NOVA IMS
Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

A Data Mining Approach to Predict urban Fires in Lisbon using H2o.ai python Library

Luis António Hill Quinta (M20190067)

Project Work presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

A DATA MINING APPROACH TO PREDICT URBAN FIRES IN LISBON USING H2O.AI PYTHON LIBRARY

by

Luis Antonio Hill Quinta

Project Work presented as partial requirement for obtaining the Master's degree in Information Management/ Master's degree in Statistics and Information Management , with a specialization in Business Intelligence and Knowledge Management.

Advisor / Co Advisor: Miguel de Castro Neto

November 2021

ACKNOWLEDGEMENTS

I would like to thank my advisor, Miguel Castro e Neto and the Department of Urban Analytics at Nova IMS, for the help and guidance throughout this thesis project.

I would also like to thank all my family and friends for supporting me and motivating me to achieve my goals throughout my thesis and my master’s degree.

ABSTRACT

Technologies have enabled societies to socially and economically prosper and to be more interconnected. With the decreasing cost of data storage and processing, cities are now trying to extract actionable information from the available data to improve and optimize their resource allocation and planning.

This thesis aims to develop a data-mining approach to predicting urban fires in Lisbon, leveraging both climate, building, and population data available to predict where a fire will happen in the future within a particular period. To aid RSB in reducing their overall response time to fires by predicting probable positive emergency event areas and understand the driving factors that lead to these events in Lisbon.

This supervised learning task developed using the CRISP-DM methodology makes use of standard machine learning estimators using the h2o.ai python module to incorporate parallel distributed computing combined with an AutoML package, evaluated using cross-validation, PR-AUC and F-0.5 score. The main conclusion from this paper is that applying predictive methods of data mining in the prediction of emergency events has a large potential to aid in resource allocation and understanding of drivers to combat emergency events, however requires large amounts of data to for algorithms to learn and extract actionable insights from their predictions.

KEYWORDS

Lisbon; Emergency Incident Prediction; Urban Fires; Predictive Modelling; H2oai;

INDEX

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

CML - Câmara Municipal de Lisboa (Lisbon’s Municipality)

INE - Instituto Nacional de Estatística (National Institute of Statistics)

IPMA - Instituto Português do Mar e da Atmosfera (Portuguese Weather Institute)

RSB - Regimento Sapador de Bombeiros (Lisbon’s Municipal Firefighters)

LIMP - Lisbon Intelligent Management Platform (PGIL)

VUCI - Veículo de Combate a Incêncio Urbano (Vehicle to Fight Urban Fires)

ICT – Information Communication technologies

IoT – Internet of Things

CRISP-DM – Cross Industry Standard Process for Data-Mining

SVM – Support Vector Machines

RF – Random Forest

PR-AUC – Area-under precision-recall curve

AutoML – Automatic Machine Learning

AUC – Area-under curve

1. INTRODUCTION

1.1. BACKGROUND

The first records of firemen started in ancient Rome around the 1st century, shortly after 'The Great Fire of Rome' took place where two-thirds of Rome had been destroyed. During these times, fires were a major concern to the population, most homes were timber-framed, and the destruction was enormous. These groups of men were run by private business owners and incurred a price per service. If negotiations fell through, firefighters would let the fires burn. It was only around the 3rd century that Emperor Nero created the 'Vigiles,' a state-run brigade of firefighters (History of Firefighting, 2016).

The main issue that arose when battling fires was the scarcity of means to combat fires (both in human resources & tools). In Oxford 872, the first fire alarm system was created (Cipriano, 2012). Even though fire brigades were privately owned by insurance companies rather than an organized fire protection system, this leads to faster response times, significantly reducing the devastating effect of wildfires.

In Lisbon, the first organization dedicated to fire response was created by King D. João I in 1395. Before forming the first firefighting organization, fires in Lisbon were put out by local carpenters and caulkers. As there were no water pumps at the time, locals would use axes to clean nearby areas surrounding the fires. (Bombeiros de Gouveia,2004). In 1678, three fireman stations were created around Lisbon to make sure the necessary people and equipment to respond to emergency incidents quickly. Since then, as technologies evolve, the role of firefighters evolve as well. In Portugal, firefighters' scope has evolved, and today they respond to many more types of emergencies such as floods, road accidents, and damaged infrastructures. In 2018 alone, RSB responded to more than 9000 incidents.

By nature, when fires start, they are easily controllable flames, however, as a fire spreads out, temperatures rise, and smoke affects visibility, making these harder to control. A critical factor in containing the spread and eliminating of a fire is the response time from firefighters, pivotal to saving lives and minimizing loss of property (Xin, 2013).

There are numerous studies and models that, over the years, have focused on the importance of quick response times and how to improve them to help emergency services (Police, Ambulances, Fire Brigades). From the development of GPS systems to find the best route to any location to other technological improvements that enable faster resource allocation (Eidam, 2016). More recently, the development of information and communication technologies (ICT's) and other IoT (Internet of things) devices have allowed for much more extensive data collection and advanced analytics in this industry.

1.2. DATA MINING IN CITY MANAGEMENT

Compared to the more traditional statistical analysis, automated data mining applications can deal with more complex data to extract actionable knowledge, identifying hidden trends and patterns. We now can process real-time data at very low costs and provide efficient decision support systems to cities' resource/urban management (Hand, Mannila, & Smyth, 2001).

Thanks to the advancement of IOT's and open data repositories and the increase in data storage and analytics capacity, cities can now take advantage of predictive analytics to aid in the decision-making process of resource management planning. Learning how factors influence the severity of emergency incidents and which locations or timeframes are most at risk of an emergency event can lead to a more efficient allocation of resources and lower response time on behalf of emergency response services (medical services, firemen, police).

Machine learning tools have helped deal with the more complex problems within emergency event prediction in urban management. Accounting for both temporal and spatial fluctuations in emergency events such as human-caused fires, road accidents, or damaged infrastructures has made data mining processes and tools the preferred approach in tackling these issues.

There are many studies relating to the use of data mining for fire prediction. The majority of studies done to date focus on wildfire risk prediction and severity since wildfires tend to have a more critical impact on the environment. These studies use climate & spatial-temporal data combined with machine learning algorithms such as random forests and neural networks (NN) to predict areas where fires are more likely to happen and how extreme a given fire can be. However, few studies have focused on predicting urban fire risk across cities all over the world. These studies focus on predicting urban areas or properties which are more prone to fires or at higher risk due to overdue inspections and overpopulation. These studies and models have allowed firefighters to take smarter approaches to firefighting.

This work project will make use of the acquired skills in both practical application of python, as well as theoretical knowledge of algorithms and predictive methods of data-mining learned throughout the information management masters specifically in the area of business intelligence and data mining modules to provide a data-mining approach to predicting urban structural fire risk in Lisbon, using population and building demographic data collected from the 2011 Portuguese census survey and real-time meteorological data collected across three stations in.

2. LITERATURE REVIEW

Many researchers have developed data-mining models around the world to improve resource planning and decision-making of emergency responding entities. From predicting the location of crimes in cities to predicting forest fires across an entire country, these models have proved to be useful in predicting future emergency events and, in turn, reducing the negative impact on society and the environment (Mukhopadhyay, et al., 2020).

2.1. WILDFIRE PREDICTION MODELS:

In Portugal, a study made in Trás-os-Montes by Cortez & Morais (2007), to take a supervised learning approach by attempting to predict the burnt area of any given fire in Matosinhos natural park, using machine learning algorithms such as Support Vector Machines (SVM) and random forest (RF). The study used daily climate data collected from sensors around Matosinhos natural park combined with historical incidents logs to understand the relationship between meteorological indicators and the size of fires.

A related study was conducted across the entire USA in 2017 by Xiong (2017) used 24 years of fire records (1.8 million fires) combined with climate data to predict the size of any given wildfire within the USA. The main difference was this problem was framed as a multi-class classification problem, while the Portuguese study framed the problem as a regression. The main goal of this study was to understand the driving factors of wildfire severity across the USA. Similarly, a study to understand the wildfire driving factors for each of the six regions of China by Ma, Feng, Cheng, Chen, & Wang (2020) used climate, socio-economic, and spatial data to create data-mining models for each geographic region to predict locations where wildfires will take place daily.

2.2. PROPERTY LEVEL URBAN FIRE PREDICTION

For this project, we focus more on urban fire prediction rather than wildfires. Urban fires relate to structural fires, and in this field, there are two types of models to predict fires. Property level refers to predicting specific buildings at risk of fire, and community level which refers to predicting areas where fires are more likely to happen.

In New York, a property-level urban fire prediction model named 'Firecast' developed by the New York City Fire Department (FDNY) highlights buildings that are more vulnerable to fires (Heaton, 2015). The model analyses data for each building in the city considering over 7500 distinct risk factors, and using data-mining techniques, compute a risk score for each building. The risk score allows FDNY to prioritize inspections of commercial buildings that are most at risk, significantly reducing the risk of fires. This is an example of 'smart firefighting’ where machine learning algorithms help transform the firefighting methods from reactive to pro-active (Roman, 2014).

A very similar property-level urban fire prediction model was created in Atlanta, GA, named 'FireBird,' in the same way as 'Firecast'; the main objective of this model is to predict the risk level of each of the over five thousand properties in Atlanta, GA. To do this, they collect data from different sources relating to building information, socio-demographic and financial information on residents, commercial licenses, and fire permits to obtain a total of 252 variables per property (Madaio, et al., 2016) and apply predictive algorithms to predict risk scores.

The property-level urban-fire prediction has proved to be a more useful tool than the community level. Despite having lower model accuracy due to fewer positive events per location, understanding which buildings are most at risk and the main factors that explain the risk at a building level allow Fire departments and government agencies to act more proactive rather than reactive. (Walia, et al., 2018) The issue with developing models of such a nature is the granularity of the data available. Most building and population data is mainly available through census surveys and, therefore, usually aggregated at census block levels.

2.3. Community Level Urban Fire Prediction

In San Diego, California, a model was developed to predict areas more susceptible to any kind of emergency incident (Medical, fires, road accidents). The general goal was to highlight which areas are at the highest risk tomorrow based on previous days (Romero, Barnes, & Cipollone, 2016). Rather than using census blocks, this study split the city into 400x400 meter grids, selects the top one percent of incident risk locations, and uses those for daily planning of emergency services.

A study made in Pittsburgh is perhaps the nearest precedent for this study. In this study, two predictive models were created. One was a property-level fire prediction model aimed at commercial buildings in Pittsburgh. The second was a community-level fire prediction model aimed at residential buildings and split into three hundred and fifty census blocks. (Walia, et al., 2018), Like the previous studies analysed, the property level model is used to evaluate risk scores so commercial inspections can be prioritized. On the residential model, since properties do not tend to be inspected, each census block's risk scores are used to prioritize education efforts on fire safety.

As a whole, most models and studies that relate to fire predicting, whether wildfire or urban fire prediction, show that a risk score for each location, given a specific time frame, is the optimal output for firemen. This score allows for fire departments to select the risk cut-off point to obtain the more at-risk locations. The selection of the cut-off point is based on the availability of resources to Fire departments (i.e., Number of VUCI available). Therefore, this study will also aim to develop a model that returns each census block's periodical fire risk levels in Lisbon. The main difference between all urban fire-prediction models analysed and this study is the inclusion of climate data. Therefore, it combines features used in wildfire prediction (weather and topography) and features used in urban fire prediction (infrastructure and population).

2.4. Optimization of firefighter response in Lisbon with predictive analytics

A previous master thesis study was conducted in 2020 by, using historical data from RSB, census data, and climate data to develop a machine learning model to predict positive emergency events that RSB responds to. The main purpose of the study was to highlight which features have high importance to predict the target variable.

The granularity levels of the data that were used were the most detailed levels of granularity available which meant that all 3351 census-blocks were used as locations and the time frame for predictions was hourly, which meant to predict hourly events the overall size of the training data frame had over 150 million data points, which ultimately affected the computational performance of the model.

Having a very large dataset means that a severe imbalance of the target variable was present within the data. To overcome the severe imbalance, over and undersampling techniques were used to test which sampling method obtained better performance on the test set for each estimator.

The three machine learning estimators that were used in the study were; Logistic regression, Decision Trees, and Random Forests, of which the best performing estimator was a Random forest algorithm with random undersampling, achieving an AUC ROC score of 0.67 and an F1-score of 0.41.

In an attempt to build and improve on what has been previously done, this study will use the same data sources used in the previous 2020 study, however rather than predicting any type of event that RSB responds to. A model will be created to predict only 1 type of event in the belief that each type of event has different characteristics, optimal predictive features, and estimator algorithms for a single type of event can vary differently from another and therefore should be modeled separately.

2.5. Spatial & Temporal Resolutions:

In the literature review, many different spatial and temporal resolutions are used. A 2019 study on the optimal resolution of space and time for machine learning concluded that the resolution depends on the specific needs of the problem at hand (Bao, Liu, & Ukkusuri, 2019).

A study that reviews several incident prediction models, concluded that when dealing with a low amount of positive events within the data, increasing the spatial or temporal resolutions discretization might reduce the accuracy of various methods, such as tree-based and deep learning algorithms. The loss in accuracy occurs because if some areas have 0 counts, this will cause a bias within models that use statistical learning (Mukhopadhyay, Pettety, Vazirizadey, & Lu, 2020).

2.6. FEATURE SELECTION:

Feature selection is also an essential aspect of studies relating to emergency incident prediction since model accuracy is dependent on selected features (Mukhopadhyay, Pettety, Vazirizadey, & Lu, 2020). Including too many features in a dataset will make the model 'noisy' and become more prone to overfitting (Saurav, 2016). Since this is a supervised machine learning problem, we can apply wrapper methods of feature selection which include, backward elimination & recursive Feature elimination, as these are seen to generally lead to better scores of classification & regression models than traditional unsupervised filter methods. (Cai, Luo, Wang, & Yang, 2018). This type of feature selection is also included in the FireBird model.

For wildfire predictions, where climate data plays an important role, the key features focused on rain, humidity, and temperature, not only instantaneous but also year and season averages. However, the rain variable is more critical in determining wildfire severity rather than the cause. For both levels of urban-fire prediction types, the main features that are considered essential to predict positive events are mainly building characteristics (year built, land area, property type, percent occupied, smoke detector) population demographics, and financial data such as (tax amount, land value). Since these variables are consistent across other models, we will assume these variables have fundamental importance in Lisbon's fire prediction.

2.7. MODELS:

Both linear and multi-linear regression has been tested in several models relating to the prediction of emergency events. In an analysis of emergency prediction models, it has been concluded that this type of estimator fails at modeling the complexity of emergency incidents (Mukhopadhyay, Pettety, Vazirizadey, & Lu, 2020), and therefore will not be considered in this study.

Based on the analysed studies, the main algorithms that resulted in the highest accuracy scores for emergency incident predictions were; Random Forests and other tree-based algorithms. The benefits found are that tree-based algorithms are better than logistic regression and support vector machine algorithms at handling severely imbalanced datasets (Haixiang, et al., 2017) and can disregard non-important information faster which is more cost-effective. Furthermore, Neural networks and other deep learning algorithms have also been used in the prediction of emergency events within cities, most studies indicate that deep learning is more effective with datasets that contain a large number of positive cases (Bao, Liu, & Ukkusuri, 2019).

2.7.1. H2o.ai

H2o is an open-source application that is used for machine learning. The python module of H2o has functions that automate much of the training and evaluation of supervised machine learning problems, and increase transparency of models. The use of h2o has not been extensive in the prediction of emergency events. However, this technology has become a leading tool in the development and deployment of many machine learning models. The ease of connection to a cluster and the simple scalability into distributed computing without extra code, whilst having functions such as AutoML, that run multiple estimators and ensembles in a standardized way has made h2o the preferred tool for many data scientists (Devisschere, 2021). The H2o.ai platform is also a preferred tool due to the APIs offered for users to integrate with other technologies.

2.8. MODEL EVALUATION:

In the analysed models, a majority of the analysed studies use random cross-validation methods to assess the model performance. Nevertheless, some frame their models as a time series problem where cross-validation is not random. Framing the task as a time series implies that events and features are dependants on time, and therefore the best approach to test if a model will be able to deal well with future data is to check if past data can predict future data.

When it comes to model selection and evaluation, cross-validation ensures the model scores are consistent against unseen data (Brownlee, 2020). A report which analyses different cross-validation techniques for various parametric models provided a greater insight into the benefits and disadvantages of each type of split for cross-validation. The general approach was to test random versus block split; block split refers to applying cross-validation with the dependant variables. The main conclusion of this study was that block-split cross-validation can cause bias in the model due to the user selection of blocks, but when properly selected the results to tend to be closer to the true performance of a model against unseen data compared with Random split. Random split is the preferred cross-validation technique when the user is not very familiar with the data.

2.8.1. Evaluation Metrics:

Due to this project's nature and the negative implications of a poorly accurate model, it is essential to consider the metrics chosen to evaluate model performance. Based on most studies analysed, emergency incident prediction, datasets tend to be severely imbalanced, where positive events have a one to one hundred ratio to negative events. Therefore common classification metrics such as accuracy and precision can be misleading about the actual performance of the model (Haixiang, et al., 2017).

When considering the business case at hand, the weighting of false positives and false negatives is different, depending on the model's application. For optimizing fire inspections, it is essential to prioritize classifying the positive class rather than minimizing false positives because false negatives would result in no inspection and, therefore, adverse outcomes (Walia, et al., 2018), while false positives result in inspections that have no adverse outcome. On the other hand, if a model is used to optimize resource allocation (which is the case for this model), then false positives will allocate resources where no event takes place, and if resources are scarce, this will have an adverse outcome on the business.

A commonly used metric to evaluate imbalanced models is the and area-under-receiving operating curve ("AUROC"), where the relationship between specificity (True negative rate) and sensitivity (true positive rate), at different threshold levels is plotted (Narkhede, 2018). However, AUROC weights false positives and false negatives equally, therefore depending on the problem at hand, AUROC can also be misleading.

To evaluate the model taking into consideration the business case to focus primarily on improving true positives and reducing false positives, 2 metrics are used; Area under the precision-recall curve (‘PR-AUC’) and F0.5 score.

The precision-recall curve (‘PRC’) shows the ratio between precision and recall for different threshold levels (Pedregosa, Varoquaux, Gramfort, & Michel, 2011). The main difference between the PRC and the ROC is that precision-recall curves do not consider true negatives, which helps understanding how well the model predicts the positive events (Ekelund, 2017). Unlike the AUC ROC curve where the baseline used to compare any model is 0.5, the PR-AUC curve’s baseline has to be calculated based on the number of samples tested.

F0.5-score is also seen as a critical metric when dealing with severe imbalance data in emergency incident prediction. The F1-score places equal importance on both precision and recall, but changing the beta from 1 to 0.5, places higher importance on precision and less to recall resulting in a better understanding of how the model predicts positive events (Brownlee, A Gentle Introduction to the Fbeta-Measure for Machine Learning, 2020).

3. METHODOLOGY

The methodology that will be used for this data mining model is the 'Cross Industry Standard Process for Data-Mining' (known as CRISP-DM). This is one of the most widely used methodologies in data mining (Brown, 2015). This process was developed in 1999 with funding from the European Union to develop a universal process for data mining.

The main concept of the CRISP-DM structure is the development of a model through 6 phases (Chapman, Clinton, Kerber, Khabaza, & Reinartz, 2000). Beginning with the business understanding where the developer identifies the business goals defined before the development of the model. It is important to consider that the business goals are always subject to changes as time progresses and therefore a robust model needs to account for potential changes in the overall projected applications of the model.

The second phase of the process is understanding the data. This refers to the discovery of the raw data and conducting exploratory data analysis to define whether the available data is relevant for the model and if more data is required for analysis.

The third phase refers to the preparation of the data for modelling, this includes cleaning (missing variables, outlier treatment), feature selection & engineering, and integrating data sources. It is important to consider the modelling stage ahead since some models benefit from extra steps in the data preparation stage such as feature scaling for linear machine learning algorithms (Roy, 2020).

The fourth phase of the process is the modelling section where the data is split into training and testing samples, machine learning models are selected and fine-tuned to train the data, and the model is built and run.

The fifth phase of the process is the evaluation of the results generated from the models and whether these answer the business goals. These results always have to be put into context since depending on the mining goal some metrics are more relevant than others.

The final phase of the CRISP-DM methodology process is the deployment of the model, where the model is documented, reviewed, and actionable information is extracted or the model is put into a repeatable mining process.

The following section describes the CRISP-DM process of the model completed for this thesis.

3.1. Business Understanding

RSB currently responds to over five thousand emergency incidents per year in the city of Lisbon, of which 25% are fires. This paper, however, focuses on predicting fires in the city of Lisbon. Despite there being some seasonality to fires around Lisbon, there is currently no system to help predict where or when fires will happen. Firefighters rely on their intuition and expertise to manage their resources.

Based on the evidence found on the importance of a rapid response time to reduce the severity of an emergency event, the main goal for this model is to be able to reduce the response time to emergency events to within five to ten minutes. To achieve this response time, firefighters need to understand which areas are at greater risk of a fire constantly to allocate first response emergency vehicles (VUCI) to high-risk areas.

The concept of the model will be to predict where and when fires will take place. However, the model is designed not specifically for fires, but for all types of emergency, events responded to by RSB. Fires have been chosen as the preferred emergency type to predict, due to the negative externalities caused by fires compared with other emergency events. Fires as an emergency event have a sporadic nature, and therefore models which can predict complex events, will in theory perform well on less complex emergency events, such as floods.

3.2. Data Understanding

The available data comprises three separate datasets. The first dataset is the Historical log of incidents between 2013-2018 provided by the RSB of Lisbon. The second dataset is provided by the national institute of sea and atmosphere with weather recordings from 3 different stations around Lisbon for the same period. The third dataset is the census data aggregated at the census block level provided by the Portuguese national institute of statistics. In this section a further description of each dataset and the connection to the target variable.

3.2.1. Event Dataset

- 67,656 rows relating to all events (emergency & non-emergency) responded by the RSB Fireman in Lisbon, between the dates of 01-08-2013 to 31-12-2018.
- 9 columns, relating to time, date, and location of event, type of event (i.e. Fire, Flood, Road Accident, etc..), and relevant ID's.

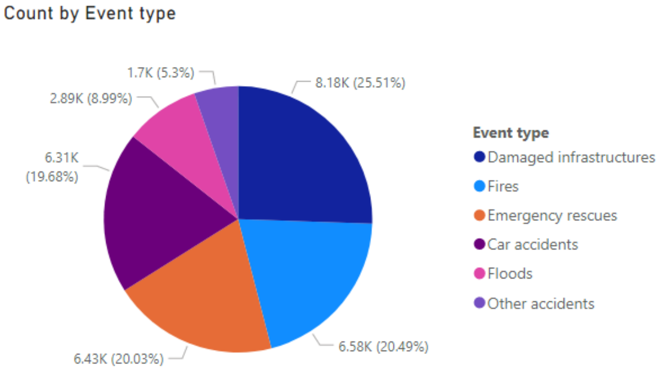


Figure 2: Distribution of event types

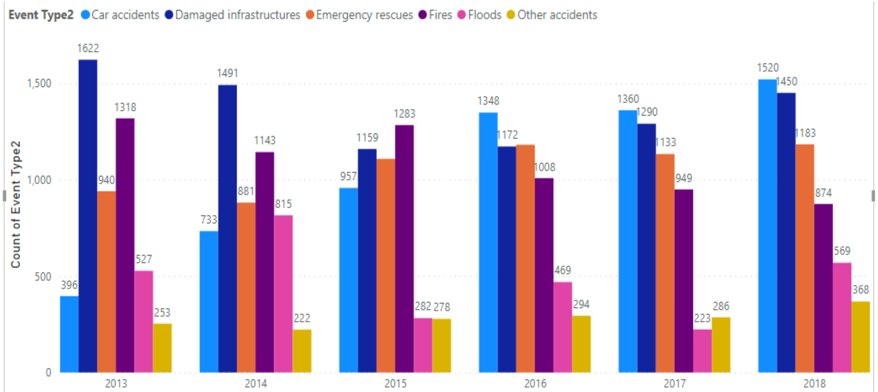


Figure 3: Emergency events count per year

The above figures show the distribution of event types responded by RSB. For this model, we disregard the 'Low severity' as an event type, since RSB operates differently for non-emergency events. From the yearly distribution, we can visualize how events alter throughout the years. For the case of fires, these are decreasing year on year, however, the opposite occurs in car accidents. It doesn't necessarily mean that there are more car accidents in general only that there are more car accidents where RSB responded. For this study, the target variable will be chosen as fires (0 equals no event, 1 equals positive event).

3.2.2. Weather Dataset

The second dataset was provided by the Instituto Português do Mar e da Atmosfera, containing the hourly meteorological data between 2013 and 2018 across three different weather stations in Lisbon.

- 156,652 rows where each row represents hourly timestamps (between 01-01-2013 to 31-12-2018) for each of the 3 weather stations in Lisbon.
- 56 columns, relating to weather indicators (i.e Humidity, Temperature, precipitation, etc..) and time characteristics (i.e. Day of the week, season, period of the day, etc..)

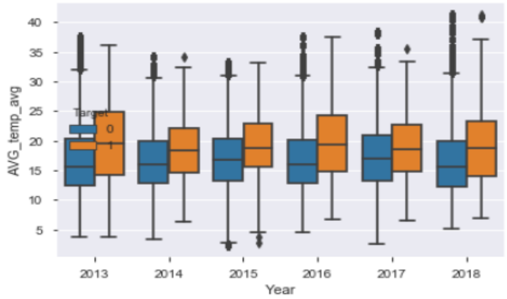


Figure 4: Average temperature per year per target

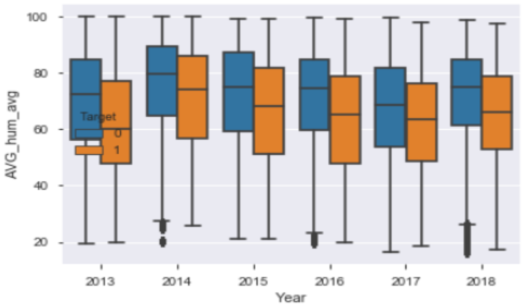


Figure 5: Average humidity per year per target

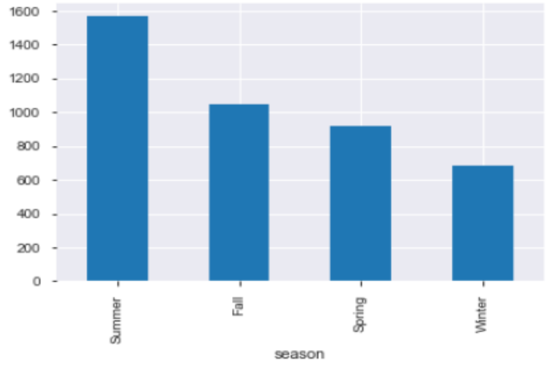


Figure 6: Count of fires per season

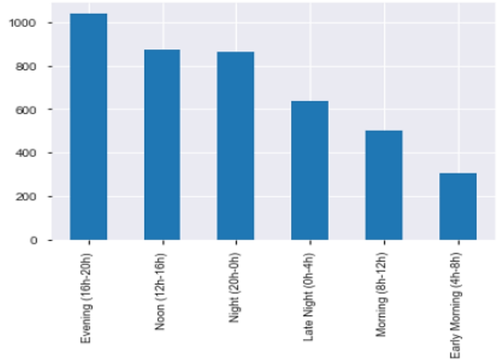


Figure 7: Count of fires per period

To better understand if the available data has useful information to make predictions, we explore our variables to discover relationships between them and the target variable.

Starting with the weather dataset, we can see from the charts above that our target variable has some seasonality to it. We can see that the target variable occurs more often in the summer, which can also be seen by the average temperature and humidity of positive and negative values. From this we can assume there is some sort of seasonality, this can be interpreted in two ways. The first is we can assume that high temperatures and low humidity levels cause fires to propagate faster and therefore cause positive events, or we can assume that for an unrelated reason there are more fires in summer, where coincidently temperatures are higher and humidity levels are lower.

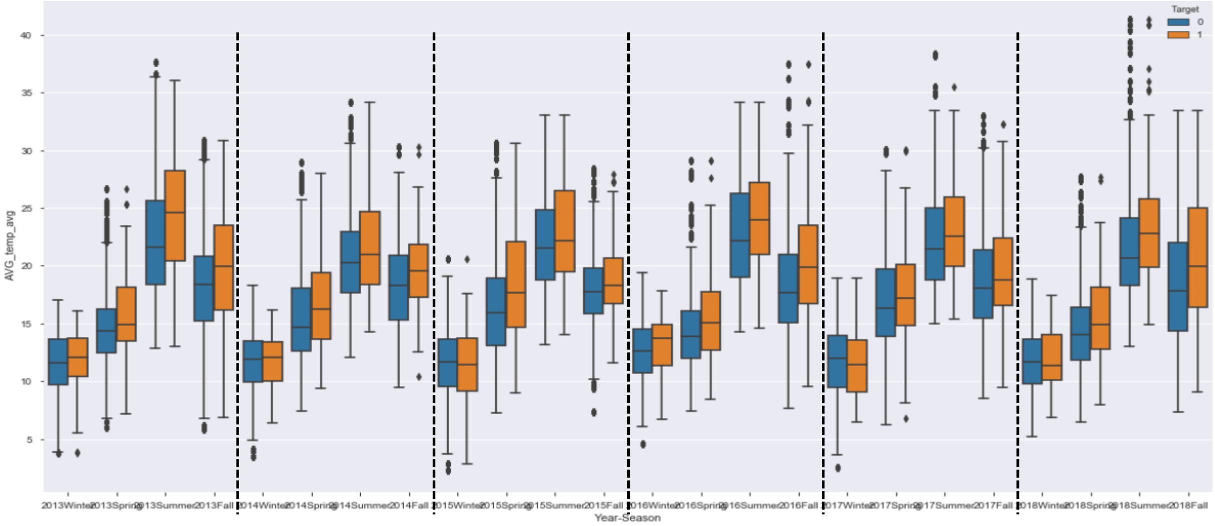


Figure 8: Average temperature per season per year per target (blue = no event, orange = positive event)

When viewing temperature and humidity graphs per month-year, we find evidence that temperatures are on average higher and humidity levels are on average lower for positive events, independently of the season, suggesting that these features can be important in predicting a positive event.

3.2.3. Census Dataset

The third dataset is the 2011 census survey, which includes a snapshot of 2011 building characteristics and population data within Lisbon and is grouped geographically into 3,551 subsections.

The Census data consists of:

- 3351 rows which are the 3351 census blocks at the lowest level of granularity.
- 102 columns, which are the characteristics and demographics of each census block. These are broken down into:
  - ➢ 9 Columns relating to IDs, coordinates, and associated weather stations.
  - ➢ 56 Columns relating to population demographics (n. men & woman, n. residents w/ secondary education, etc..)
  - ➢ 37 Columns relating to building demographics (n. classical buildings, n. concrete buildings, etc..)

Since columns are highly correlated with each other, there won't be a major shift between trends of predictive columns and the target variable. Within the aggregations, some census blocks are much larger in area than others and therefore have more buildings and larger population size, which in turn correspond to a higher number of positive events. Figure 9 below shows that as the area increases the number of positive events also increases however not exactly linearly. The same can be verified in figure 10, graphing the number of residents against the target variable.

Since both figures above are almost identical, to better understand the relationship between these two variables and the target variable a new graph is created that combines both area and population into 'Population Density (number of residents / Area of c-block), in an attempt to extract further insights.

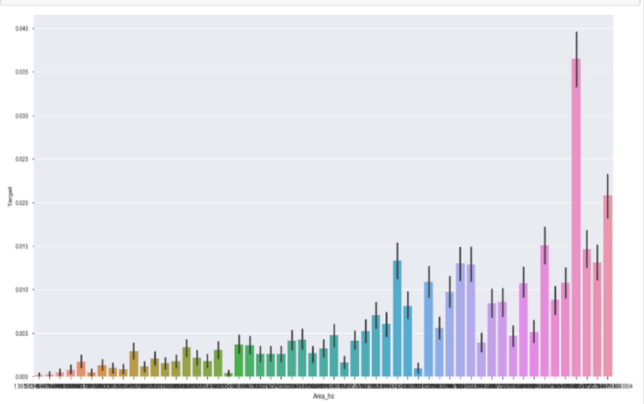


Figure 9: Count of fires per Area

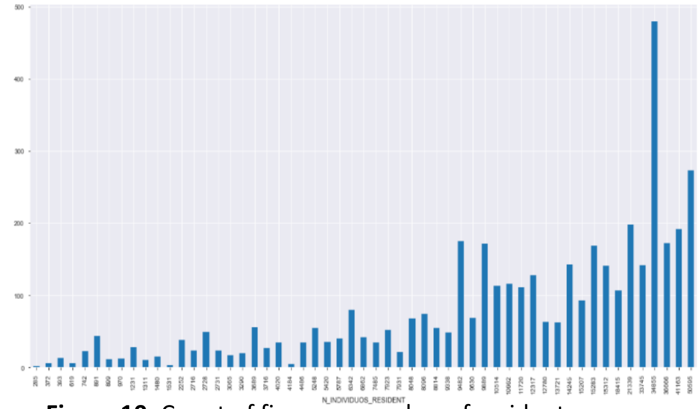


Figure 10: Count of fires per number of residents

3.3. Data Preparation

3.3.1. Data Cleaning

3.3.1.1. Missing Values

The data cleaning section of this model was initially done in a previous thesis using the same data sources (Teixeira, 2018). The only dataset that contained missing values was the IPMA dataset. As there are two other stations, the data was filled using the closest weather station data. In the case where all stations had missing data for the same period, then the averages of the previous and the next hour were inputted. In the case of wind direction, 0.5% of records were not recorded on either station and therefore the authors used a random forest algorithm to predict the missing values. There were no missing values for the census or the RSB historical dataset.

3.3.1.2. Multicollinearity

To reduce a potential bias within the data, it is common practice to remove features with high multicollinearity. Within the weather data, for each meteorological measurement (humidity, temperature, wind speed, etc..), there are three features, the instant recording, the one hour & two hour average of the indicator measurement. These columns are very highly correlated with each other, so we chose to remove the instant and the 2hr measurement, leaving only the 1hr average.

Within the census data, many features are highly correlated to other features. Some features overlap and double count the statistics. (i.e., n.Residents aged 20_65 & n.Residents aged 25_65), Other columns are highly correlated with other features, (such as students attending secondary education vs n.Residents aged 14_20. The treatment of multicollinearity between features will be the same as in the previous study where Pearson's correlation coefficient is greater than 0.8 were removed from the dataset, a full list of the included/removed features is in appendix 3.

3.3.2. Data Integration

3.3.2.1. *Geospatial* interpolation

To combine the weather dataset with the census dataset, a weather station was assigned to each of the 3,351 subsections of the census. This was done using simple interpolation using ArcGis based on the shortest Euclidean distance between the weather station and subsection. No topography was taken into consideration in this interpolation.

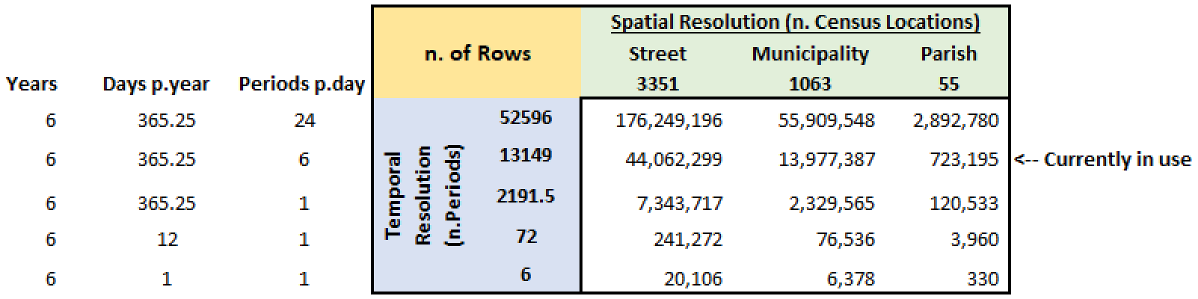
3.3.2.2. Spatial-Temporal Resolutions

Based on the analysed literature, it is essential to create the right balance between spatial and temporal discretization but also to consider the implications that changing resolutions and granularity of the data might have on the model and real-world application. Larger datasets might require under-sampling or dimensionality reduction to predict efficiently; however, increasing the resolution of the data can also create bias and cause overfitting. (Mukhopadhyay, et al., 2020)

For the weather data, to reduce the size of our dataset, we have decided to rather than use hourly meteorological data, we aggregate a day into six blocks of four hours each (see appendix 1). The main reason for this is that when put into context the application of this model, firefighters tend to work in shifts of four hours. When considering future predictions the data that is being used will be IPMA forecasted weather, which itself contains a predictive error, rather than the actuals used for training the model. Therefore using 4-hour averages reduces the potential error that these forecasts might have. Since weather indicators are already one-hour averages, the aggregating function will be to take the mean values of indicators within the blocks. This reduces our weather dataset by 75% from 52,584 to 13,149 rows.

In terms of the census data, there are three possible aggregations, 3351 streets, 1063 municipalities, or 55 parishes.

Figure 11 shows the potential space and time combinations that could be tested for the available dataset.



				n. of Rows	Spatial Resolution (n. Census Locations)			
					Street	Municipality	Parish	
Years	Days p.year	Periods p.day			3351	1063	55	
6	365.25	24	Temporal Resolution (n.Periods)	52596	176,249,196	55,909,548	2,892,780	
6	365.25	6		13149	44,062,299	13,977,387	723,195	<-- Currently in use
6	365.25	1		2191.5	7,343,717	2,329,565	120,533	
6	12	1		72	241,272	76,536	3,960	
6	1	1		6	20,106	6,378	330	

Figure 11: Diagram showing the potential breakdown of spatial and temporal resolutions.

When defining the optimal breakdowns of the spatial resolution, it is vital to consider the performance and the computational time of the model. It is also important that the spatial resolution breakdown is still relevant to solve the problem at hand. For the particular case of this study, the main objective is to reduce the first response vehicle travel time to five minutes. Therefore, for the lowest spatial resolution available, which corresponds to the 55 parishes of Lisbon, we created an interactive map that creates a five-minute radius for the ten largest parishes in terms of area (km2) and when compared to the radius to the coordinates of the historical fire log to check if any emergency event lied outside the 5minute radius. All events were captured within the ranges of the five-minute drive, which means that using the lowest census special resolution would still be useful when considering the problem at hand.

For this methodology, we assume that the spatial resolution used will be 55 location blocks, and the temporal resolution used is 10,950 periods (5years, 365 days, six 4-hour blocks). These resolutions were chosen because they are the optimal resolutions where predictions are most useful for RSB.

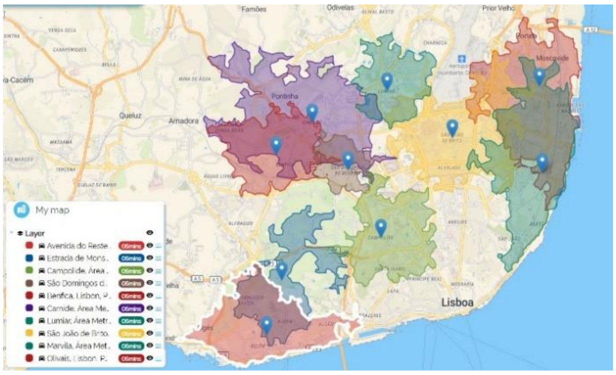

My map
Layer
Avenida do Reste.. 05mins
Estrada de Mons.. 05mins
Campolide, Área .. 05mins
São Domingos d.. 05mins
Benfica, Lisbon, P.. 05mins
Carnide, Área Me.. 05mins
Lumiar, Área Metr.. 05mins
São João de Brito. 05mins
Marvila, Área Met.. 05mins
Olivais, Lisbon, P.. 05mins
Queluz
Amadora
Lisboa

Figure 12 Oalley Map, showing 5min radius from largest parishes

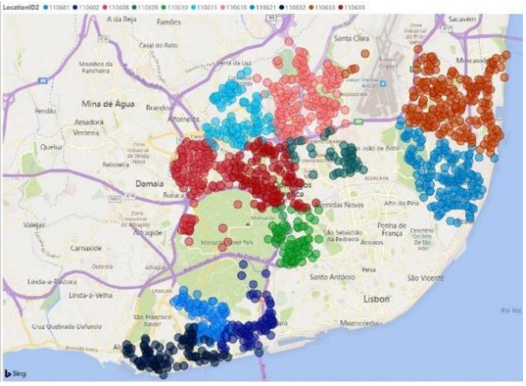


Figure 13 Map of Fireman Incidents locations in largest parishes

Similarly, as seen in the methodology section with the census dataset we concluded that predictions are more practical for RSB if the data is aggregated at parish level rather than street or district level. Therefore reducing the number of rows from 3351 to 55, aggregating features by summing the values rather than taking the mean.

3.3.2.3. Cross Joining data frames

Once the granularity level of each dataset has been decided and aggregated, the next step is to cross-join both tables, so that we create a paired combination of each row of one table with the other, obtaining the product of both tables. With the granularity levels that were decided the final cross joined table has 55(parish's)*13,149(periods) = 723,195 rows. Figure 14 below shows the selection of granularity levels.

With the main data frame is created, joining weather indicators for the past 6 years for each of the fifty-five census block locations, we now need to identify rows where a positive event had occurred. Using the information provided on the Event dataset, a left joined is used to identify which rows (census-block, date-period) correspond to a positive event. In this case, the positive event was 'Fires', however, any other emergency event which RSB responds to can be used as the target variable.

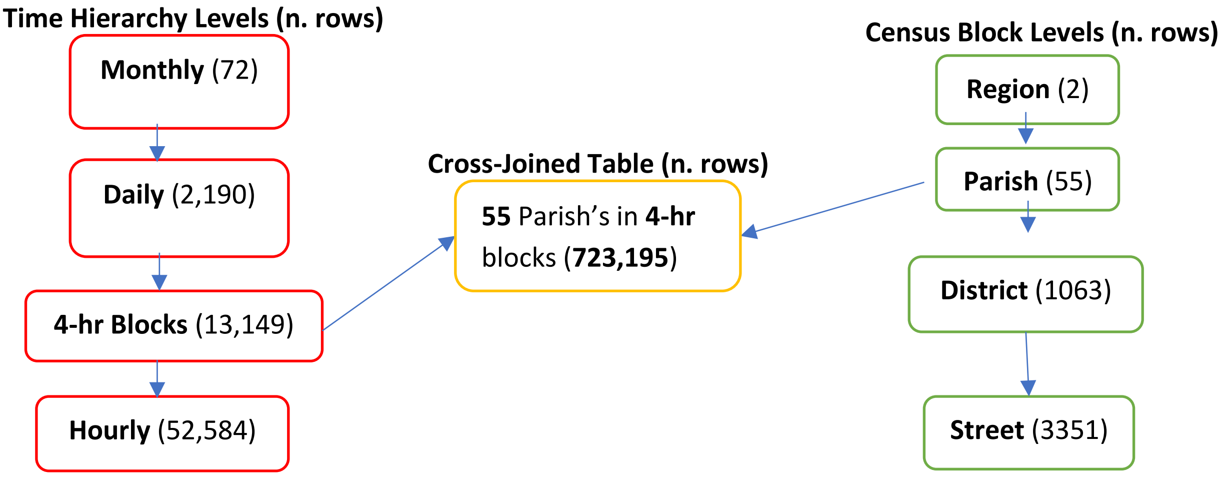

Time Hierarchy Levels (n. rows)
Monthly (72)
Daily (2,190)
4-hr Blocks (13,149)
Hourly (52,584)
Cross-Joined Table (n. rows)
55 Parish's in 4-hr blocks (723,195)
Census Block Levels (n. rows)
Region (2)
Parish (55)
District (1063)
Street (3351)

Figure 14: Diagram showing the selection from hierarchy levels of datasets and cross joining dataframes

3.3.3. Data Pre-processing

3.3.3.1. Imbalanced Dataset

Now that our data set is created and clean, we need to prepare it to be run through the algorithms. A vital step to consider is the bias in our data. Currently, within our dataset, our positive target variable accounts for 0.58% of our dataset (4220 positive events). It is expected that having an imbalanced dataset will affect the performance of the model and will limit the metrics that can be used to evaluate model performance.

Therefore three different balancing approaches are tested to understand their significance towards the model performance. The first strategy is to oversample the minority class, which means we synthetically increase the number of positive records to obtain a 50/50 split and a total of 1,437,074 records. The second strategy is to under-sample the majority class, which means removing records of the majority class to end up with a 50/50 split and a total of 8,438. The third strategy is to use a combination of both over and under-sampling, which has proven to increase the accuracy of predictions (Nutthaporn Junsomboon, 2017). The final strategy is not to make any changes to the created dataset and leave the imbalanced. All strategies will be tested and selected based on the best performance metrics.

To test these methods of balancing the data, a grid search will be run where different balancing ratios are used. Essentially the 3 tests will be to use no sampling, under-sampling, and oversampling. Despite the previous study's best-performing model being a random forest with under-sampling, since a different library and estimators will be used, there is no certainty the best model will be the same.

3.3.3.2. Feature Selection

From the literature review, we have seen examples of feature selection. Feature selection is the process of selecting variables to keep or remove from the dataset. Removing some variables can reduce noise and improve the computational performance of the model whilst increasing the interpretability of results by reducing the complexity of the model (Saurav, 2016). Nevertheless, removing too many variables might cause model bias.

3.3.3.3. Unsupervised feature selection

Within the pre-processing phase of the process, a manual selection of features that are not relevant for the study is removed, such as survey responses that relate to the timeframe of the questionnaire (a timeframe not in the scope of this project). Features that also have high multicollinearity will be removed or used to create new features that reduce multicollinearity because as seen in the literature review, features that are highly correlated to other features can cause some bias in the model. It is important to consider some correlations are not removable due to their perceived importance to the model therefore in the modelling section, models that are not affected by multicollinearity are used.

3.3.3.4. Supervised feature selection

Since the model is computationally intensive, wrapper methods of feature selection are not used. The h2o module doesn't have any specific feature selection function however, most algorithms used to train the model, especially tree-based models and neural networks, use feature ranking as part of the training process and therefore reduce the need for feature selection in the pre-processing stage. The results from the training algorithms also show the variable importance in predicting the target variable, which can be used to interpret the model itself and for further analysis to prevent future positive events.

3.4. Modelling

3.4.1. Train test split

The holdout method will be used for training and testing the model. Since the target variable has an even distribution across years, we define time as a good feature to split the data. Therefore, all data before 31-12-2017 is be used to train the model using cross-validation. Since all our features are independent of time, a time series cross-validation is not used, instead, k-fold cross-validation is used.

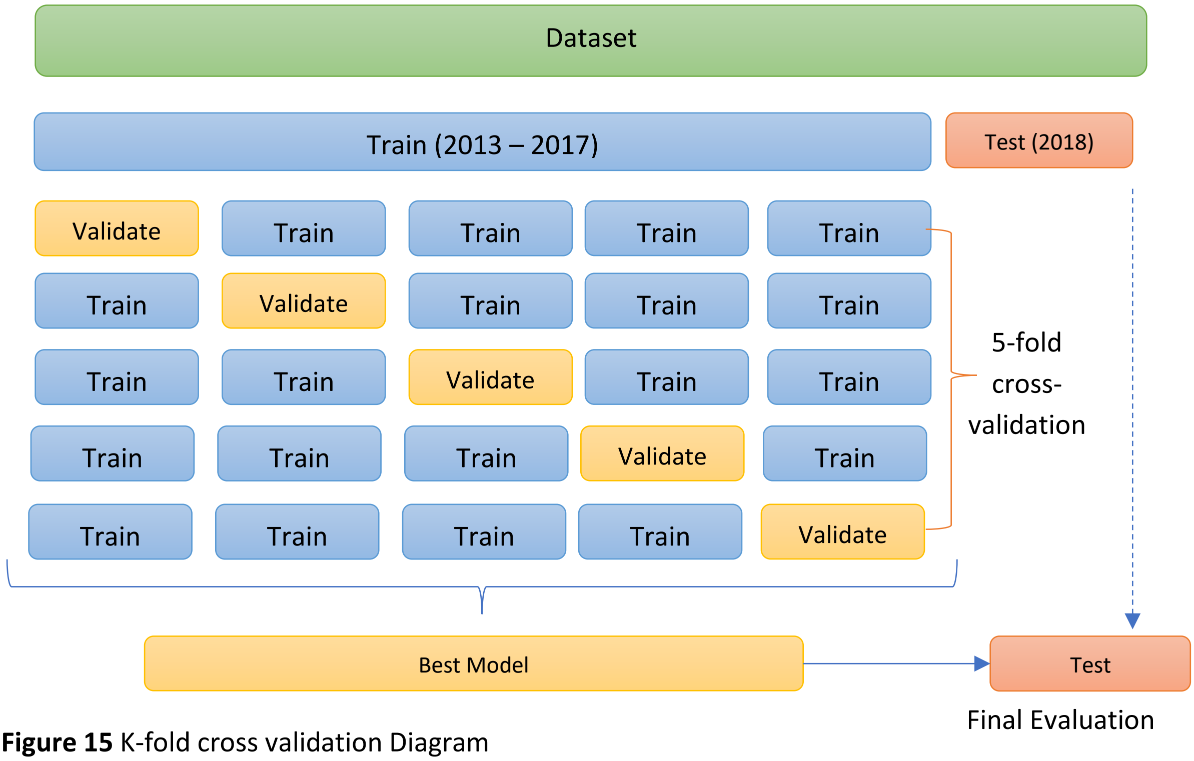


Figure 15 K-fold cross validation Diagram

3.4.2. Auto ML

In the predictive section of the model, the single package used for the supervised machine learning problem is called H2o. The main benefit of using this package versus other machine learning packages such as 'Scikit-learn is due to the easy scalability and integration to enterprise-level deployments such as spark, HDFS, and databricks. This means that machine learning algorithms can run faster and therefore increasing the amount of data will not greatly impact the computational performance whilst continuously improving the predictive performance of the model. Most tree-based algorithms are used to deal with large severely imbalanced datasets.

A commonly used feature of the h2o package is the 'AutoML' command, The AutoML algorithm of the H2o package automates the machine learning training process by creating a pipeline of estimators and ensembles. The full process is described as "trains and cross-validates the following algorithms (in the following order): three pre-specified XGBoost GBM (Gradient Boosting Machine) models, a fixed grid of GLMs, a default Random Forest (DRF), five pre-specified H2O GBMs, a near-default Deep Neural Net, an Extremely Randomized Forest (XRT), a random grid of XGBoost GBMs, a random grid of H2O GBMs, and a random grid of Deep Neural Nets." (LeDell & Poirier, 2020)

Since the Lisbon Intelligent Management Platform (LIMP) is running in a distributed cluster, the use of the h2o package will reduce computing time and simplify the code necessary to achieve parallel computing. Therefore, the approach is to run the 'Automl' as a baseline which retrieves a list of best models based on chosen criteria which in this case is Area under the curve of precision and recall (PR-AUC). The result of the AutoML is then analysed to select the estimator that best performed. A model is then run with a grid search in an attempt to fine-tune the model.

3.4.3. Hyperparameter Tuning

Regarding hyperparameter tuning, a grid search will be included in the cross-validation pipeline to reduce overfitting and improve model performance.

The grid search is used to determine which treatment (over/undersample) is best for the imbalance in our dataset. Using the 'balance_classes' function to determine if the model generates better performance metrics when over/undersampling the data. To define the best sampling strategy within the grid search, the criteria chosen is 'max_balance_after_size', which refers to the ratio change of the minority class (1< for undersampling, 1 for no sampling, >1 for oversampling). The grid search is also used to tune other attributes of the best performing algorithm in the AutoML stage, such as 'max depth' and 'sample rate' for tree-based algorithms. The max depth relates to the maximum vertical splits of a tree. By limiting the amount that the trees can grow the model tends to increase its generalization which decreases the chance of overfitting the training data. However, reducing the max depth by too much can cause the model to become too generalized and can negatively affect the performance. The sample rate relates to the proportion of rows that are sampled. For larger datasets with low positive events larger sample rates tend to have better performance. When reducing the dataset size through sampling the model generalization is also increased, so the optimal rate is tested within the grid search.

4. RESULTS AND DISCUSSION

As seen from the methodology section, the optimal evaluation metric to be used for this problem given the severe imbalance of the dataset is the area under the precision-recall curve. The first calculation is understanding what the actual baseline PR-AUC is for this dataset. To calculate this the formula is the proportion of positive examples to the total number of samples:

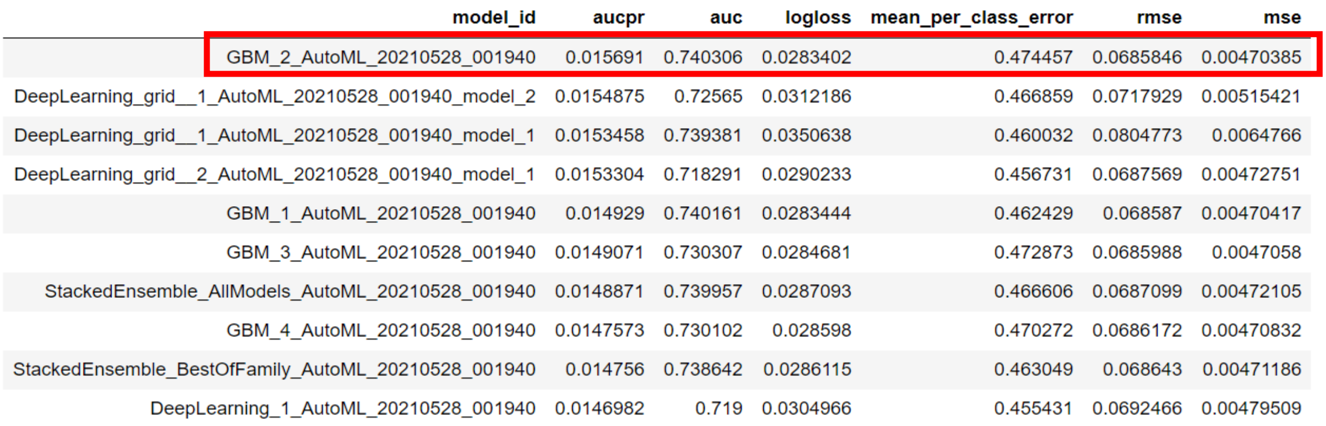
$$Train\ Baseline\ AUCPR = \frac{N.Positive\ examples\ (P)}{Total\ N.of\ Training\ Data\ (N)} = \frac{4{,}220}{723{,}030} = 0.005837$$

Likewise, for our testing sample, the baseline PR-AUC is given by;

$$Test\ Baseline\ AUCPR = \frac{572}{120{,}450} = 0.0047489$$

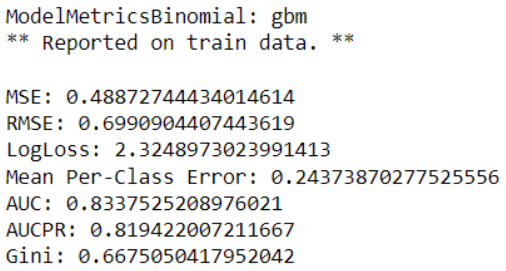
Another key metric that is also used for evaluating the performance of the models with severe imbalance is the 'Fbeta'-score which similarly to the PR-AUC calculates the relationship between precision and recall, the h2o module has a specific f0.5-score which score places more importance on false positives than false negatives.

Figure 16 shows the summary of the top 10 models run within the AutoML on the validation set.



model_id	aucpr	auc	logloss	mean_per_class_error	rmse	mse
GBM_2_AutoML_20210528_001940	0.015691	0.740306	0.0283402	0.474457	0.0685846	0.00470385
DeepLearning_grid__1_AutoML_20210528_001940_model_2	0.0154875	0.72565	0.0312186	0.466859	0.0717929	0.00515421
DeepLearning_grid__1_AutoML_20210528_001940_model_1	0.0153458	0.739381	0.0350638	0.460032	0.0804773	0.0064766
DeepLearning_grid__2_AutoML_20210528_001940_model_1	0.0153304	0.718291	0.0290233	0.456731	0.0687569	0.00472751
GBM_1_AutoML_20210528_001940	0.014929	0.740161	0.0283444	0.462429	0.068587	0.00470417
GBM_3_AutoML_20210528_001940	0.0149071	0.730307	0.0284681	0.472873	0.0685988	0.0047058
StackedEnsemble_AllModels_AutoML_20210528_001940	0.0148871	0.739957	0.0287093	0.466606	0.0687099	0.00472105
GBM_4_AutoML_20210528_001940	0.0147573	0.730102	0.028598	0.470272	0.0686172	0.00470832
StackedEnsemble_BestOfFamily_AutoML_20210528_001940	0.014756	0.738642	0.0286115	0.463049	0.068643	0.00471186
DeepLearning_1_AutoML_20210528_001940	0.0146982	0.719	0.0304966	0.455431	0.0692466	0.00479509

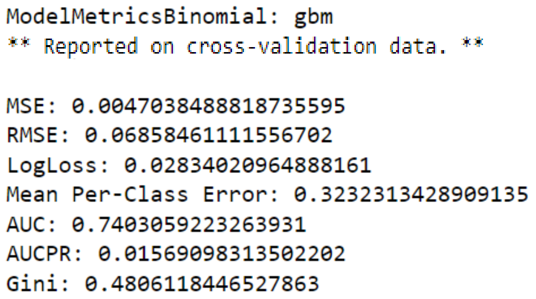
Figure 16: AutoML Leader board



```
ModelMetricsBinomial: gbm
** Reported on train data. **

MSE: 0.48872744434014614
RMSE: 0.6990904407443619
LogLoss: 2.3248973023991413
Mean Per-Class Error: 0.24373870277525556
AUC: 0.8337525208976021
AUCPR: 0.819422007211667
Gini: 0.6675050417952042
```

Figure 17: AutoML best model train set results



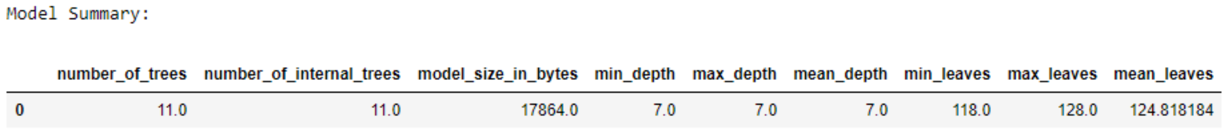
```
ModelMetricsBinomial: gbm
** Reported on cross-validation data. **

MSE: 0.0047038488818735595
RMSE: 0.06858461111556702
LogLoss: 0.02834020964888161
Mean Per-Class Error: 0.3232313428909135
AUC: 0.7403059223263931
AUCPR: 0.01569098313502202
Gini: 0.4806118446527863
```

Figure 18: AutoML best model cross validation results

The results show that overall the best performing estimator on the training set with the problem is the Gradient Boosting Machine. To check for overfitting, the results of cross-validation are also analysed to determine if there is a large discrepancy between train set metrics and cross-validation metrics. With cross-validation, sample sizes of each fold can vary, and therefore a baseline PR-AUC is not calculated for each fold, however, it is assumed that the proportion of positive target events is similar across all folds, and therefore the baseline PR-AUC for the training will also be considered for the cross-validation.

The optimal model from the AutoML function is described as following

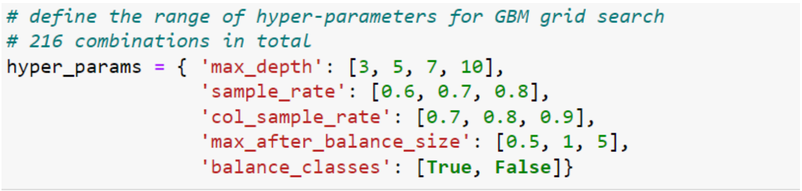


Model Summary:

	number_of_trees	number_of_internal_trees	model_size_in_bytes	min_depth	max_depth	mean_depth	min_leaves	max_leaves	mean_leaves
0	11.0	11.0	17864.0	7.0	7.0	7.0	118.0	128.0	124.818184

Figure 19: AutoML Selected model characteristics

Using the information obtained from the optimal model summary, there is a better understanding of which values should be used in the grid search when fine-tuning the model. The Figure 20 shows the hyperparameters selected in the grid search for the gradient boosting machine estimator of the h2o module. 216 combinations were selected to be tested. Specifically the ‘max dept’ levels to reduce computation and reduce potential overfitting, ’column sample rate’ and ‘sample rate’ which defines the ratio of columns and rows, this is mainly used for large datasets and finally the balancing of classes to deal with the severe imbalance of the dataset.



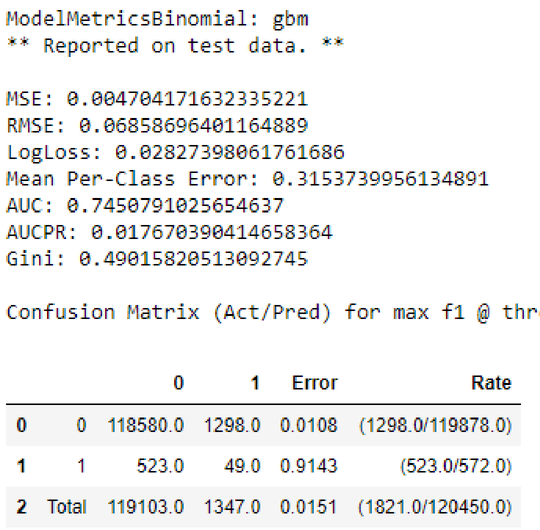
```
# define the range of hyper-parameters for GBM grid search
# 216 combinations in total
hyper_params = { 'max_depth': [3, 5, 7, 10],
                 'sample_rate': [0.6, 0.7, 0.8],
                 'col_sample_rate': [0.7, 0.8, 0.9],
                 'max_after_balance_size': [0.5, 1, 5],
                 'balance_classes': [True, False]}
```

Figure 20: GBM grid search hyperparameters

The optimal model based on the grid search, improved the PR-AUC of both the cross-validation and the test set, as seen in the figure below.

Sample	Baseline (PR-AUC)	Best Model (PR-AUC)	F-0.5 Score
Cross-Validation	0.005	0.0169	0.042614
Test-Set	0.004	0.0176	0.042716

Figure 21: Baseline to actual comparison table



```
ModelMetricsBinomial: gbm
** Reported on test data. **

MSE: 0.004704171632335221
RMSE: 0.06858696401164889
LogLoss: 0.02827398061761686
Mean Per-Class Error: 0.3153739956134891
AUC: 0.7450791025654637
AUCPR: 0.017670390414658364
Gini: 0.49015820513092745

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.03168617395264105:
```

		0	1	Error	Rate
0	0	118580.0	1298.0	0.0108	(1298.0/119878.0)
1	1	523.0	49.0	0.9143	(523.0/572.0)
2	Total	119103.0	1347.0	0.0151	(1821.0/120450.0)

Figure 22: Grid search results on test data

The output from the grid search shows an improvement in all metrics against the test set compared with the AutoML run before the hyperparameter tuning. The best performing model, did not balance classes which means for this particular case that over or under sampling did not improve the chosen metrics.

Comparing the above figures, it is clear that the models can improve on the baseline PR-AUC, however analysing the confusion matrix there are still a significant amount of false positives and false negatives, which is a practical approach that is not yet sufficient to use as a decision-making tool to effectively allocate resources since the potential cost of a false positive is very high. Nevertheless, it is important to consider that the false positive/ negative rate using this methodology is significantly lower compared to the methodology used in a previous paper using the same data sources. This suggests that this method of running machine learning algorithms for specific event types rather than all emergency events can lead to better predictions. Since the model has shown that it is considered better than the baseline, the variable importance tool can give some further insights into the features that are considered to have the most information gained towards predicting the target variable.

After a model is selected based on its F0.5-score and the PR-AUC against the test dataset, the feature importance ranking of the model is plotted to understand which features have the most predictive power. This allows for some explanation of the sporadic nature of urban fires in Lisbon.

4.1. Variable Importance

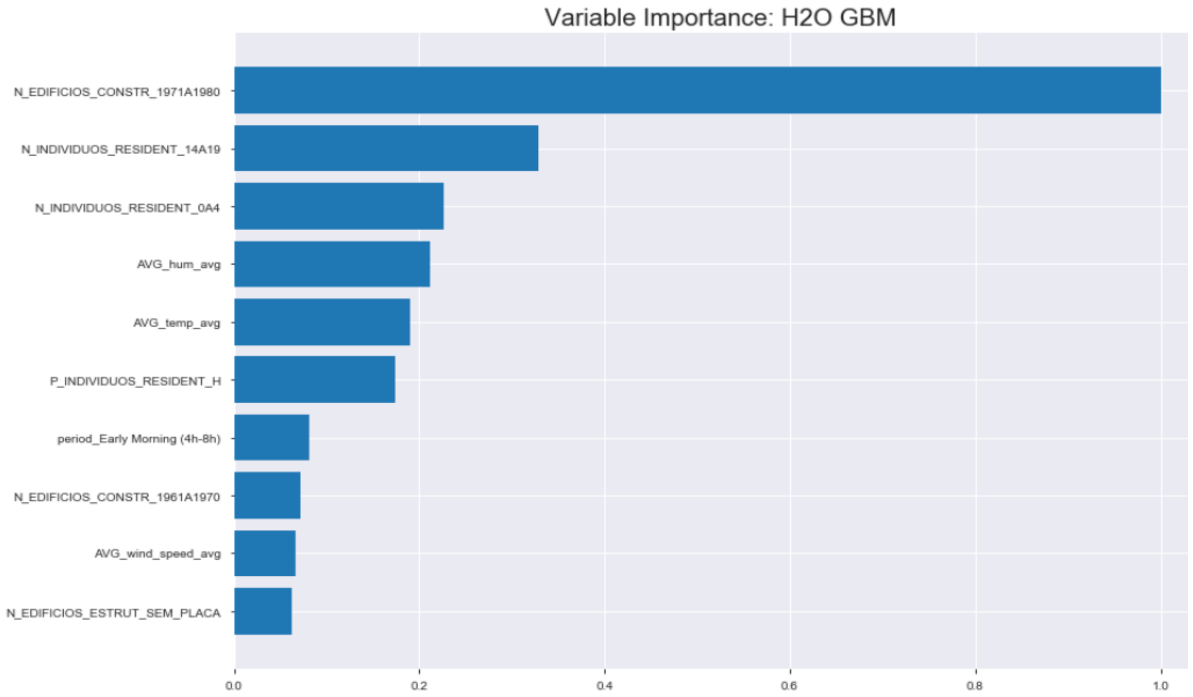


Figure 23: Top 10 variables ranked by variable importance

Considering the variable importance, we can see that all subsections of data appear as important variables. If only the census data (population & buildings) appeared within the top variables then we could conclude that the model is only predicting hotspots. Since the model also includes climate data such as temperature and humidity this means that there is some sort of correlation between these indicators and the fires. Another key variable is the time variable 'period_early_morning' because this also puts the model working in terms of time frame. Since the variable importance only shows how important features are, it does not show the relationship between the feature and the target variable (i.e if it contributes positively or negatively.) Another key takeaway from the variable importance plot is the building characteristics. The highest-ranked feature is the number of buildings between 1970 and 1980 and the 10th ranked feature is Nº of buildings of masonry walls without steel structure. This information alone can be used in decision-making for building inspections. Prioritizing building inspections to structures that contain characteristics represented in the variable importance graph could lead to more preventive actions to reduce the risk of future emergency events. Understanding which areas or demographics are most likely to be involved in fires can also help firefighters educate people on how to prevent emergency events from happening.

Analysing the Variable importance of some key takeaways can help not only predict future fires but also raises awareness on which factors might make a location more prone to fires than others. Once relevant factors have been identified then, the model can be used as a proposal for an IoT device that allows getting more frequent updates on these figures.

4.2. OUTPUT

The final stage of the model is predicting future events for RSB to use in their resource planning. The model is created such that a user can upload a weather forecast for any future period (day, week, month), provided the file is in the same format as the IPMA weather log, and the model will predict probabilities of a positive event per location per 4-hour block. The users can then depending on available resources select the top n locations per 4-hour block, and assign first-response vehicles to the area, so that in the case that an emergency event takes place, they can respond within a 5-minute window.

5. CONCLUSIONS

To summarize the project, many studies and articles were reviewed to determine the best approach to tackle the issue of predicting emergency events. The overall goal of this project was use skills learnt during the masters to build on top of a similar project that was undertaken as part of a masters project and to try and use the same data sources and data treatments along with state of the art algorithms and functions to develop a machine learning model which can predict individual emergency events within a specific time frame for each of the defined subsections of the city of Lisbon. Whilst being able to be re-trained and scalable with more data sources and different areas. Despite most of the approaches taken is based on the literature review, some changes arose from discussions with the Urban analytics lab at NOVA IMS, which is in close contact with the council of Lisbon, to give both theoretical and practical methods to solving the problem at hand. In terms of the modelling section, various data treatments and algorithms were tested and selected based on appropriate evaluation measures. The model itself is not yet sufficient to be used as the only decision-making tool, however, the groundwork has been developed for the model to constantly improve as more data is fed to it, both event data, but also new predictive variables can be added to the model, as these are available. Overall, the lessons learnt during the masters were applied successfully and the project improved from the previous study conducted using skills learned throughout the Information Management masters and applications of state of the art technologies and methodologies and has set up for future works in the emergency prediction industry.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

6.1. LIMITATIONS

6.1.1. Weather Data Limitations

A potential limitation for this project is the reliability of the unseen data is useful to make predictions. The data used to train the model was actual weather data collected from 3 separate weather stations around Lisbon however, the weather data that will be used to make predictions will be forecasted weather indicators. These indicators are not 100% accurate and therefore will have some error, this error will then carry on into the model and potentially affect the prediction.

In terms of the geospatial interpolation of the weather stations to the geographical locations, there are also some limitations. Using the Euclidean distance between location and weather station means that topography is ignored in general, and taking into consideration that Lisbon is a hilly city, not using topography in spatial interpolation might not give an actual representation of the weather indicators. A better spatial interpolation can be used to get more accurate climate data for each census block. Another way to improve the accuracy of the weather data used for training is to increase the number of stations that collect climate data. Currently in Lisbon, more IoT devices are being set up across the city to collect climate data. The data from these can be used moving forward to get more accurate data sources for training the model.

6.1.2. Census Limitations

The main limitation of this model is the census data. Since the data is static and has been collected 2 years before the first event of the historical data, it doesn’t take into account changes over time, and therefore can be outdated. Given that the census survey is only collected once every ten years the next available survey will be conducted in 2021, however, according to the Portuguese national institute for statistics, the census blocks in Lisbon will be altered. Specifically, more parishes have been created and the borders have changed, so some districts will have changed their associated parish. This will require the user to recreate the spatial interpolation for emergency events and weather station allocation to use the historical data combined with the new data available.

6.2. FUTURE WORKS

As per most supervised machine learning problems, the larger the observation size, the better the model predicts the target variable. Therefore, the best approach to improving the output of the model is to keep adding event data and train the model regularly. By doing so, the model will have more examples to learn from and will in theory generate better predictions. Since the computation time of the entire model running with a single node is about 2.5 hours, running the model weekly or monthly with a cluster of nodes will not be operationally expensive, even though the dataset is larger.

Additionally, other data sources can be added to make the model more robust. As seen in the literature review, using economic data such as building valuation and tax amounts are also seen as useful in predicting emergency events. Another potential data source that could be added to the report, is the population in each area. The department for urban analytics is developing a dataset that compiles cellular network data to count the number of people in each location per time frame. If these data sources are at the same granularity level (both spatial and temporal) as the event data, then this can be added to make the model more robust. Despite this not being captured in the census survey this type of information will potentially add another perspective to the data (financial & population movement) and can lead to better performance or better indication of which drivers contain higher importance in predicting each type of emergency event

6.2.1. Predicting other emergency events.

In a previous study, using the same data sources, a model was created to predict all emergency events that RSB responded to. In comparison, this model only predicts fires, however, it is built so users of this model and easily change the model so it predicts any other emergency type that RSB responds to. For example, as seen from the variable importance the current model places greater importance on the season being summer. It can be that this model predicts better fires that take place in the summer than in the winter (where there are fewer fires), however, in the winter there are more floods and therefore one can easily switch the target variable to floods. This also allows to take better actionable insights into what are contributing factors for each positive emergency event and has the potential to lead RSB in taking a more preventive approach in tackling emergency events and improving operational efficiency.

7. BIBLIOGRAPHY

Akandea, A., Cabral, P., Gome, P., & Casteleyn, S. (2019). The Lisbon Ranking for Smart Sustainable Cities in Europe. *Sustainable Cities and Society, vol. 44*, 475–487.

Albino, V., Berardi, U., & Dangelico, R. M. (2015). Smart Cities: Definitions, Dimensions, Performance, and Initiatives. *Journal of Uban Technology*, 3-10.

Aniruddha Bhandari. (2020). Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization. *Analytics Vidhya*.

Bao, J., Liu, P., & Ukkusuri, S. V. (2019). *A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data.* New York City: Elsevier.

Brown, M. S. (2015, July 29). *What IT needs to know about the data Mining process.* Retrieved from Forbes: https://www.forbes.com/sites/metabrown/2015/07/29/what-it-needs-to-know-about-the-data-mining-process/?sh=788ff743515f

Brownlee, J. (2020, August 3). *A Gentle Introduction to k-fold Cross-Validation.* Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/k-fold-cross-validation/#:~:text=Cross%2Dvalidation%20is%20primarily%20used,the%20training%20of%20the%20model.

Brownlee, J. (2020, Feb). *A Gentle Introduction to the Fbeta-Measure for Machine Learning*. Retrieved from Machine Learining Mastery: https://machinelearningmastery.com/fbeta-measure-for-machine-learning/

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: a new. *Neurocomputing*.

Cargaliu, A., Bo, C. D., & Nijkamp, P. (2011). Smart Cities in Europe. *Journal of Urban Technology*, 62-65.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., & Reinartz, T. (2000). CRISP-DM 1.0. *SPSS Inc.*

Cipriano, S. (2012, 09 19). *Historia dos Bombeiros.* Retrieved from Bombeiro.pt

Combining Over-Sampling and Under-Sampling Techniques for Imbalance Dataset. (n.d.). *Publication:ICMLC 2017: Proceedings of the 9th International Conference on Machine Learning and Computing*.

Cortez, P., & Morais, A. (2007). *A Data Mining Approach to Predict Forest Fires.* Minho, Portugal: Associação Portuguesa para a Inteligência Artificial (APPIA).

Devisschere, P. K. (2021, 09). *H2O in practice: a Data Scientist feedback*. Retrieved from Adatlas: https://www.adaltas.com/en/2021/09/29/h2o-automl-data-scientist-feedback/

Driving-radius map application. (2021, January 2). Retrieved from Oalley: https://www.oalley.net/

Eidam, E. (2016). *Cincinnati Predictive Analytics Project Takes Aim at Emergency Medical Services.* Cincinnati: Government Technology.

Ekelund, S. (2017, March). *Precision-recall curves – what are they and how are they used?* Retrieved from acutecaretesting: https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

European Commission. (2018). *Smart Cities Initiatives.* Retrieved from European Commission: https://ec.europa.eu/info/eu-regional-and-urban-development/topics/cities-and-urban-development/city-initiatives/smart-cities_en

H2o AI Platform. (2021, 03 1). Retrieved from H20 AI: https://www.h2o.ai/hybrid-cloud/

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & G., B. (2017). Learning from class imbalanced data: Review of methods and applications. *Expert Systems with Applications 73*, 220-239.

Hand, D., Mannila, H., & Smyth, P. (2001). Principles of Data Mining. *MIT Press*.

Heaton, B. (2015). New York City Fights Fire with Data. *Justice and Public Safety*.

History of Firefighting. (2016, 01). Retrieved from Firefighter Foundation: https://www.firefighterfoundation.org.uk/history/#:~:text=The%20history%20of%20the%20firefighter,in%20the%203rd%20Century.&text=In%20AD%2060%20Emperor%20Nero,also%20as%20a%20police%20force

Kapkar, B. (2020, May). *Which Machine Learning requires Feature Scaling(Standardization and Normalization)?* Retrieved from Kaggle: https://www.kaggle.com/getting-started/159643

LeDell, E., & Poirier, S. (2020, July). *H2O AutoML: Scalable Automatic Machine Learning.* Retrieved from 7th ICML Workshop on Automated Machine Learning (AutoML): https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf.

Ma, W., Feng, Z., Cheng, h., Chen, S., & Wang, F. (2020). Identifying Forest Fire Driving Factors and Related. *Forests*.

Madaio, M., Chen, S.-T., Haimson, O. L., Zhang, W., Cheng, X., Hinds-Aldrich, M., . . . Dilkina, B. (2016). *Firebird: Predicting Fire Risk and.* San Francisco: KDD.

Marsal-Llacuna, M. C.-L.-F. (2015). Lessons in urban Monitoring taken from sustainable and livable cities to better address the Smart Cities Initiative. *Technological Forecasting and Social Change*, 611-622.

Mukhopadhyay, A., Pettet, G., Vazirizade, S., Lu, D., Jaimes, H., Alex, B., . . . Dubey, A. (2020). A Review of Emergency Incident Prediction,. *arXivLabs*, 2-10.

Mukhopadhyay, A., Pettety, G., Vazirizadey, S., & Lu, D. (2020). *A Review of Emergency Incident Prediction,.* arXiv:.

Mukhopadhyay, A., Wang, K., Perrault, A., Kochenderfer, M., & Tambe, M. (2020). Robust Spatial-Temporal Incident Prediction. *Centre for Research on Computation and Society*.

Narkhede, S. (2018, June 26). *Understanding AUC - ROC Curve*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

NECPortugal. (2020, 05). *NEC cooperates with the Lisbon City Council in the development of the Lisboa.24 App.* Retrieved from NEC Portugal: https://uk.nec.com/en_GB/press/202005/20200514_01.html

Nutthaporn Junsomboon, T. P. (2017). Combining Over-Sampling and Under-Sampling Techniques for Imbalance Dataset. *ICMLC 2017: Proceedings of the 9th International Conference on Machine Learning and Computing*, 243–247.

Pedregosa, F., Varoquaux, G., Gramfort, A., & Michel, V. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 2825-2830. Retrieved from Scikit-Learn: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html#:~:text=The%20precision%2Drecall%20curve%20shows,a%20low%20false%20negative%20rate.

Priano, F. H., & Guerra, C. F. (2014). A framework for measuring smart cities. *International Conference on Digital Government*, 44-45.

Roberts, D., & Boyce, M. S. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogentic structure. *Ecography*, 913-929.

Roman, J. (2014). Data Driven Information and the brave new world of smart firefighting. *National Fire Protection Association*, 1-12.

Romero, T., Barnes, Z., & Cipollone, F. (2016). *Predicting Emergency Incidents in San Diego.* Stanford.

Roy, B. (2020, April 6). *All about Feature Scaling*. Retrieved from Towards Data Science: https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

Saurav, K. (2016). Introduction to Feature Selection methods. *Analytics Vidhya*.

SharingCities.eu. (2020). *Lisbon's Profile*. Retrieved from SharingCities.eu: http://www.sharingcities.eu/sharingcities/city-profiles/lisbon

Srikanth, C. S., Rayudu, T. B., Radhika, J., & Anitha, R. (2019). Smart Waste Management using Internet of Things (IOT). *Journal of Innovative Technology and Exploring Engineering*, 2518-2522.

Teixeira, L. B. (2018). *Optimization of firefighter response with.* Lisbon: NOVA IMS.

Walia, B. S., Hu, Q., Chen, J., Lee, J., Kuo, N., & Narang, P. (2018). A Dynamic Pipeline for Spatio-Temporal Fire Risk Prediction. *Association for Computing Machinery.*

William, D. E., & John, S. (2018). Forces of change: Smart cities. *Deloitte Insights*, 4.

Xin, J. a. (2013). Fire Risk Analysis of Residential Buildings Based on Scenario Clusters and Its Application in Fire Risk Management. *Fire Safety Journal*, 62, 72-78.

Xiong, Y. (2017). *Machine Learning Wildfire Prediction.* California, USA: NoiselabUSCD.

8. APPENDIX

8.1. Appendix 1: Table showing Breakdown of daily periods

Order	Period
1	Late Night (0h-4h)
2	Early Morning (4h-8h)
3	Morning (8h-12h)
4	Noon (12h-16h)
5	Evening (16h-20h)
6	Night (20h-0h)

8.2. Appendix 2: Pair correlations and the removed variables:

Variable 1	Variable 2	Correlation	Eliminated
N_INDIVIDUOS_RESIDENT_20A64	N_INDIVIDUOS_RESIDENT_25A64	0,9987	1
N_INDIVIDUOS_RESIDENT_M_20A64	N_INDIVIDUOS_RESIDENT_M_25A64	0,9986	1
N_INDIVIDUOS_RESIDENT_H_20A64	N_INDIVIDUOS_RESIDENT_H_25A64	0,9983	1
N_IND_RESID_EMPREGADOS	N_IND_RESID_EMPREG_SECT_TERC	0,9977	2
N_INDIVIDUOS_RESIDENT	N_INDIVIDUOS_RESIDENT_M	0,9959	1
N_INDIVIDUOS_PRESENT	N_INDIVIDUOS_PRESENT_M	0,9958	1
N_INDIVIDUOS_RESIDENT_25A64	N_INDIVIDUOS_RESIDENT_M_25A64	0,9951	1
N_INDIVIDUOS_RESIDENT_65	N_INDIVIDUOS_RESIDENT_M_65	0,9941	2
N_INDIVIDUOS_RESIDENT_14A19	N_INDIVIDUOS_RESIDENT_15A19	0,9939	2
N_INDIVIDUOS_RESIDENT_65	N_IND_RESID_PENS_REFORM	0,9925	1
N_INDIVIDUOS_PRESENT_M	N_INDIVIDUOS_RESIDENT_M	0,9920	1
N_INDIVIDUOS_RESIDENT_M_14A19	N_INDIVIDUOS_RESIDENT_M_15A19	0,9907	2
N_INDIVIDUOS_PRESENT_H	N_INDIVIDUOS_RESIDENT_H	0,9891	1
N_INDIVIDUOS_RESIDENT_H_14A19	N_INDIVIDUOS_RESIDENT_H_15A19	0,9887	1
N_INDIVIDUOS_RESIDENT_H	N_INDIVIDUOS_RESIDENT_H_25A64	0,9879	2
N_INDIVIDUOS_RESIDENT_M_25A64	N_IND_RESID_EMPREGADOS	0,9856	2
N_INDIVIDUOS_RESIDENT_H	N_INDIVIDUOS_RESIDENT_M_25A64	0,9842	1
N_INDIVIDUOS_RESIDENT_M	N_INDIVIDUOS_RESIDENT_M_25A64	0,9832	2
N_INDIVIDUOS_RESIDENT_20A24	N_INDIVIDUOS_RESIDENT_H_20A24	0,9744	2
N_INDIVIDUOS_RESIDENT_20A24	N_INDIVIDUOS_RESIDENT_M_20A24	0,9735	1
N_INDIVIDUOS_RESIDENT_M	N_IND_RESID_SEM_ACT_ECON	0,9730	1
N_INDIVIDUOS_RESIDENT_H_65	N_IND_RESID_PENS_REFORM	0,9721	1
N_INDIVIDUOS_RESIDENT_10A13	N_IND_RESIDENT_FENSINO_2BAS	0,9713	1
N_EDIFICIOS_1OU2_PISOS	N_EDIFICIOS_CLASSICOS_1OU2	0,9711	1

N_INDIVIDUOS_RESIDENT_0A4	N_INDIVIDUOS_RESIDENT_H_0A4	0,9682	2
N_IND_RESID_PENS_REFORM	N_IND_RESID_SEM_ACT_ECON	0,9681	2
N_INDIVIDUOS_RESIDENT_0A4	N_INDIVIDUOS_RESIDENT_M_0A4	0,9670	1
N_INDIVIDUOS_RESIDENT_5A9	N_INDIVIDUOS_RESIDENT_H_5A9	0,9614	1
N_INDIVIDUOS_RESIDENT_14A19	N_IND_RESID_ESTUD_MUN_RESID	0,9609	1
N_IND_RESIDENT_ENSINCOMP_1BAS	N_IND_RESIDENT_ENSINCOMP_2BAS	0,9568	2
N_EDIFICIOS_CLASSICOS_1OU2	N_EDIFICIOS_CLASSICOS_EMBANDA	0,9464	1
N_INDIVIDUOS_RESIDENT_H_10A13	N_IND_RESIDENT_FENSINO_2BAS	0,9406	1
N_INDIVIDUOS_RESIDENT_H_5A9	N_IND_RESIDENT_FENSINO_1BAS	0,9390	2
N_IND_RESIDENT_FENSINO_SEC	N_IND_RESID_ESTUD_MUN_RESID	0,9363	2
N_INDIVIDUOS_RESIDENT_M_10A13	N_IND_RESIDENT_FENSINO_2BAS	0,9290	2
N_IND_RESIDENT_ENSINCOMP_SUP	N_IND_RESIDENT_FENSINO_SUP	0,9285	1
N_IND_RESIDENT_ENSINCOMP_1BAS	N_IND_RESIDENT_ENSINCOMP_3BAS	0,9177	2
N_IND_RESIDENT_ENSINCOMP_SEC	N_IND_RESIDENT_FENSINO_SUP	0,9080	2
N_IND_RESIDENT_ENSINCOMP_1BAS	N_IND_RESID_DESEMP_PROC_EMPRG	0,9078	2
N_INDIVIDUOS_RESIDENT_M_14A19	N_IND_RESIDENT_FENSINO_SEC	0,9038	2
N_INDIVIDUOS_RESIDENT_H_15A19	N_IND_RESIDENT_FENSINO_3BAS	0,9001	2
N_IND_RESIDENT_ENSINCOMP_SEC	N_IND_RESID_PENS_REFORM	0,8961	1
N_INDIVIDUOS_RESIDENT_H_15A19	N_INDIVIDUOS_RESIDENT_M_20A24	0,8827	1
N_INDIV_RESIDENT_N_LER_ESCRV	N_IND_RESIDENT_ENSINCOMP_1BAS	0,8742	1
N_INDIVIDUOS_RESIDENT_M_10A13	N_INDIVIDUOS_RESIDENT_M_20A24	0,8690	1
N_IND_RESID_TRAB_MUN_RESID	N_IND_RESID_PENS_REFORM	0,8639	1
N_ALOJAMENTOS_VAGOS	N_EDIFICIOS_CLASSICOS_3OUMAIS	0,8602	2
N_INDIVIDUOS_RESIDENT_M_0A4	N_INDIVIDUOS_RESIDENT_M_5A9	0,8477	2
N_INDIVIDUOS_RESIDENT_H_5A9	N_INDIVIDUOS_RESIDENT_M_0A4	0,8419	2
N_INDIVIDUOS_RESIDENT_M_20A24	N_IND_RESID_EMPREG_SECT_SEQ	0,8382	1
N_ALOJAMENTOS_VAGOS	N_EDIFICIOS_CLASSICOS	0,8313	2
N_INDIVIDUOS_RESIDENT_M_14A19	N_IND_RESID_EMPREG_SECT_SEQ	0,8164	1
N_IND_RESIDENT_ENSINCOMP_1BAS	N_IND_RESID_EMPREG_SECT_SEQ	0,8138	1