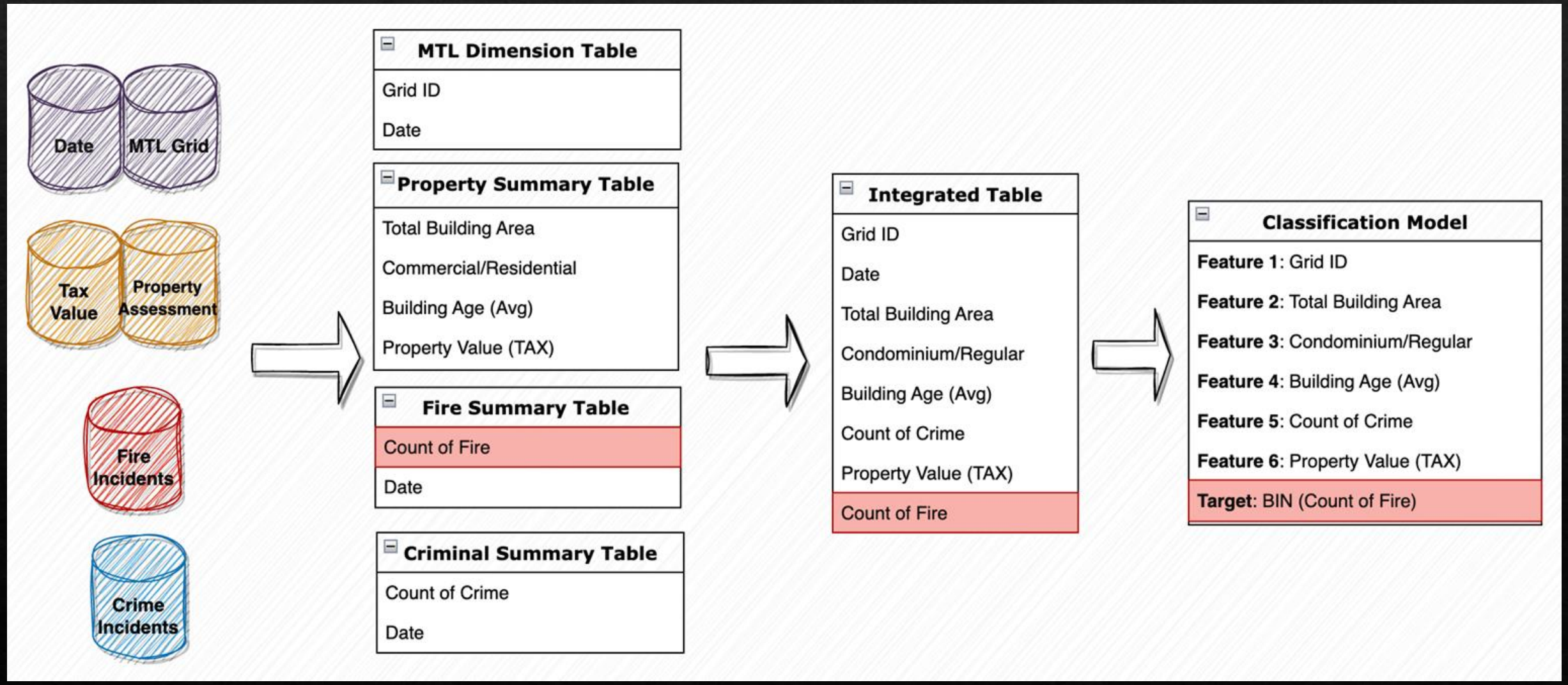
Feature and Model creation



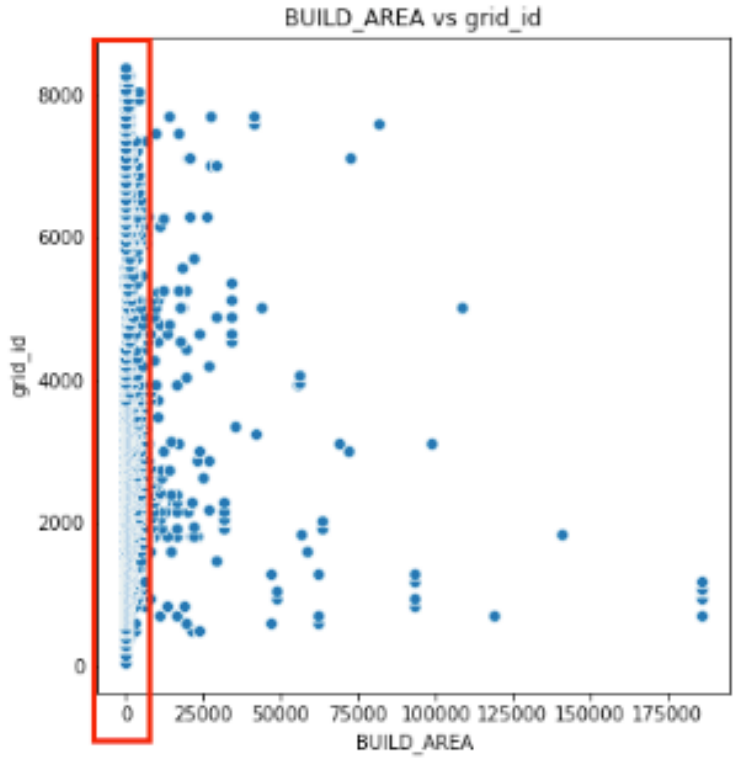
Attempt 1: Classification

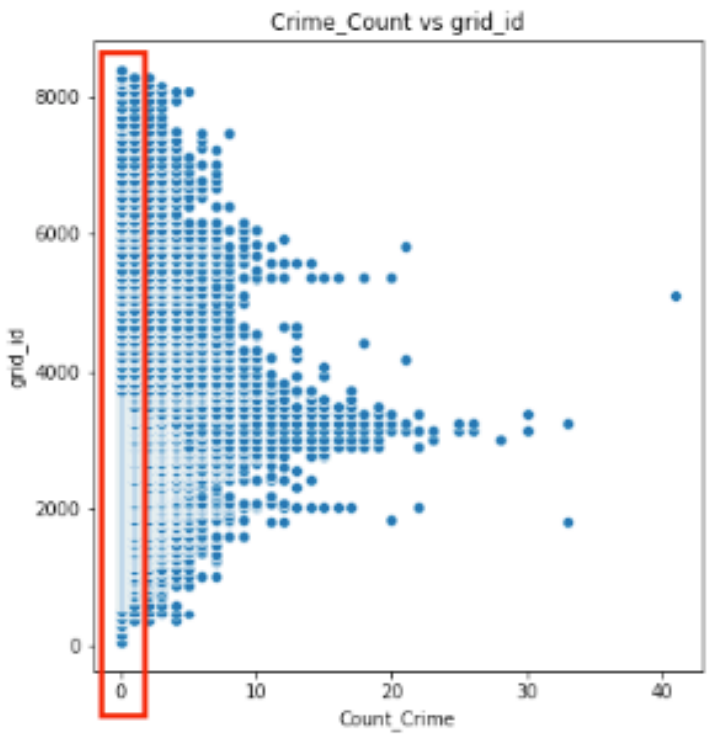
Accuracy Score with **25% Test size**:

Decision Tree: **0.85**

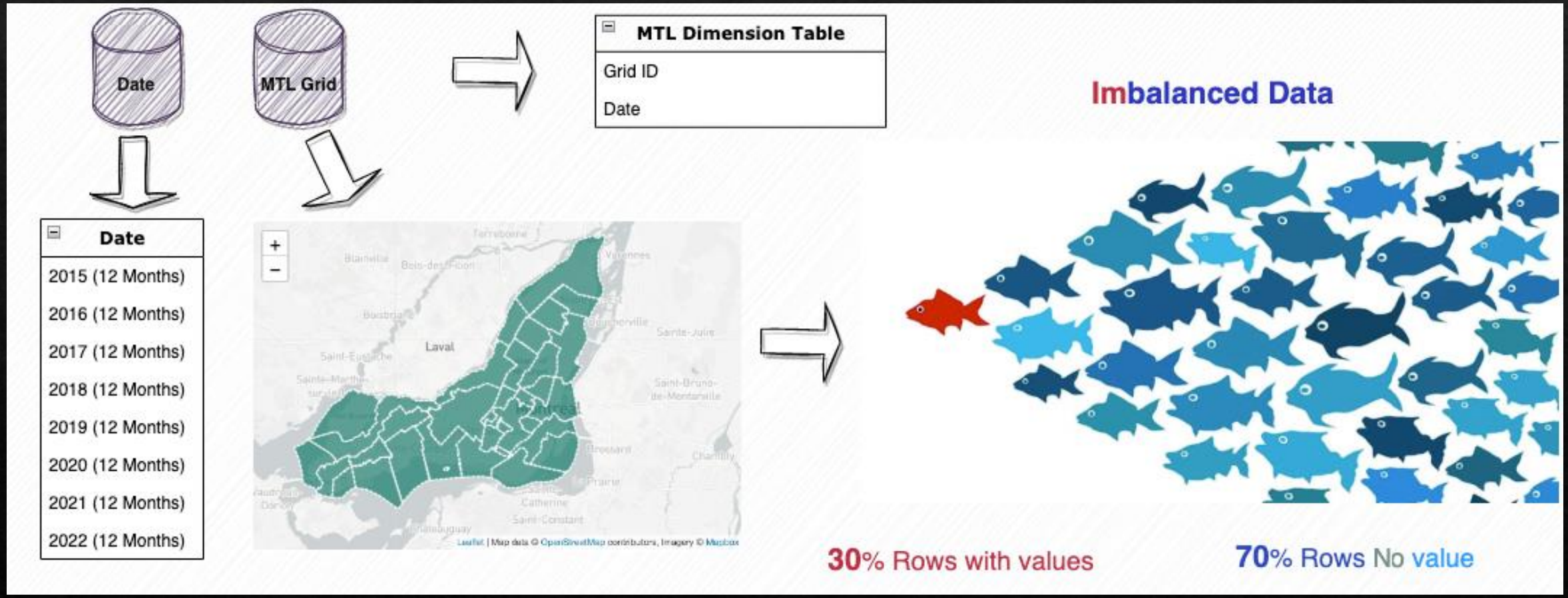
GradientBoosting: **0.81**

Random Forest: **0.83**



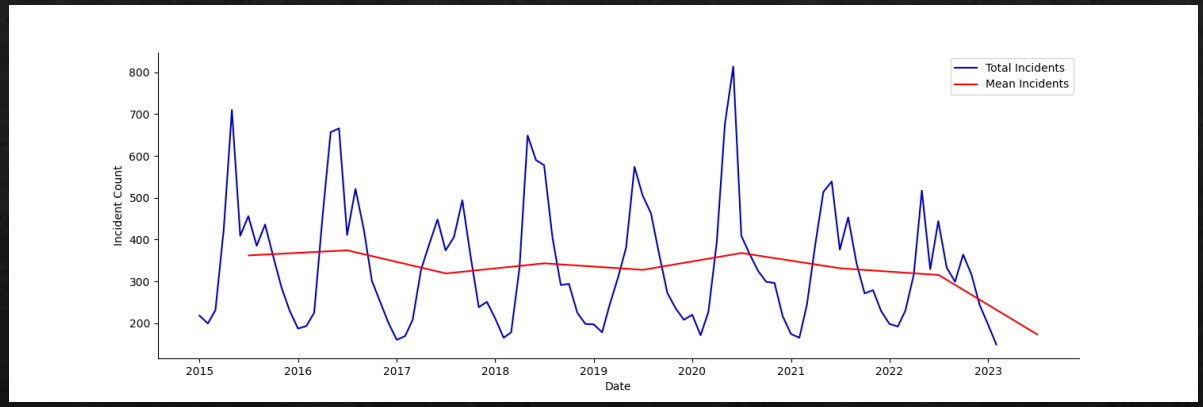


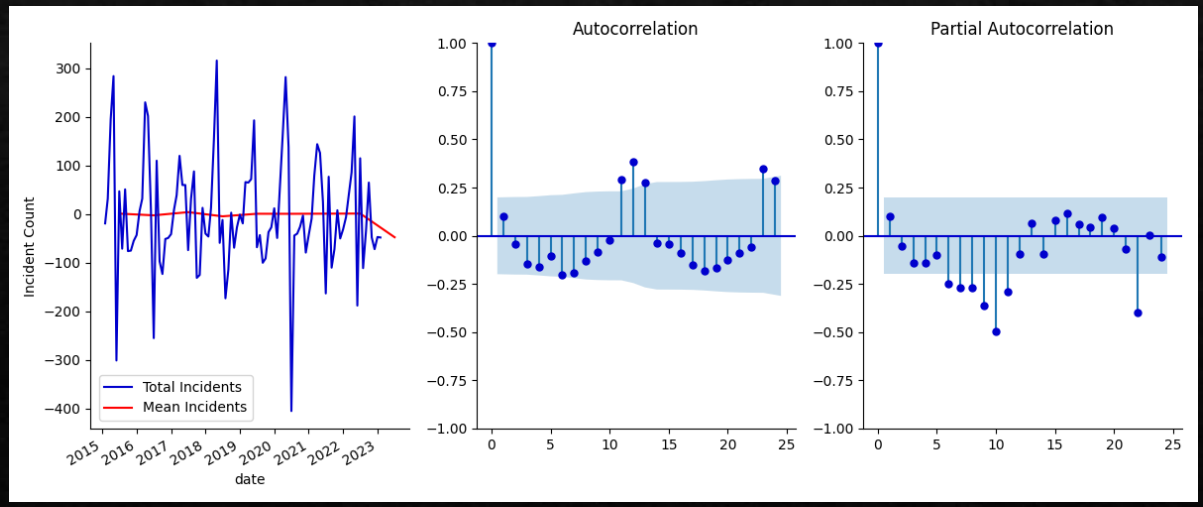
Imbalanced Data



Time Series Analysis

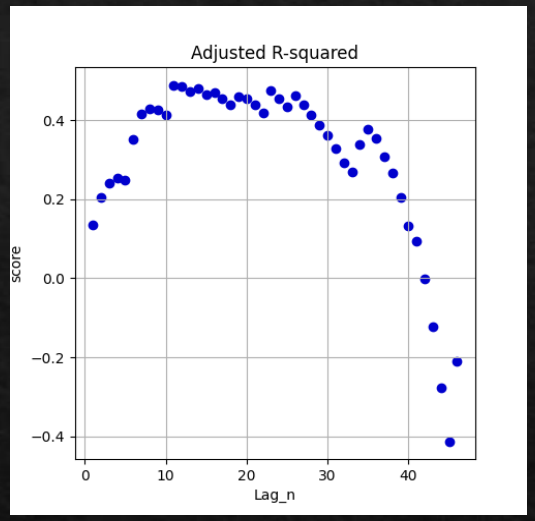
- Our goal was to try to predict monthly incidents in the whole island.
- Only real fires were consolidated in the data and aggregated by month.
- Transformation required since the data is not stationary.
- Create columns from lag_1 to lag_n and assign values by using shift() method.

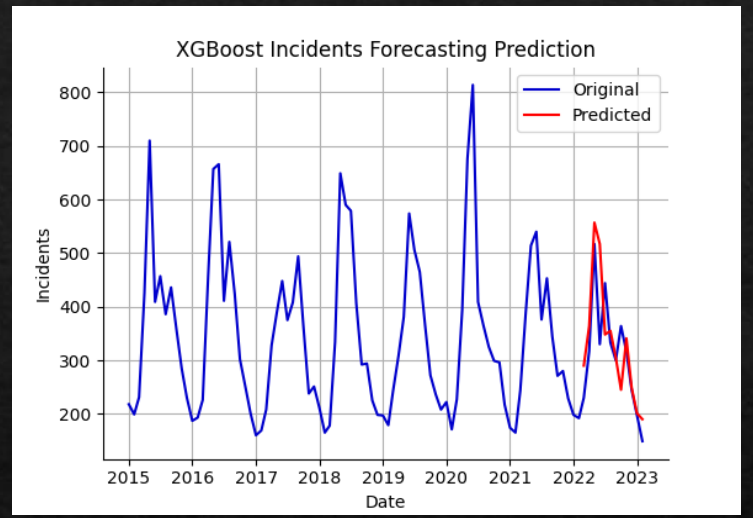


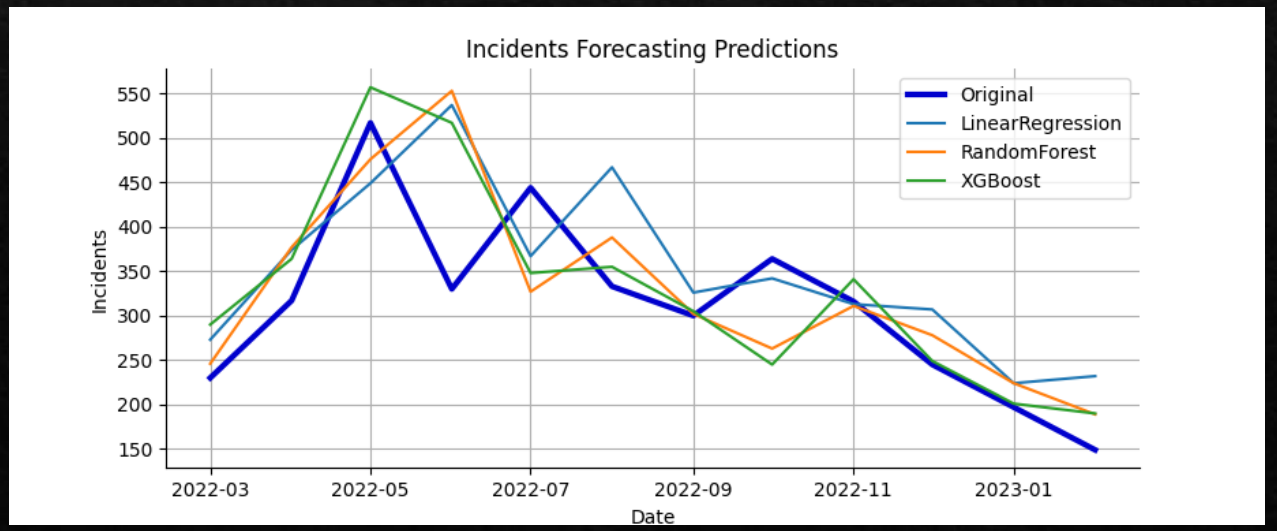


Time Series Models

- To what extent do the features (lags) contribute to prediction accuracy?
  - Adjusted R-squared
- Train test split: Testing only with the last 12 months.
- 3 Models:
  - Linear Regression: R2 = 22.1%
  - Random Forest: R2 = 14.4%
  - XGBoost: R2 = 40%







Attempt 2

- Dataset aggregation: data was aggregated monthly and grid_id basis
- Target Variable: Number of fire incidents
- Model Type: Regression
  - Algorithms tried: multiple regression, ridge/lasso regression, catboost, **XGboost**
- Best model accuracy: 0.42 %

Attempt 3

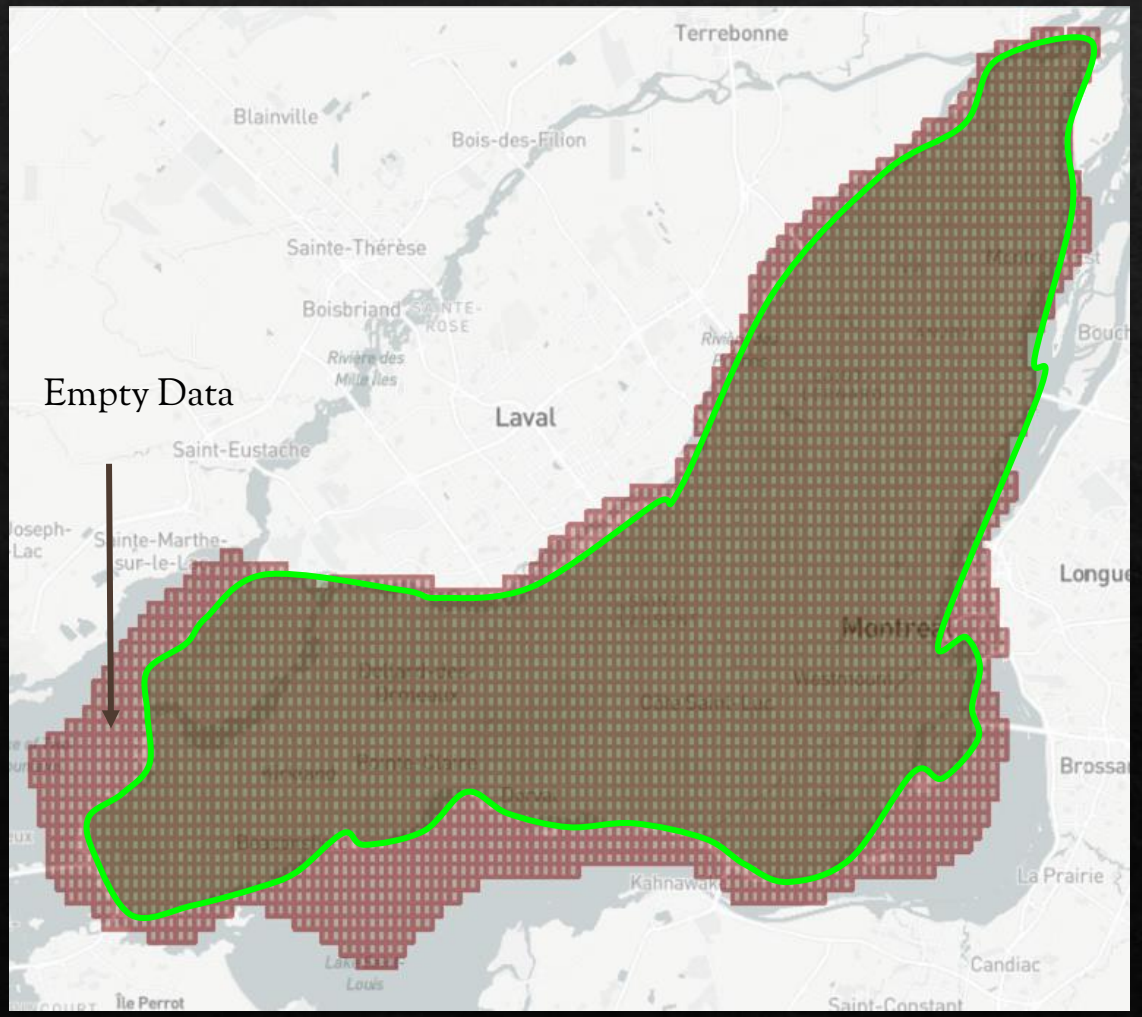
- Dataset aggregation: data was not aggregated, each fire incident had a date and was mapped to a grid_id (~900,000 data points)
- Target Variable: Type of fire incident
  - This was a categorical variable built based type of fire incident (Class 1 - building fire, Class 2 - other fire, Class 3 - any response other than fire)
- Model Type: Regression
  - Algorithms tried: logistic regression, KNN, random forest, catboost, XGboost
- Best model accuracy: >0.4 %
- Problem: Data set was unbalanced, 90% of the dataset was of class 3 incident type

Data Balancing

- The MTL Administrative database used to create the grid, covers more area than the island's land limits.
- On certain areas, the dataset was empty over a period of 12 months multiplied by 7 years, thereby exacerbating the imbalance.

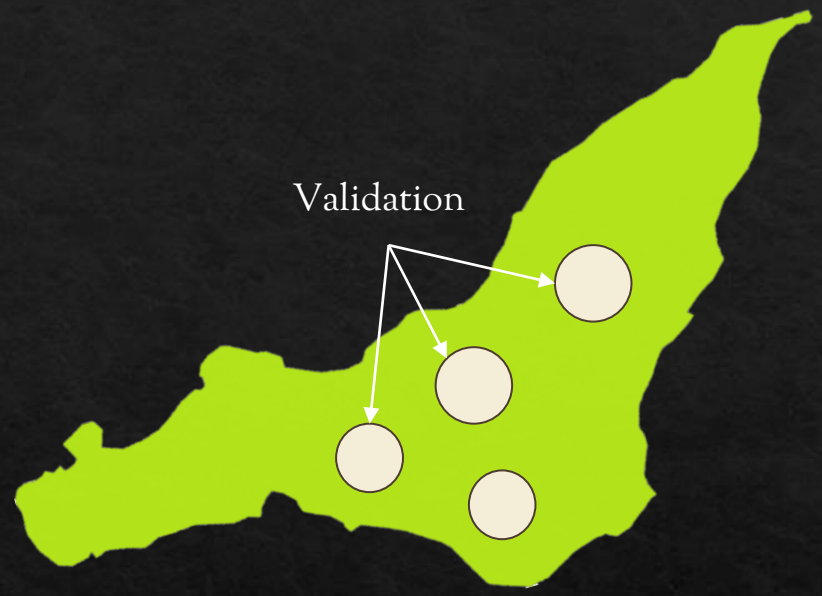
Solution:

- Eliminate the affected grid points by using the 2021 Census Subdivisions boundary file (SHP) provided by Statistics Canada.
- Increase the grid size (currently 500m)



Performance

- Use K-mean Clustering to spot hot areas
- Split the data-set Train/Test into different regions of the city:
  - For better testing with regions never seen before to validate accuracy
- Reserve 2020-22 data for Validation
- Reduce Imbalance by using larger Grid or TimeStamp to capture more Fire Incidents (1)



Feature (Re)Engineering

Predictors					Response
Dataset	**Grid**	**Property Assessment (2023)**	**Tax Roll (2021-2023)**	**Crimes (2015-2023)**	**Fires (2015-2023)**
Variables	- Grid ID - Geometry - Date - Year - Quarter	- Avg lot area - Total No. of lots - Count of residential label lots - Count of non-residential label lots - Total No. of residential accommodations - Avg of building age - Total built area	- Total area evaluation - Avg are evaluation - Total No. of tax parcel entries at address	- Crime occurrence by type	- Fire occurrence (Risk)
Merge field	Grid ID	Grid ID	Grid ID	Grid ID	Grid ID

Thank you