FINAL PROJECT PROPOSAL

AY6020 – Predictive Analytics (CRN: 82556)

A picture containing black, photo, red, sitting

Description automatically generated

Instructor: Valeriy Shevchenko

Submitted by: - Dhanashri Jadhav

**Group BETA**

**INTRODUCTION**

The purpose of the final project proposal is to build and develop a predictive model which will help in identify the real-world problem for the dataset that we will be using for our analysis. For the analysis purpose we will be develop a predictive model using various predictive modelling techniques that will be able deliver insights. There is multiple predictive modelling technique such as GLM (Generalized Linear Models, Classification KNN (K-nearest Neighbors), Tree based methods, Clustering, Naïve Bayes Algorithm, and Support Vector Machines. We will be building a model and a hypothesis that would help in significantly find the challenges and roadblocks and identify a suitable predictive model to address the real – world problem or scenario. Our final report will be based on the below structure and the following roadmap: -

![A close up of a device

Description automatically generated]()

We will be following the CRISP-DM methodology to train the model and output for our dataset. The cross industry standard process for data mining or CRISP-DM model is a process framework which helps in designing, creating, framing, building and testing and deploy machine learning algorithm.

## **DATASET**

The dataset we can select and decide to work on is ‘Fire incidents data of Boston.’ This data explains the fire accidents which occurred inside the house properties of Boston. It’s a big dataset with over 230K observational columns and 25 variables. Following is the path of data -

[https://data.boston.gov/dataset/fire-incident-reporting/resource/0f0b5646-f7db-46fb-9951-](https://data.boston.gov/dataset/fire-incident-reporting/resource/0f0b5646-f7db-46fb-9951-b1251cbd9453/view/40394758-048f-4f40-8f29-c30fb6156c14) [b1251cbd9453/view/40394758-048f-4f40-8f29-c30fb6156c14](https://data.boston.gov/dataset/fire-incident-reporting/resource/0f0b5646-f7db-46fb-9951-b1251cbd9453/view/40394758-048f-4f40-8f29-c30fb6156c14)

**PROJECT GOALS**

This dataset is shared on <https://data.boston.gov/>. From the data, we will be able to analyze ‘2016-2019’ data using different statistical and predictive models in either R and develop a Business Intelligence tool in Tableau/RShiny to assist Fire Department of Boston City to highlight the cause, trend and solutions for the unexpected House Fire incidents by answering all / some of the following questions after treating the data -

* Does the safety protocols for condos and apartment towers are standardized in Boston?
* Which areas of Boston houses don’t follow house construction safety policy of fire?
* What’s the trend of house fire incidents in Boston?
* Which are the properties having top fire exposed in Boston?
* Which incidents mostly occurs in Boston?
* What are the main reasons for fire in a multifamily house?
* Top 5 streets of Boston that are affected by fire?
* Trendline of incidents according to Time?

Apart from answering the above-mentioned questions our model will be able to predict the presence of a fire incident in particular area for a given month with a predictive performance better than the prior state.

## **GROUP MEMBERS**

Group BETA consists total of 5 members in the group which are listed below: -

Member 1: - Dhanashri Shivaji Jadhav

Member 2: - Riya Shrivastava

Member 3: - Sangram Thombre

Member 4: - Krupali Rameshbhai Gami

Member 5: - Fabiana Santos

We in a group of 5 will be building a predictive model to identify the real-world business problem and will be using the algorithms and techniques introduced to us in the Predictive Modelling Class. The model implementation will be done in the steps as mentioned below: -

## **WORKFLOW**

* Descriptive analysis
* Multivariate Logistic Regression
* Clustering
* Data Prep
* Time Series models- ARIMA and SARIMA

1. **Clustering**

* Elbow Method to determine number of clusters (k=6).
* K means (Take the Euclidian Distance metric between Incident type and Property use) to assign clusters to the type of incidents.

1. **Multivariate Logistic Regression**

* Predict the likelihood of fire for a given address by training the model on the historical data of fire incidents and evaluate the model on the test set which was not used for training.

1. **Data Preparation**

* We will be removing the null values, duplicated values, missing values and creating dummy variables for digitizing the data
* Stationary Test (Augmented Dickey Fuller Test)- A **stationary time series** is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over **time.**

1. **Data Visualization**

* To perform the data visualization we will be using Tableau, R and RShiny to predict various fire incidents happening in Boston, cause/reason behind the same, Top 10 Properties that are highly exposed to fire and make a time series graph for the fire incidents which provides the time line of the fire incidents happening.

1. **Model Building**

* We will be using Incident Type as an outcome to be predicted whereas all other attributes will be used as a feature in the model.

1. **Model Comparison**

* We will be using factors to evaluate the performance of the model using the Recall, Precision Rate

1. **Improvise the model further if required**

## **GROUP TASK**

Our task will be divided into a group of 5 members. Each of the group members will be working based on their skillset. Each member having a certain type of skillset that they are good will be undertaking and allotted the task accordingly. For the draft purpose we will be dividing each task equally. Below would be the task allotment for each member in a group:-

Data Preparation and Cleaning:- Sangram Thombre (Member 3)

Data Visualization in Tableau and R :- Dhanashri Jadhav( Member 1)

Model Construction :- All members of the group will be actively working and taking part in building the predictive models. (Group Task- Focus on coding and building models).

## **TOOLS**

R, RShiny, and Tableau

## **CONCLUSION**

This model will help the Fire Department of Boston City to analyze the data deeply and thereby reducing the fire incidents, rather than taking help from concerned Analysts and Engineers. The model will also help in saving time with reading the complex reports. This will ensure in providing business insights and smoothen, speed up with the business decisions. With the use of dashboard fire inspectors can easily identify the figures and numbers and take quick business decision, identify the cause and provide numerous ways in which there can be reduction in the number of fire incidents.

## **REFERENCES**

* Retreived from <https://data.boston.gov/dataset/fire-incident-reporting/resource/0f0b5646-f7db-46fb-9951-b1251cbd9453/view/40394758-048f-4f40-8f29-c30fb6156c14>
* Retrieved from Mass.Gov (2020). Annual Fact Sheets 2016-2018. Retrieved from [https://www.mass.gov/service-details/fire-data-and-statististics](https://www.mass.gov/service-details/fire-data-and-statist#istics)