**Statistics Final** – Project

D00251757

Khubim Kumar Chhetri

**Student Name: Student ID:**

**Predicting Forest Fire and Temperature**

**Project Title:**

1. Background Research/ Problem Statement

A forest fire is unplanned and uncontrollable and can occur by lightning or human carelessness in forests, grasslands or shrubbery (Government of Canada, 2020). The vast majority of forest fires are human-caused, however dry climate, hot temperatures, lightning, and volcanic eruption can also lead to their occurrence (National Park Service, 2018). The past decade of climate change has only exacerbated the amount of forest fires, leading to more frequent and extreme occurences.

Wildfire agencies use many variables to indicate an imminent wildfire and the evolution of machine learning has provided us the ability to predict future events by analyzing these variables. Thus, we pose the predictive question: do certain variables allow us to determine if a forest fire has or will occur and if so, how accurate will they be?

To support our hypothesis, we used a dataset on Algerian Forest Fires from UCI (Faroudja & Izeboudjen, 2020). The dataset contains a culmination of forest fire observations and data in two regions of Algeria: the Bejaia region and the Sidi Bel-Abbes region. The timeline of this dataset is from June 2012 to September 2012. In this project, we focused on whether certain weather characteristics could predict forest fires in these regions . we used a dataset on **Algerian Forest Fires from UCI**. The dataset contains a culmination of forest fire observations and data in two regions of Algeria: the Bejaia region and the Sidi Bel-Abbes region.

The timeline of this dataset is from June 2012 to September 2012. In this project, we focused on whether certain weather features could predict forest fires in these regions using few Machine Learning algorithms.

The dataset used can be found here from the UCI Machine Learning Repository.

We analyse its are measured by taking into considering six broad categories.

|  |  |
| --- | --- |
| Research questions | Project Goals |
| Predict the possibility of fire and predict temperature? | Learn more about Forest fire happening and check its impacts over the years    To detect future fires based on certain Weather report.   1. Understand different factors affecting the Fire and check for accuracies to find out the cause?      1. Gain Hands-on experience with visualization, Hypothesis testing and the Machine learning models with accuracy score |
| Which Years are going to generate more Temperature coming years?  Predict the change in atmospheric Humidity over the coming years. |
| Check and find out if there is any seasonality in Fire changes? |

1. Research Plan/Design

I Plan to spend nearly 12 days for collection, cleaning and preparation of data. The data can be collected from various sources from worldindata- where I can get the downloadable csv file, and Google- the clear guidance and inference drawn from the data. With the documentation reports.

Once the data is collected , I can go for initial or preliminary investigation .I will plan to check for misleading , broken data that is need to be fixed or removed for further analysis. After cleaning data and integrating dataset I will compare all dataset together to analyze correctly for the changes in mobility before and during the Pandemic.

Apart from dataset, I may look for articles ,helpful information which are relevant and required for to find the result or findings of the project. Consequently once the dataset are arranged ,the data then analyse and identify for results. The data needs to be displayed and visualized with graphs, plots, hypothesis, conclusions once the statistical results are valid and clear from the analysis.

***Collect Dataset(Using a downloadable csv from worldindata)***

***Clean,Organize and prepare data(Look for Broken,missing and anomalies in dataset)***

***Integrate data for Research***

***Analyse and Elaborate data***

***Visualize data***

***Graphs,Charts and tables***

***Findings ,Solutions and Conclusions***

***Presentation/Results***

**Aim and Objectives**

The paper aims to develop a trustworthy and interpretable model that will predict the customers that will churn from a Telecom Company based on historical customer telecom data. The identification of the customers that churn will aid telecom companies in significantly reducing expenditure on customer relations.

The objectives of the research are based on the above aim and are as follows:

* To analyze the relationship and visualize patterns of customer behaviour to indicate to the telecom company if a customer is going to churn
* To suggest suitable feature engineering steps to extract the most value from the data including picking the most significant features
* To find appropriate balancing techniques to enhance the model performance on the dataset
* To compare the classification or predictive models to identify the most accurate model to determine the customers that will churn
* To understand the factors and behaviour of consumers that leads to customer attrition in the telecom industry
* To evaluate the performance of the models to identify the appropriate models

**Significance of the research**

The research is contributing to the explanation and interpretation of the prediction of various predictive models to support decision making and increase the bottom line of the company by flagging customers that are going to churn. This will help customer allocate budget and time to the customers that are likely to churn by running targeted campaigns. The sales team will be able to offer value-adds to the high-risk and high-value customers. This can help the company document the pain points faced by its customers and can ultimately help aid in fundamental policy changes that can increase the overall profit.

**Scope of the study**

Due to the limitation of the time frame in this research, the scope of the research will be limited to the below points:

* The data for the study has directly been obtained from the authorized source, and data validation will not be part of this research
* The research will include the development and evaluation of various machine learning algorithms. The latest algorithms such as Neural Networks and Deep learning will not be considered as a part of this study due to a lack of resources and time
* The study will limit the use of classification algorithms such as logistic regression, decision tree, K-nearest Neighbour as a part of interpretable models, whereas random forest, support vector machine, gradient boosting and XGBoost will be leveraged as black-box models for this study
* We will focus on models that are interpretable. If time permits, we will attempt to use other models to perform customer attrition analysis

**Research Methodology**

In this paper, we were able to identify that the telecom industry is an extremely competitive industry where customers have the free will to move across companies if they believe they are getting more value with another service provider. We also noted that based on the customer's behaviour patterns, we would have indicators to note if a customer might churn or not. Since the cost of retention is much higher than customer acquisition, it is vital to the company's survival to identify the customers likely to churn and run campaigns to retain the existing customer base. It was also observed that a reduction of customer attrition of 5% could lead to profit margins increasing from 25% to 95% (Hadden et al., 2006). In the telecom industry where the approximated annual cost of customer attrition is $ 10 billion annually (Castanedo et al., 2014), and 30% customers churn on average, there is a substantial need to perform active targeting to retain the customer base.

**Data Understanding**

There are various data sources used to predict churn in the telecom industry through the literature survey. In this research, we shall be using the IBM Watson Telecom churn data found on the Kaggle website. The telecom churn data consists of 8043 rows and 21 attributes at a customer id level. The data has a combination of numerical and categorical variables that can be used as feature variables to predict the target variable churn. Churn is indicated within the dataset as a "Yes" or a "No" indicating if a customer has churned or not churned respectively. This data presented is for the last month based on which predictions are to be made.

The given data consists of multiple factors about the customers regarding lifestyle, behaviour in a Yes or No format that can be leveraged post-processing. It is presented in a .csv format with customer attributes information as metadata.

1. Special Requirement and deliverable of the project

It is recommended to analyse the data for less than 6 months period .Also we must be aware to analyse and compare weekend data with weekdays.

The deliverables are mobility of visitors with respect to different places during pandemic .As we see the number of cases rising now –then possibility to suggest some solution if required.

**Algerian Forest Fires**

**Data set Available at:** [link text](https://archive.ics.uci.edu/ml/datasets/Algerian+Forest+Fires+Dataset++)

***Data Set Information:***

* The dataset includes 244 instances that regroup a data of two regions of **Algeria**,namely the
* **Bejaia region** located in the **northeast of Algeria** and the **Sidi Bel-abbes region** located in the **northwest of Algeria**.
* 122 instances for each region.
* The period from June 2012 to September 2012.
* The dataset includes 11 attribues and 1 output attribue (class)
* The 244 instances have been classified into **fire** (138 classes) and **not fire** (106 classes) classes.

**Attribute Information:**

**1. Date :** (DD/MM/YYYY) Day, month ('june' to 'september'), year (2012)

**Weather data observations**

**2. Temp :** temperature noon (temperature max) in Celsius degrees: 22 to 42

**3. RH :** Relative Humidity in %: 21 to 90

**4. Ws :** Wind speed in km/h: 6 to 29

**5. Rain:** total day in mm: 0 to 16.8

**FWI Components**

**6. Fine Fuel Moisture Code (FFMC) index from the FWI system:** 28.6 to 92.5

**7. Duff Moisture Code (DMC) index from the FWI system:** 1.1 to 65.9

**8. Drought Code (DC) index from the FWI system:** 7 to 220.4

**9. Initial Spread Index (ISI) index from the FWI system:** 0 to 18.5

**10. Buildup Index (BUI) index from the FWI system:** 1.1 to 68

**11. Fire Weather Index (FWI) Index:** 0 to 31.1

**12. Classes:** two classes, namely **Fire** and **not Fire**