SSA team with their historical incident data (2009-2017). Additional demographic data can be further blended to expand the precision of the Fire Risk model coupled with providing both the Ministry and Fire officials more detailed dimensions into high risk area by population characteristics. The UK Office of National Statistics derives annual national population estimates by key demographic characteristics. SSA resources will have to blend London population estimates by neighborhood/borough information with LFB internal data to enhance the provided base data structure to support this level of expanded BI functionality.

In this mock case, our group take the role of SSA and are expected to develop a valid quantitative model with satisfactory accuracy in prediction on fire, with the aid of historical incident data and other data which may have impacts on fire – specifically, weather data (temperature and precipitation) for each day in London, demographic characteristics (total population, area, population age, median pay) for each borough in London, and average building age for each borough in London; and all of the data are all from UK authorities as required.

2. Data Preparation

2.1 Data Preparation for Separate Data Tables

As is mentioned in introduction, we basically have four data tables in hand initially:

a. fire incident data
b. weather data
c. demographics
d. property building age

After checking the missing value in each table, we find missing values only occur in few days in weather data table; to impute these more accurately, we manually make them be equal to the average of corresponding same days in other years. For example, we found the weather data for 11/11/2016 was missing, and thus we took the average of weather figures in 11/11 of other years and utilized the mean value for the missing one in 2016.

For incident dataset, which does not have missing values in the columns we need, it contains various incidents, e.g., false alarm, special service, and fire. So firstly, we select out all the records which are from actual fire, and then only keep the date, borough code columns for further study. Following we provide a screenshot of original incident dataset for reference (not all columns are included), and it is shown that from the dataset we can basically know where (borough), when (date), and which kind of incident was happened.

DateOfCall	CalYear	TimeOfCall	HourOfCall	IncidentGroup	StopCodeDescription	SpecialServiceType	PropertyCategory
01/01/2009	2009	0:00:37	0	Special Service	Special Service	RTC	Road Vehicle
01/01/2009	2009	0:00:46	0	Special Service	Special Service	Assist other agencies	Outdoor
01/01/2009	2009	0:03:00	0	Fire	Secondary Fire		Outdoor
01/01/2009	2009	0:04:27	0	Fire	Secondary Fire		Outdoor
01/01/2009	2009	0:05:39	0	Fire	Secondary Fire		Outdoor
01/01/2009	2009	0:06:03	0	False Alarm	AFA		Dwelling
01/01/2009	2009	0:12:31	0	Special Service	Special Service	RTC	Road Vehicle
01/01/2009	2009	0:13:42	0	Fire	Secondary Fire		Outdoor Structure

With the incident table, we can easily get the fire number in a borough in one day, by using SQL query saying that count the rows when group by date and borough code, and then we get a table, the first several rows are shown below (Table 1):

	Date	Borough_Code	Fire_Number
1	01/01/2009	E09000002	8
2	01/01/2009	E09000003	1
3	01/01/2009	E09000004	2
4	01/01/2009	E09000005	4
5	01/01/2009	E09000006	3
6	01/01/2009	E09000007	1
7	01/01/2009	E09000008	6

Table 1

	Date	Borough_Code
1	01/01/2009	E09000001
2	01/01/2009	E09000002
3	01/01/2009	E09000003
4	01/01/2009	E09000004
5	01/01/2009	E09000005
6	01/01/2009	E09000006
7	01/01/2009	E09000007

Table 2

However, since Table 1 is from incident table, which only records incidents that have happened but does not directly tell when and where there is no fire, i.e., there is no record for fire number equal to zero in Table 1; under the circumstance, we manually create a blank record table (see Table 2) with all data & borough combinations, and as is highlighted, there are rows which will never be matched into Table 1 by date and borough code since under the particular date & borough combination, there is no fire recorded. For example, in borough E0900001 on 01/01/2009, there was no fire, and that is why this data & borough combination is not shown in the first row of Table 1.

After getting blank record table with all of the date & borough combinations (Table 2), we make it left join Table 1 on date and borough code, and get a table shown below (Table 3); now we clearly see that since there is no fire record for borough E0900001 on 01/01/2009 in Table 1, we get null value in Table 3; and there is also no fire record for some other date & borough combinations, which can be reflected by "NA" in Fire_Number column of those rows. As a result, we only need to change NA to zero for those rows to indicate that there was no fire for that borough on that day, see Table 4.

	Date	Borough_Code	Fire_Number
1	01/01/2009	E09000001	*NA*
2	01/01/2009	E09000002	8
3	01/01/2009	E09000003	1
4	01/01/2009	E09000004	2
5	01/01/2009	E09000005	4
6	01/01/2009	E09000006	3
7	01/01/2009	E09000007	1

Table 3

	Date	Borough_Code	Fire_Number
1	01/01/2009	E09000001	0
2	01/01/2009	E09000002	8
3	01/01/2009	E09000003	1
4	01/01/2009	E09000004	2
5	01/01/2009	E09000005	4
6	01/01/2009	E09000006	3
7	01/01/2009	E09000007	1

Table 4

So far, Table 4 we get can then tell both no-fire "records" and fire records, which is crucial to our quantitative model to predict whether a fire happens or not.

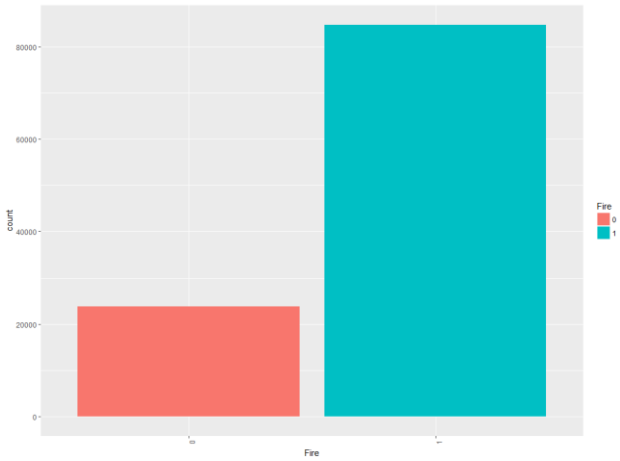
Keeping in mind that we need a merged dataset for modelling to predict the risk of fire in each borough for a given day, we start to make incident data (Table 4) join other tables: join weather data on date, join demographics on borough code, and join property age data on borough code. Eventually, we get the following table (Table 5) to build models, with a dummy variable column "Fire" (Yes = 1, No = 0, based on the Fire_Number column).

	Date	Borough_Code	Fire_Number	TAVG	PRCP	Total_Population	Inland_Area	Population_Density	Population_Age	Median_Pay	Building_Age	Fire
1	01/01/2009	E09000001	0	1.9	0	8800	290.3934	30.30372	43.2	56288	57.08133	0
2	01/01/2009	E09000002	8	1.9	0	209000	3610.7817	57.88220	32.9	28439	71.96974	1
3	01/01/2009	E09000003	1	1.9	0	389600	8674.8314	44.91154	37.3	32119	72.92509	1
4	01/01/2009	E09000004	2	1.9	0	244300	6058.0668	40.32640	39.0	30611	69.29530	1
5	01/01/2009	E09000005	4	1.9	0	332100	4323.2637	76.81697	35.6	30133	79.21127	1
6	01/01/2009	E09000006	3	1.9	0	327900	15013.4892	21.84036	40.2	29848	69.55453	1
7	01/01/2009	E09000007	1	1.9	0	242500	2178.9295	111.29318	36.4	38147	84.19400	1

Table 5

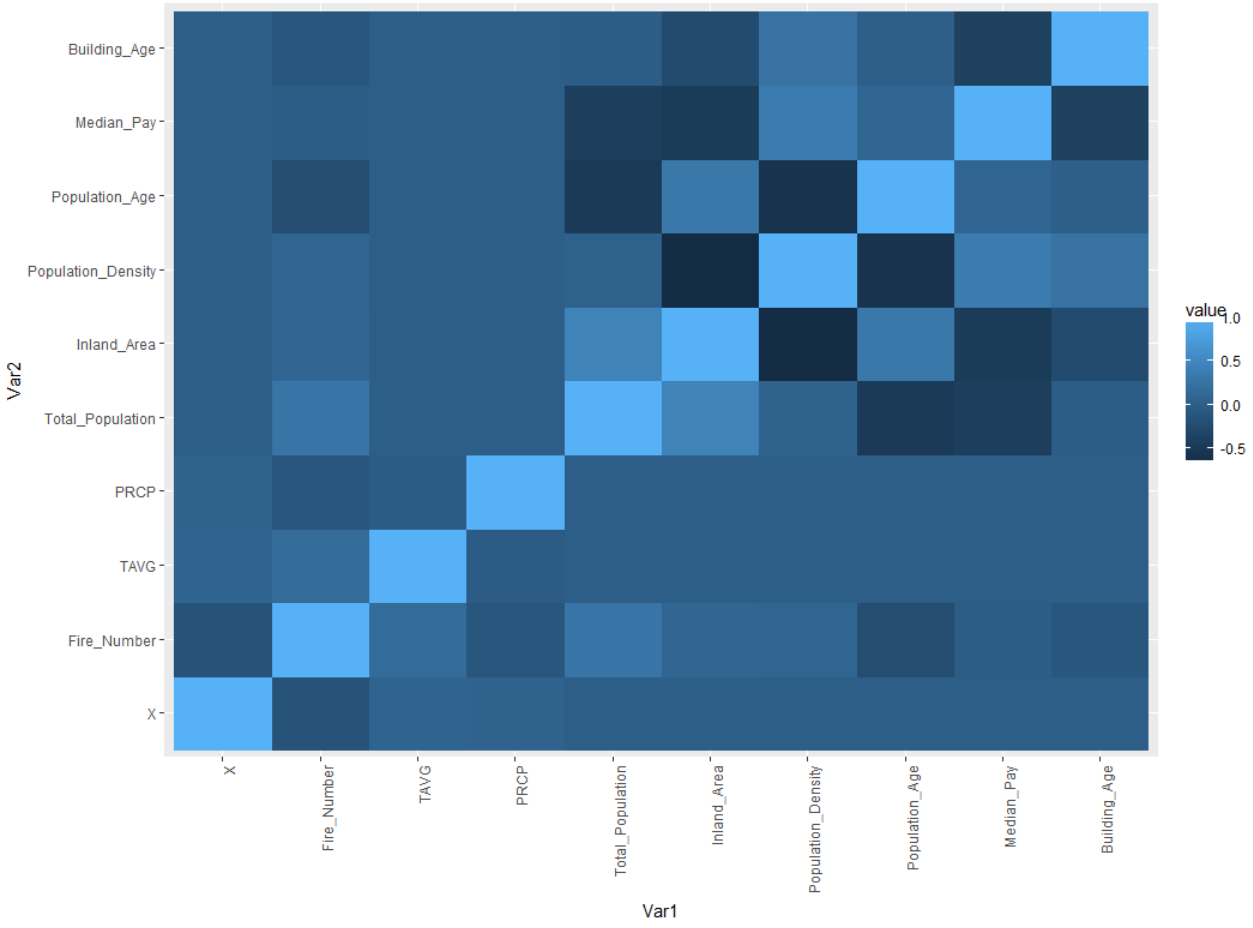
2.2 Data Preparation for Model Building

As the graph of the target variable Fire (binomial) showing below, we could find that this dataset for modeling has the Imbalanced Target Variable issue, to deal with that, over/undersampling is required. Using "ROSE" package in R to create a more balanced target distribution, we choose under-sample since it is a large dataset for R to deal with.



The correlation heat map showing below indicates that for the independent variables, both population_age and population_density, Inland_area and population_density have strong multicollinearity issue since population_density = total population/area, which might cause unstable prediction for the model we build. Thus, we will drop population_density variable in further analysis.

Correlation Heat Map



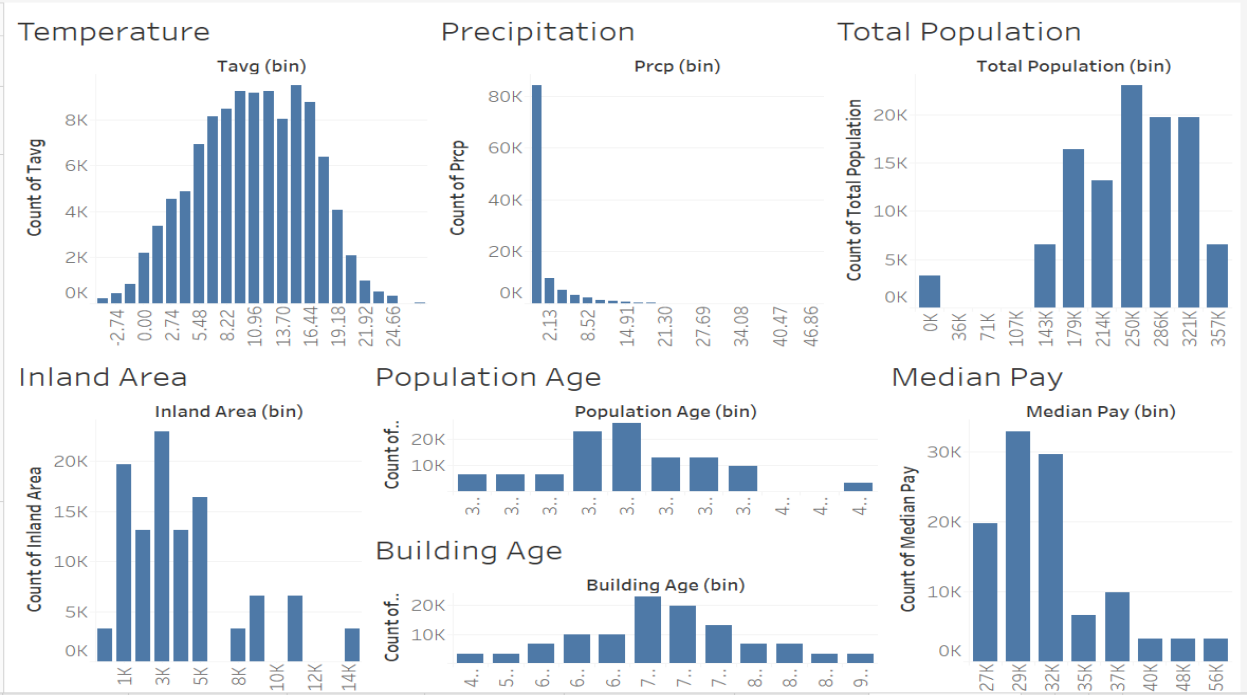
3. Model Building

After all the data preparation process above, we list all the variables that will be included in model building for predictive analysis.

Variables for Model Building

Variable	With Regard to	Role	Description
Fire	Date & Borough	Dependent Variable	1 for Yes and 0 for No.
Temperature	Date	Independent Variable	Average London daily temperature.
Precipitation	Date	Independent Variable	London daily precipitation.
Total Population	Borough	Independent Variable	Total population for each borough.
Inland Area	Borough	Independent Variable	Inland area for each borough.
Population Age	Borough	Independent Variable	Average people's age in each borough.
Median Pay	Borough	Independent Variable	Median pay for each borough.
Building Age	Borough	Independent Variable	Average building age for each borough.

Distributions for Independent Variables



We mainly use H2O package in R to construct GLM (binary classification), random forest and neural network model to see which model can do a better job.

We split the whole dataset into train(70%), valid(15%), and test(15%) to make our model more stable and accurate, prevent it from overfitting and give us an honest feedback of the model.

We use AUC (a measure of sensitivity in relation to specificity: at low levels of sensitivity we want high levels of specificity (and vice versa)) as our main measure to do model selection.

3.1 Logistic Regression

Using GLM in H2O to perform binary classification to conduct logistic regression. We did 3 GLM in this part: 1. Logistic regression without standardization; 2. Logistic regression with lambda search; 3. Logistic regression with standardization. Please see details in R codes.

Standardization seems quite important since it gives us the best result in GLM, which is also necessary to compare the relative contribution of different predictors to the model, we will analyze the details of standardized coefficient of GLM model in next chapter.

Best AUC for Logistic Regression: 0.7075298 (3. Logistic regression with standardization)

3.2 Random Forest

For the Decision Tree model, first we train a basic Random Forest model with default parameters with 'ntrees = 50'; next we increase the number of trees used in the forest by setting 'ntrees = 100', using test dataset to retrieve AUC, we get results around 0.715; increase ntrees to 200, it seems the result does not change much. Please see details in R codes.

Best AUC for Random Forest: 0.7285115 (2. 'ntrees = 100')

3.3 GBM grid search

We use Grid (Hyperparameter) Search in H2O and train a standard supervised prediction model for a better result.

We conduct a GBM grid search with parameters showing below:

learn_rate = c(0.01, 0.1)

max_depth = c(3, 5, 9)

sample_rate = c(0.8, 1.0)

col_sample_rate = c(0.2, 0.5, 1.0)

Here is the search result:

#	col_sample_rate	learn_rate	max_depth	sample_rate	model_ids	AUC
1	0.2	0.1	5	1.0	gbm_grid1_model_27	0.72666
2	0.5	0.1	3	1.0	gbm_grid1_model_22	0.72664

Notice the AUC result above is the AUC for validation dataset ('valid'). Grab the top 2 GBM model, chosen by validation AUC, and use test dataset ('test') to give an honest feedback for the deep learning model.

```
best_gbm1 <- h2o.getModel(gbm_gridperf1@model_ids[[1]])
best_gbm_perf1 <- h2o.performance(model = best_gbm1, newdata = test)
h2o.auc(best_gbm_perf1)
```
The test AUC for the top1 model: 0.7359763.

```
best_gbm2 <- h2o.getModel(gbm_gridperf1@model_ids[[2]])
best_gbm_perf2 <- h2o.performance(model = best_gbm2, newdata = test)
h2o.auc(best_gbm_perf2)
```
The test AUC for the top2 model: 0.7356292.

Best AUC for GBM grid search: 0.7359763.

Model Summary for the best GBM model:
number_of_trees: 100
number_of_internal_trees: 100
model_size_in_bytes: 25304
min_depth: 0
max_depth: 5
mean_depth: 3.44
min_leaves: 1
max_leaves: 31
mean_leaves: 15.06.

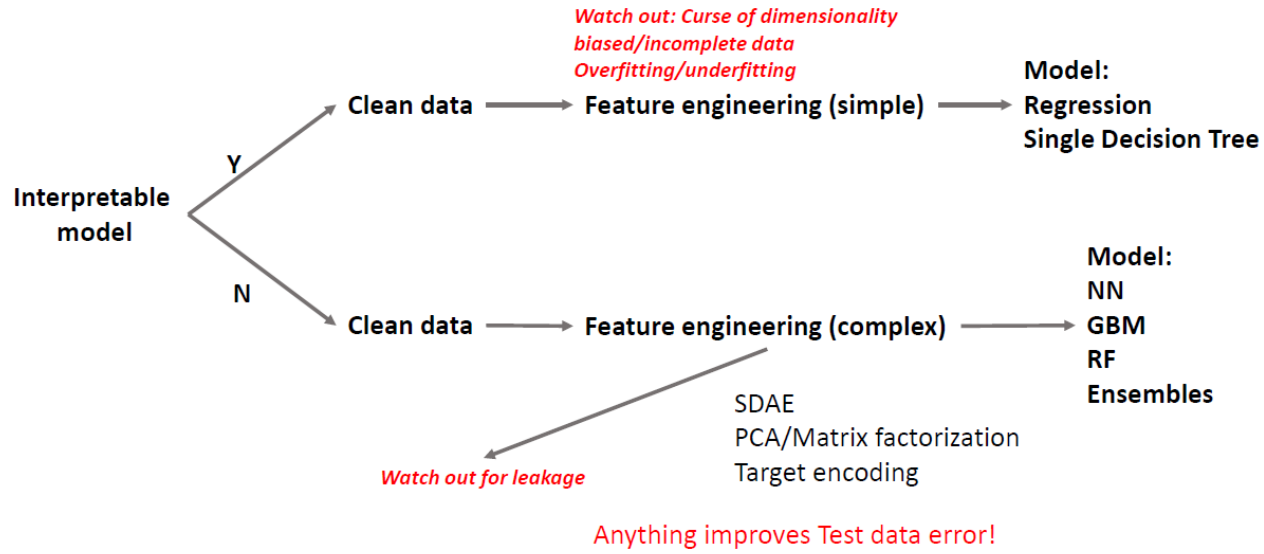
4. Results and Conclusion

4.1 Model Results

Here is the best result we obtained in model building part:

a. Best AUC for Logistic Regression: 0.7075298;
b. Best AUC for Random Forest model: 0.7285115;
c. Best AUC for GBM grid search model: 0.7359763.

According to the graph showing below, it is reasonable for us to interpret our model by logistic regression or single decision tree and do predictive analysis by random forest and GBM grid search.



Though Random Forest model and GBM grid search model have better predictability, to interpret factors which have big effect on Fire or give business suggestions to LFB, we should use results from interpretable model to explain what happened in LFB Fire dataset.

Thus, we also do a single decision tree model in SAS JMP, compare it with GLM model result (standardized coefficient) to determine the most important variables (factors to which LFB should pay attention to improve its Fire Warning System).

Here is the variable selection result:

Single Decision Model in SAS JMP

Column Contributions

Term	Number of Splits	SS	Portion
Total_Population	3	1120.39941	0.7256
Building_Age	2	154.578262	0.1001
PRCP	1	119.479583	0.0774
TAVG	2	92.5125447	0.0599
Median_Pay	1	57.1650573	0.0370
Inland_Area	0	0	0.0000
Population_Density	0	0	0.0000
Population_Age	0	0	0.0000

Logistic Regression Model in R (H2O)

```
Standardized Coefficient Magnitudes: standardized coefficient magnitudes
             names coefficients sign
1 Total_Population     0.534461  POS
2   Population_Age     0.342168  NEG
3             TAVG     0.239368  POS
4       Median_Pay     0.179502  POS
5      Inland_Area     0.178745  POS
6             PRCP     0.160851  NEG
7     Building_Age     0.016024  POS
```

Compare these 2 variable contribution results, we could find that 'Total_Population', 'TAVG', 'Median Pay' play a leading role in both 2 models, 'Population_Age' is not statistically significant in logistic regression model and has little contribution in single decision tree model, 'PRCP' and 'Building_Age' seem a little bit tricky —— it plays an important role in one of these 2 models but not both of them. We still take them into account to give interpret to the following factor analysis.

Influential Factors for Fire Risk Analysis

Important Variables	With Regard to	Role	Description
Temperature	Date	Independent Variable	Average London daily temperature.
Precipitation	Date	Independent Variable	London daily precipitation.
Total Population	Borough	Independent Variable	Total population for each borough.
Building Age	Borough	Independent Variable	Average building age for each borough.
Median Pay	Borough	Independent Variable	Median pay for each borough.

4.2 Business Suggestions

4.2.1 Weather-related Variable: “TAVG” (Temperature) & “PRCP” (Precipitation)

Monitor real-time weather condition, especially for temperature and precipitation. It is the common sense that the more the precipitation, the less the possibility for a Fire, and we estimate that this possibility increases as the temperature increases. For example, when extreme weather events happen, such as continuous high temperature and abnormal drought, it is likely some small fuse might be the trigger of a huge disaster.

4.2.2 Population-related Variable: “Total Population” & “Median Pay”

Take population density into account besides area/distance to optimize the arrangement of fire equipment/facility for boroughs with high population density to meet the high-variability demand of firefighting, thus, improve its fire readiness. Boroughs having larger population will not only increase the possibility to cause fire, but also render it harder to fight the fire and save lives/property, resulting in higher loss.

It is strange the median pay has a positive relationship with the possibility of fire. Normally, we will consider that boroughs with higher median pay will have better infrastructure, resulting in less possibility to cause fire. It might need further study to find out the causal relationship between them or this might be a just spurious relationship. Since there are no collinearity issues or obvious interactions in our dataset, further analysis based on domain knowledge for their relationship is required.

4.2.3 Facility-related Variable: “Building_Age”

Set a safety period for the building (most British building have 132-year average service life limit, fire possibility increases with the increase of building age, thus a safety period, for example 70 years, can be set for LFB to give alarm/monitor to those exceed such safety-period limit but not reach a certain level service life limit.).

Rating the building with an extremely long service life, for example, longer than 130 years, urge the government to discard those buildings exceed its service life as worthless. We could find in our case, more than 50% buildings’ ages exceed 70 years which can easily lead to a fire disaster.

4.3 Ideas for Further Analysis

The 4 parts above is our main content for London Fire Risk Analysis, it seems that there is still room for improvement, however, it is also hard for us to do them in a short term. We list these directions below for reference:

a. Combine not only common sense but also domain knowledge to detailed analyze the causal relationship between Fire and all the independent variables;
b. Find data about casualties and property loss to define the level of the fire disaster;
c. Based on the data in b, find the outliers in fires with extremely high casualties and property loss and do further analysis to see why these catastrophic disasters happened;
d. Based on domain knowledge and relevant papers about researches on fire disasters, add more variables to develop a more accurate fire prediction model.