PREDICTING FIRE RISK AND PRIORITIZING FIRE INSPECTIONS IN FAIRFAX COUNTY

ABSTRACT

To help Fairfax County address the fire incidents and help in reducing the occurrence of fire hazards we have come up with a predictive model to evaluate property-level fire risk and prioritize fire inspections. We used historical fire incidents and inspections data from Fairfax County as the base data for this project. In addition to that we used Property datasets from Fairfax County GIS Open data source which provided abundant information about the different properties in Fairfax County. The most challenging part was to join the different datasets. Once the datasets were joined, four classification models namely Logistic Regression, Decision Tree, Random Forest and Stochastic Gradient Boosting were built, and the model performance were compared.

Keywords: Fairfax County, Fire Risk, Logistic Regression, Decision Tree, Random Forest, Stochastic Gradient Boosting.

1. INTRODUCTION

There are several fire incidents occurring across the world. In the United States alone, fire departments are estimated to attend 1.4 million fires in 2020. These fires killed 3,500 people and wounded 15,200 more. Furthermore, property loss was estimated to exceed $21.9 billion[1]. The authorities, who examine houses on a regular basis, are attempting to enforce several fire laws in order to decrease the harm caused by fire. However, authorities follow a fixed set of rules based on pre-existing licenses, or, at best, a rule-based heuristic for evaluating eligibility, selecting and inspecting properties for inspection. [2]Nonetheless, this technique is outdated and ineffective since, as the population grows, so does the need for housing, and with such a large number of properties, the designated department cannot inspect them all on a yearly basis. With technological breakthroughs in the fields of machine learning and data, it is now possible to analyze the risk of each property, allowing authorities to adapt this method and conduct inspections on a regular basis. These existing fire inspection procedures could be considerably improved by using risk-based and data-driven processes for identifying, selecting, and prioritizing new properties to inspect. As a result, we anticipate that the suggested approach will aid in assessing the likelihood of fire risk in property

[1] "Fire loss in the United States during 2020 - NFPA." [Online]. Available: https://www.nfpa.org/~/media/FD0144A044C84FC5BAF90C05C04890B7.ashx. [Accessed: 10-May-2022].

[2] Bhavkaran Singh Walia Carnegie Mellon University, B. S. Walia, C. M. University, Q. H. C. M. University, Q. Hu, J. C. C. M. University, J. Chen, F. C. C. M. University, F. Chen, J. L. C. M. University, J. Lee, N. K. C. M. University, N. Kuo, P. N. C. M. University, P. Narang, Jason Batts Pittsburgh Bureau of Fire, J. Batts, P. B. of Fire, Geoffrey Arnold Dept. of Innovation and Performance, G. Arnold, D. of I. and Performance, M. M. C. M. University, M. Madaio, I. C. London, Ibm, and O. M. V. A. Metrics, "A dynamic pipeline for spatio-temporal fire risk prediction: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining," ACM Other conferences, 01-Jul-2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3219819.3219913. [Accessed: 10-May-2022].

structures. To address this gap in fire inspection prioritizing, we created a predictive model to evaluate property-level fire risk (i.e. the possibility of a certain property having a fire event).

2. DATA SOURCES

The Fairfax County Fire and Rescue Department (FCFRD)[3] is a combination of career and volunteer all-hazards department providing a full complement of emergency medical services, fire suppression, technical rescue, swift water rescue, hazardous material response, arson investigation and fire and life safety education. To analyze the fire risk properties for Fairfax County we use data from multiple sources as tabulated in the table.

Data	Source	Features	Records	Date Range
Fire Incidents	Fairfax County fire-ems	81	13242	2016 to 2021
Fire Inspections	Fairfax County fire-ems	24	444865	2001 to 2021
Address Points	Fairfax County GIS	29	374214	2022
Tax – Land Data	Fairfax County GIS	8	368663	2022
Tax- Dwelling Information	Fairfax County GIS	27	336486	2022

Table 1: Data Sources

2.1 FIRE INCIDENTS:

Fairfax county provided us with a fire inspections dataset which included historical fire incidents from 2016 to 2021.The dataset has approximately 13000 fire incidents reported in the area. This includes information about fire incidents, such as incident type, Caller Source, date and time of the incident, location address, geographical coordinates, respective jurisdiction and cause of fire incident. There were 81 features in the dataset, after careful analysis and consideration the number of features were reduced to 20.

2.2 FIRE INSPECTIONS:

The Fire prevention and Inspection section consists of multiple units that are geographically subdivided. They inspect buildings, institutions, and occupancies to ensure compliance with the Virginia SFPC, Fairfax County Fire Prevention Code, and the Virginia Petroleum Storage Tank regulations to maintain a safe environment for occupants. The fire inspections data has the details of the fire inspections that happened in the Fairfax County area from 2001 to 2021.We have approximately 444K reported fire inspections happened over the years. We can see an increase in the number of inspections over the years. There were 24 features in the dataset. The status of the inspection is recorded as either Pass or Fail.

2.3 GEOSPATIAL INFORMATION SYSTEMS (GIS) DATASETS:

Fairfax County GIS & Mapping Services provides access to many free GIS datasets. The below datasets from GIS were found useful for our project. These datasets provide structural level information about the

[3] Fairfax County GIS Open Datasets; ‘https://www.fairfaxcounty.gov/maps/open-geospatial

properties in Fairfax County. They also help to identify which year the property was built, remodeled, the area occupied and type of property such as Residential, Commercial, Agricultural etc.

2.3.1 Address Points

This data contains the point features representing the address points, with attributes that include full address and parcel ID number for Fairfax County. This dataset is used for Joining the Fairfax country Fire Inspection/Incidents dataset with other GIS and Property related dataset. The Address Points dataset joins the Fire Inspection/Incident dataset by X and Y Latitudes and Includes Parcel Identifier (PIN) for those available addresses. The parcel pin can be used to join with other Property datasets

2.3.2 Tax Administration's Real Estate - Dwelling Data

The Tax Dwelling data dataset contains information about the external structure on the parcel (building identifier) including year built, number of bedrooms and bathrooms and style for properties within Fairfax County. There is a one-to-many relationship to the parcel dataset. Out of the 22 features in the dataset, we are considering 16 features in the model which are described in the table below. The Tax Administration dwelling dataset is joined to Address Points dataset through Parcel Identifier (PIN).

2.3.3 Tax Administration’s Real Estate - Land Data

Tax Administration Real Estate Land Data contains the information about the land including land sizes (square feet & acres) and land property type for properties within Fairfax County. There is a one-to-many relationship to the parcel data. There were 7 features in the dataset, The code_desc is an important attribute which identifies what type of property it is – Commercial, Residential, Agricultural land etc. The Tax Administration Land dataset is joined to the Address Points dataset through Parcel Identifier (PIN).

3. DATA PREPROCESSING

3.1 Data Cleaning

After collecting the datasets, a significant amount of data cleaning was required. The first step towards data cleaning was to check for the missing values and decide how to deal with them. We have decided to impute median values to minimize the deletion of records. However, the features containing more than 25% of NA (missing) values were deleted from the data. Box-cox transformation, centering and scaling were performed on the numeric features. For the categorical features, trimming was performed to remove blank spaces. We have computed near zero variance and eliminated four features. For each dataset, we manually studied each feature to check the relevancy of the feature to the prediction. All the features that we considered irrelevant to the prediction were eliminated.

3.2 Data Filtering

The Fire incidents dataset had a total of 13,241 records. Filtering the data was necessary as the data consisted of incidents which were not related to fire. The dataset had a feature “NFIRSTypeCode” which represented the type of incident. Through these codes we were able to identify the codes which were related to fire and filtered out the other codes from the data. After the data filtration, there were 12,408 incidents

retained in the dataset. The NFIRS Type code groups and the incidents they represent can be seen in the figure 1 below.

The fire inspections dataset had a feature Application Description which contained the value "TRANSPORT VEHICLE FOR BLASTING". This inspection application is irrelevant for this project of fire risk prediction and was deleted from the data. Few other records were removed from the dataset as those types contained very less data to support the prediction.[4]

Fig. 1 NFIRS Incident Type Groups

3.3 Joining Datasets

This step is crucial for the project since the prediction of fire risk of a property will be accurate when all the aspects of the property are fed to the model. For example, the model will be able to predict the fire risk based on the details of the property such as age of the property, tax details, and structure details. Furthermore, joining fire incidents and fire inspections datasets provides insight into the status of the inspections when the fire incidents occurred at the property.

We joined datasets primarily on spatial location information. The datasets contained two main factors: location information and Parcel Identification Number. Parcel ID is a unique ID given to each property for tax purposes by the Virginia County government. The tax-land, tax-dwelling and address points datasets were joined using Parcel ID. To join the fire inspections and fire incidents datasets, the location coordinates were to be used. However, the location coordinates of fire inspections were in WGS84 coordinate system format whereas that of fire incidents were in Virginia State Plane Coordinate System (North Zone). The difference in coordinates of these two coordinate systems is illustrated in the table below.

[4] NFIRS Incident Types, Actions Taken & Property Use Codes, https://www.responserack.com/nfirs/common/

WGS84 Coordinate System		Virginia State Plane Coordinate System	
Lattitude	Longitude	X- coordinate	Y-coordinate
-77.190176	38.82572	11797289.31	7033359.24

Table 1 Illustration of the difference in coordinates of the two coordinate systems

We have identified the “rgdal” library which is a Geospatial Data Abstraction Library to convert the location information in Virginia State Plane Coordinate System to WGS84 coordinate system using the spTransform() method. A snippet of the R code used to convert the location information is shown in figure 2.

```
data1= data.frame(x=finc_clean$xCoordinate,y=finc_clean$yCoordinate)
coordinates(data1) <- ~ x + y
proj4string(data1) <- CRS("+init=epsg:2283")
finc_clean[,14:15] = data.frame(spTransform(data1, CRS("+init=epsg:4326")))
head(finc_clean)
```

Fig. 2 Snippet of location coordinates conversion R code

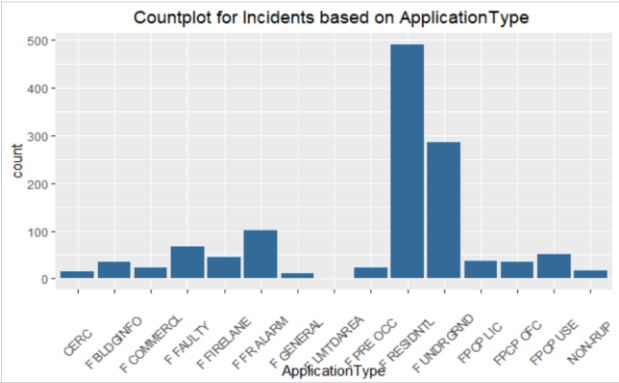
After converting the coordinates, we then reduced the coordinates of both the datasets to four decimal places to match the location in the datasets. A new feature was created in fire incidents datasets and “1” value was populated into the feature. We then joined the two datasets and populated “0” into the Incident Happened feature where there were NA (missing) values. This feature will be the response variable used in the models for fire risk prediction. We then merged all the datasets together using location coordinates. A feature in the merged dataset “EXTWALL_DESC” which contains the material of the external wall of the property was transformed into a binary variable which indicates if the external wall was built with wood or other material. The final dataset contained 1,231 observations and 20 features.

3.4 Data Splitting and Upsampling

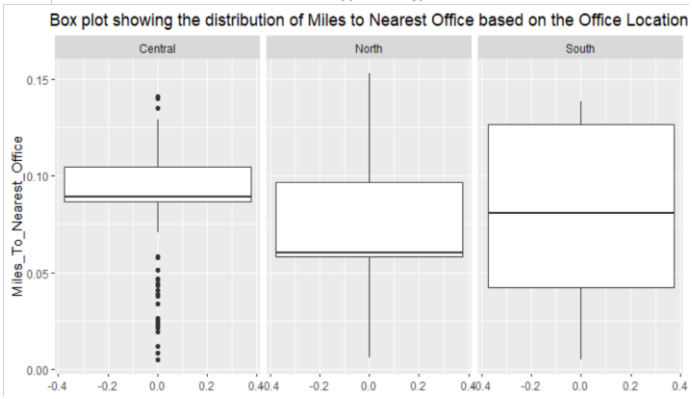
After finalizing the dataset, the data had to be split into training and test datasets. We have decided to split the data into 75% training and 25% test datasets so that we have enough test data to validate the model. Once the dataset was split, the train dataset had 924 observations and the test dataset had 307 observations. Since the training and test datasets are imbalanced, we have decided to up sample the train dataset. The upSample() method of Caret library was used to increase the samples in the data which helped in balancing the response variable data so that the model is not biased. After applying up sampling, the samples in the training data increased to 1758.

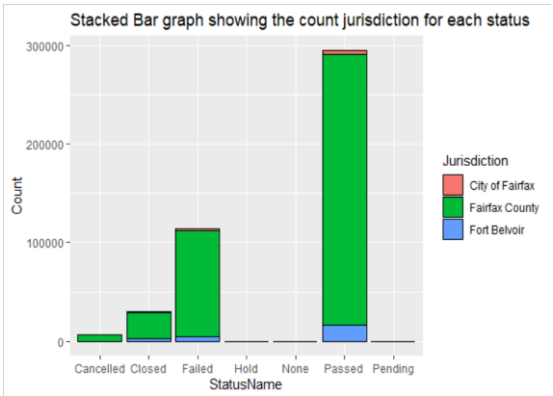
4. EXPLORATORY DATA ANALYSIS

The datasets from Fairfax County Fire and Safety Department, Fire Inspections, Fire Incidents were thoroughly examined. As performing exploratory data analysis is a critical process of examining data to discover patterns, spot anomalies and understand the correlation between different categories we extended our analysis to the secondary datasets supporting the main datasets. The parcels, tax data and buildings data were taken into consideration to perform visualizations.



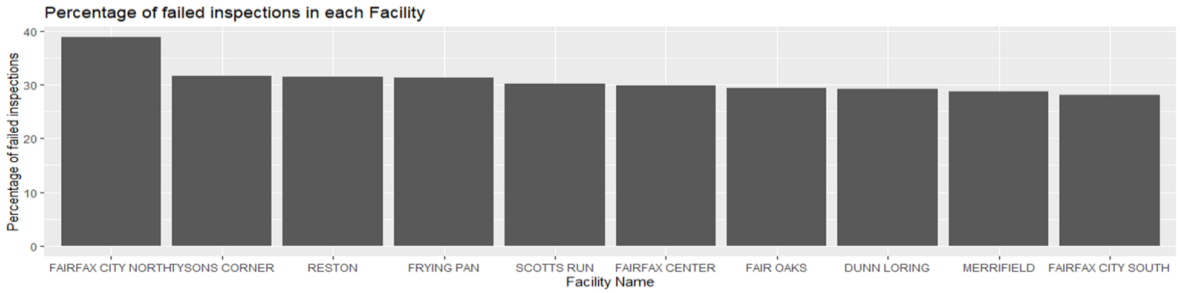
Univariate analysis was performed on the variables in the dataset. The bar graph was plotted to understand the Application Type count on the Fire Inspections dataset. As we can see from the plot above, the F_RESIDENTIAL application type has the highest number of counts. It is the Fire Prevention Code Permit which allows the permit applicant to handle, store or use substances or devices regulated by the fire department.



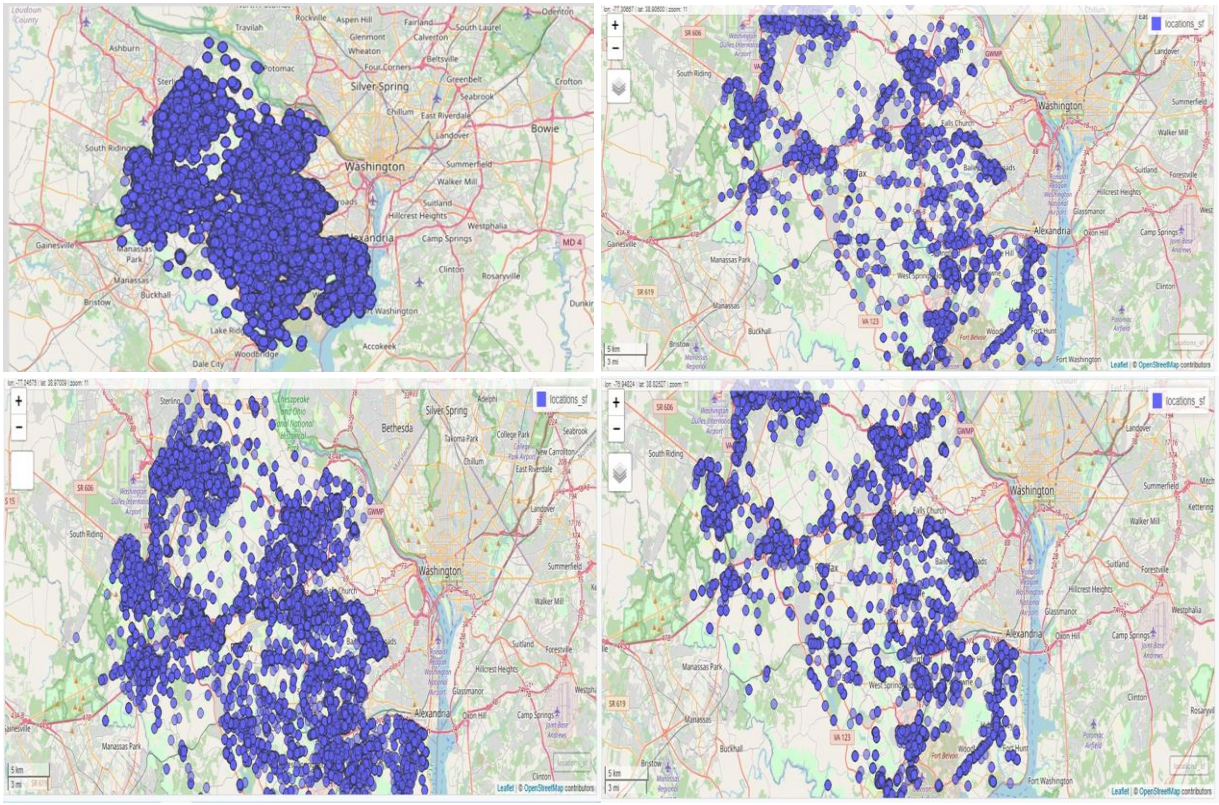


The boxplot shows distribution and skewness of the numerical data by displaying it as percentiles and averages. This plot shows the distribution of distance in miles to the nearest office based on the Office category. The south has the widest distribution whereas in Central you can see many outliers.

The stacked bar graph shows the count of Jurisdiction based on the status of the fire inspection. By the plot above we can see that Fairfax county has the maximum number of passed inspections and failed inspections. Then we began to analyze the percentage of failed inspections in each facility. It is observed that Fairfax city North has the maximum percentage of failed inspections.



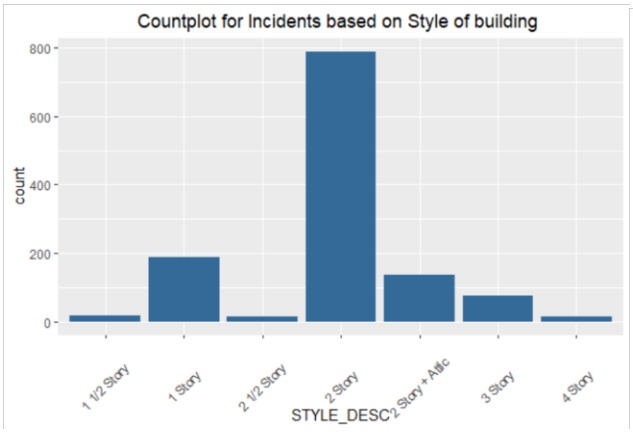
Geospatial Analysis of the dataset

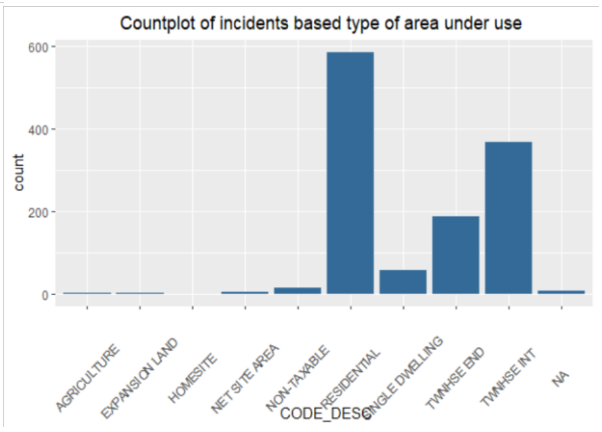


(From Top Left to Right) **Map1:** Locations where all inspections were carried out. **Map 2:** Locations where inspections have failed for at least one time. **Map 3:** Locations where inspections have been carried out more than 5 times. **Map 4:** Locations where inspections have failed more than 5 times.

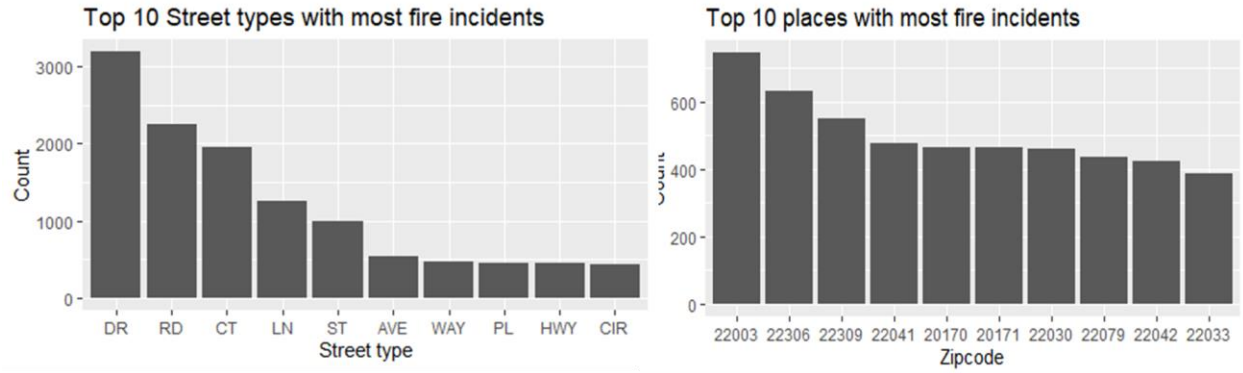
Based on Building types

When we mapped the building data and the fire incidents dataset with respect to their location, we wanted to inspect the features which had suffered more fire incidents. 2 Story buildings experienced most fire incidents over the years and Residential areas are more prone to fire risk when compared to other areas.





The top 10 street types and zip codes were calculated analyzing the fire incidents to understand the more risk prone locations.



5. PREDICTIVE MODELING

5.1. LOGISTIC REGRESSION

Logistic Regression is a supervised machine learning algorithm which is used for binary classification problems and it is used when the target variable is of Categorical data type. Logistic regression is a predictive analysis therefore we used this regression method to predict the risk of fire in a particular area of Fairfax county. The logistic regression model takes the natural logarithm of the odds as a regression function of the predictors with 1 predictor, X, which takes the form:

Since it is less difficult to apply, analyze, and train logistic regression. It can classify unknown data quite rapidly. When datasets are divided linearly, logistic regression works well. Model coefficients might be seen as indications of feature importance. This is why we made our initial option towards logistic regression when determining which algorithms to create.

$$\mathbf{ln[odds(Y=1)]=\beta 0+\beta 1X}$$

where,

ln stands for the natural logarithm,

Y is the outcome : Y= 1 when the event happens (versus Y=0 when it does not),

β0 is the intercept term,

β1 represents the regression coefficient, t

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Fig. Confusion Matrix, 0 Indicates that fire has not happened and 1 indicates that fire has happened

	Reference	
Prediction	0	1
0	249	44
1	7	8

The change in the logarithm of the odds of the event with a 1-unit change in the predictor X. The difference in the logarithms of 2 values is equal to the logarithm of the ratio of the 2 values, so by taking the exponential of β1, we obtain the ratio of the odds (the odds ratio) corresponding to a 1-unit change in X. To build our logistic model we used 10-fold cross validation and we used the probability measure to predict the risk of a given property. The image shows the confusion matrix of the logistic model and 83.44% Accuracy.

Fig. Confusion Matrix of the model

5.2. DECISION TREE MODEL

Decision tree is an inductive inference based model, which is built based on the data provided and predicts the output based on previous learnings. The tree build is robust to noise. Here we used the CART model (Classification and Regression trees) with a type of 'class', which indicates a classification tree is being built. Using which we have performed hyper parameter tuning and a complexity parameter is determined as 0.03 using tune grid function during training the model. The final model was built with an accuracy of 96.91%

	Reference	
Prediction	0	1
0	281	12
1	0	15

Fig. Confusion Matrix of the model

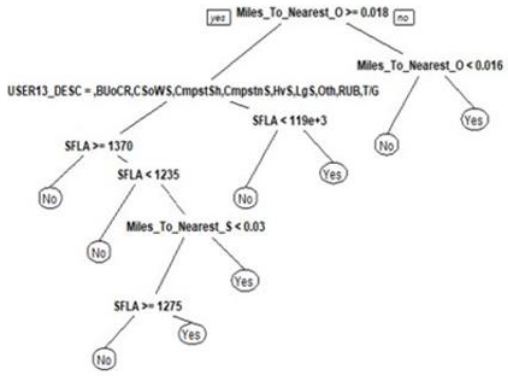


Fig: Decision Tree

5.3 RANDOM FOREST MODEL

A random forest is built over decision trees. Even though decision trees are easy to build and use but they do not perform well practically, that is they are great with the training data but lag with the testing data or with new samples. Random forest combines the simplicity of decision trees with flexibility resulting in vast increase in the accuracy of the model.

Here the original data set is split into multiple boot strapped data sets, then a model is created by randomly sub-setting the original data set at each point and this process is repeated until models are on all the data sets. Random forest is an average multiple deep decision tree that is trained using a training set and with the goal of overcoming the overfitting problem of the classification or simple decision tree. Random forest

is used when we have both continuous and categorical variables. We have implemented this model as our target variable is binary (Incident happened or not), we have used the random forest classifier to predict the class probabilities for all the data instances to check the instances that fall under which class. We ran the model with 80% of trained data and 20% of test data by which we observed an accuracy of 98.5% with 100 trees. But we observed OOB error (Out of Box) as 1.06% while considering 4 variables at every split(mtry value). We used cross-validation of 10 folds during training to avoid overfitting and underfitting.

Fig. Confusion Matrix of the model

	Reference	
Prediction	0	1
0	293	0
1	12	3

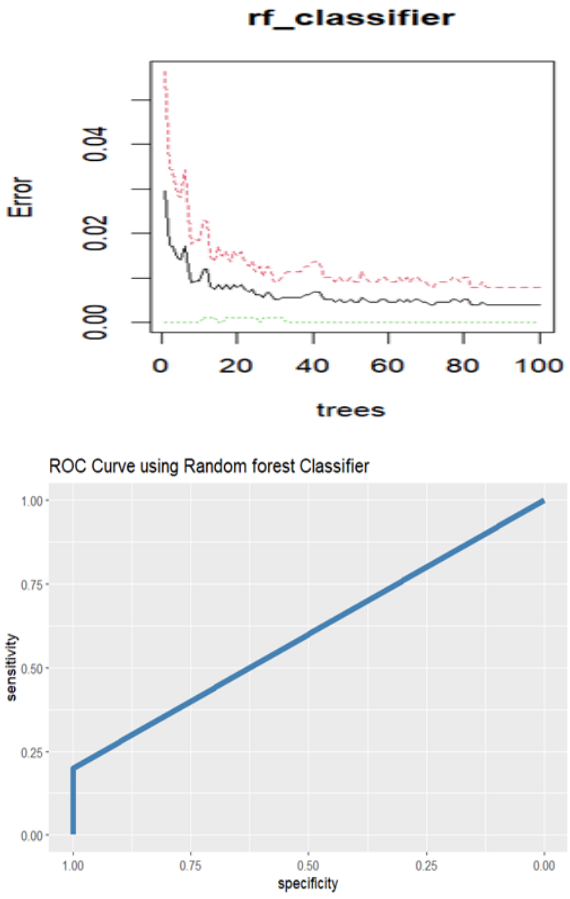


Fig: Change in OOB-Error rate based on no. of trees, ROC Curve

Random Forest has a better ROC area compared to the other models for dataset. As the sensitivity increases in the Random Forest Model, generalizations with new data can be done more effectively. So, the Random Forest Model has the best prediction accuracy compared to the other models.

```
ApplicationType           5.858755
ApplicationStatus         2.709643
StatusName                5.730930
Miles_To_Nearest_Station 14.401384
Miles_To_Nearest_Office  19.011876
Station                   8.800306
Company                   8.063267
Office                    6.143594
STYLE_DESC                5.845519
RMBED                     6.244594
FIXBATH                   2.885826
FIXHALF                   4.731759
SFLA                     10.414262
BSMT_DESC                 8.384812
HEAT_DESC                 4.048598
USER13_DESC               6.960563
Total_No_Years            9.412782
ACRES                     9.358794
UNITS                     0.000000
CODE_DESC                 3.592326
EXTWALL_DESC_With_Wood_   2.456726
```

Fig. Variable Importance Plot for the random forest model

5.4 STOCHASTIC GRADIENT BOOSTING MODEL

Since Stochastic Gradient Boosting is a generalized framework of Real AdaBoost, Gentle AdaBoost, and LogitBoost, we have decided to use this model for the fire risk prediction. To train this model, we have used 5 repeated cross validations. Once the model is trained, a confusion matrix was populated to understand the model performance. The confusion matrix can be seen in the figure below. From the figure, it can be seen that there were only six false negatives and positives in the code. These results suggest that the model is successful in predicting the fire risk. The accuracy value is 98% which also suggests that the model has good predictive performance.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 288   3
         1   3  12

               Accuracy : 0.9804
                 95% CI : (0.9578, 0.9928)
    No Information Rate : 0.951
    P-Value [Acc > NIR] : 0.006599

                  Kappa : 0.7897

Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.9897
            Specificity : 0.8000
         Pos Pred Value : 0.9897
         Neg Pred Value : 0.8000
             Prevalence : 0.9510
         Detection Rate : 0.9412
   Detection Prevalence : 0.9510
      Balanced Accuracy : 0.8948

       'Positive' Class : 0
```

Fig: Confusion matrix and Accuracy scores of Stochastic Gradient Boosting

6. PERFORMANCE SUMMARY

The models were tested after they were built to see how well they performed. Due to the imbalance in the dataset, we needed to address this issue to ensure the model could perform well on the testing data. Following training, the model calculates four metrics to assess its performance which are as follows:

• Accuracy

• ROC

• Sensitivity

• Precision

Model Type	Kappa	Sensitivity	Specificity	Accuracy	ROC
Logistic Regression	0.1766	0.8498	0.5333	0.8344	0.6916
Decision Tree	0.6952	1	0.5556	0.961	0.9795
Random Forest	0.3223	0.9607	1	0.961	0.6
Boosting	1	1	1	1	1

Table: Table shows the model comparison values of each model built to predict fire risk.

Because of the significant class imbalance in our data, we cannot utilize accuracy as a performance metric. This is because if the model just assigns "no fire" to every case, it will be right most of the time. As a result, the kappa score is a more relevant indicator because it is derived simply as the percentage agreement that compensates for class imbalances. Our classifier is also assessed by looking at its sensitivity and precision. We used a confusion matrix to test our model's performance. As a result, we should evaluate our model not just on the number of true positives and false negatives it properly classifies, but also on the number of false negatives it generates.

7. CONCLUSION

Our goal is to make this work beneficial to fire department officials. Data science and its use in decision-making processes can assist us in developing predictive models for making better decisions. However, no model perfectly predicts the outcomes. Thus, we do not intend the fire department to completely replace their process with our model. Instead, the model results could assist the fire department in making changes to their fire inspections to better prevent fire incidents.

ACKNOWLEDGEMENT

We thank the Fairfax County Fire Department for providing us with the opportunity to work with Fairfax County fire data. We extend our thanks to our Professor, Dr. Jie Xu, for supporting and encouraging us throughout the project and for providing us the guidance we needed. It was our extreme pleasure to work on this project and the knowledge we gained by working with real world dataset was wholesome.