Laurence Forget Brisson

Capstone two – Final Report

Problem statement

What is the evolution of the impact of catastrophic events on the commercial aerial traffic in Canada, between 2001 and 2018?

Context

The number and severity of catastrophic events have considerably increased in recent years, and the negative impacts could be potentially significant for several industries, particularly for aerial transportation. In order to deal with this problem and consider a possible acceleration of these phenomenon related to climate change, it is necessary to understand how catastrophic events affect aerial traffic, especially to offset economic losses caused by interruptions, breakages, etc.

Datasets

To address this question, two datasets were used: 1) the "Canadian Disaster Database – Dataset"[1], later called 'disaster'; and 2) the "Operating and financial statistics for major Canadian airlines, monthly"[2], later called 'airline'. Both datasets are open access. The first dataset is from the Government of Canada, while the second is from Statistics Canada.

Data wrangling

Preliminary examination of both datasets

Disaster

- The dataset contains 263 rows and 22 columns.
- Original columns are: 'EVENT CATEGORY', 'EVENT GROUP', 'EVENT SUBGROUP', 'EVENT TYPE', 'PLACE', 'EVENT START DATE', 'COMMENTS', 'FATALITIES', 'INJURED / INFECTED', 'EVACUATED', 'ESTIMATED TOTAL COST', 'NORMALIZED TOTAL COST', 'EVENT END DATE', 'FEDERAL DFAA PAYMENTS', 'PROVINCIAL DFAA PAYMENTS', 'PROVINCIAL DEPARTMENT PAYMENTS', 'MUNICIPAL COSTS', 'OGD COSTS', 'INSURANCE PAYMENTS', 'NGO PAYMENTS', 'UTILITY - PEOPLE AFFECTED', and 'MAGNITUDE'.
- Catastrophic events are recorded in Canada from April 13th 2000 to April 18th 2019.
- There is a considerable amount of missing values.
- Some columns do not record any valuable information for the project and should be dropped.
- Some rows are an extension of the previous one ('Comments' entry) and do not record another disaster. They should be repatriated or dropped.

[1] Government of Canada, Canadian Disaster Database, https://www.publicsafety.gc.ca/cnt/rsrcs/cndn-dsstr-dtbs/index-en.aspx. For the provinces of British Columbia, Alberta, Saskatchewan, Manitoba, Ontario and Québec. Selection of meteorological and geological disaster between 2000 and 2022.

[2] Statistics Canada. Table 23-10-0079-01 Operating and financial statistics for major Canadian airlines, monthly, https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2310007901&cubeTimeFrame.startMonth=01&cubeTimeFrame.startYear=2000&cubeTimeFrame.endMonth=01&cubeTimeFrame.endYear=2022&referencePeriods=20000101%2C20220101&request_locale=en.

  - The first column being something else then "Disaster" indicates that there is an expansion of the "Comments" data in on multiple rows, as previously described.

Airline

- The dataset contains 444 610 rows and 17 columns.
- Original columns are: 'REF_DATE', 'GEO', 'DGUID', 'Airports', 'Class of operation', 'Peak hour and peak day of movements', 'UOM', 'UOM_ID', 'SCALAR_FACTOR', 'SCALAR_ID', 'VECTOR', 'COORDINATE', 'VALUE', 'STATUS', 'SYMBOL', 'TERMINATED', 'DECIMALS'.
- Data are recorded from January 1997 to February 2022 and should be reduce from 04-2000 to 04-2019 to match with the disaster dataset.
- There is a considerable amount of missing values, but not in the columns of interest.

Data cleaning

For both dataset, missing values have been managed, time values have been formatted, relevant columns have been preserved while the others have been dropped and both datasets were set to correspond to the same timeframe.

The airline dataset has further been fragmented into two, corresponding to airline_total (at the country scale) and airline_local (at the province and territory scale). Here are the remaining cleaned dataset:

- The disaster dataset contains the information relative to the natural disasters recorded in Canada during the period between 2001 and 2018;
- The airline_total dataset contains the monthly total aerial movements for Canada for the period between 2001 and 2018;
- The airline_local dataset contains the monthly aerial movements recorded for each airport of Canada for the period between 2001 and 2018.

Exploratory data analysis

In total, 229 environmental disasters events were recorded in Canada between 2001 and 2018. It was observed that most (98%) of these events are related to meteorological or hydrological phenomenon, while the remaining 2% have a geological origin.

13 types of disaster events are distinguished in the dataset:

1. Storms and severe thunderstorms;
2. Tornado;
3. Wildfire;
4. Flood;
5. Heat event;
6. Storm - Unspecified/Other;
7. Winter storm;
8. Landslide;
9. Hurricane/Typhoon/Tropical Storm;
10. Earthquake;
11. Drought;
12. Avalanche;
13. Storm surge.

The occurrence of each of these events during the period of interest can be visualized in figure 1, where we can see that flood is the most common event and appended 74 times. Wildfire and storms/severe thunderstorms are the next more frequent disaster, with 56 and 50 occurrences, respectively. Figures 2 and 3 illustrate the distribution of each event types per year, and by provinces and territories, respectively.

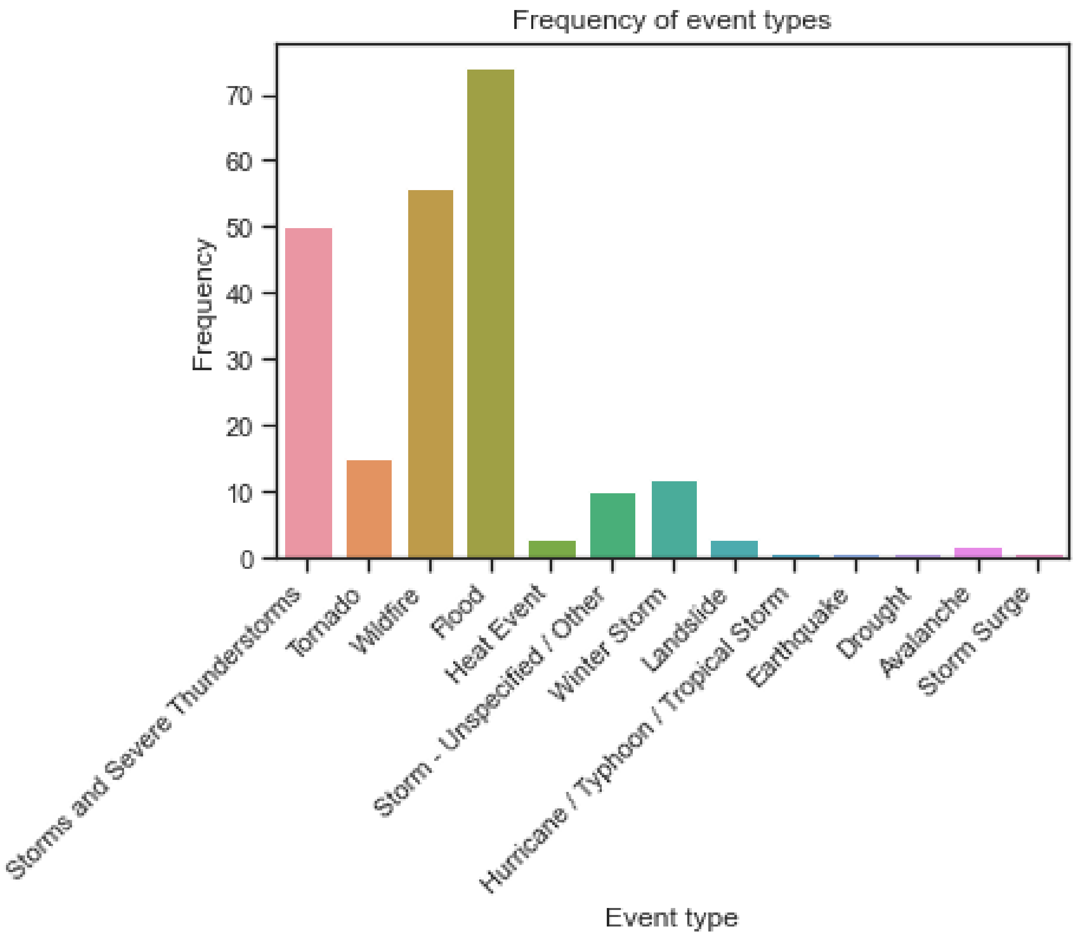

Figure 1 : Frequency of environmental disaster event types, between 2001 and 2018.

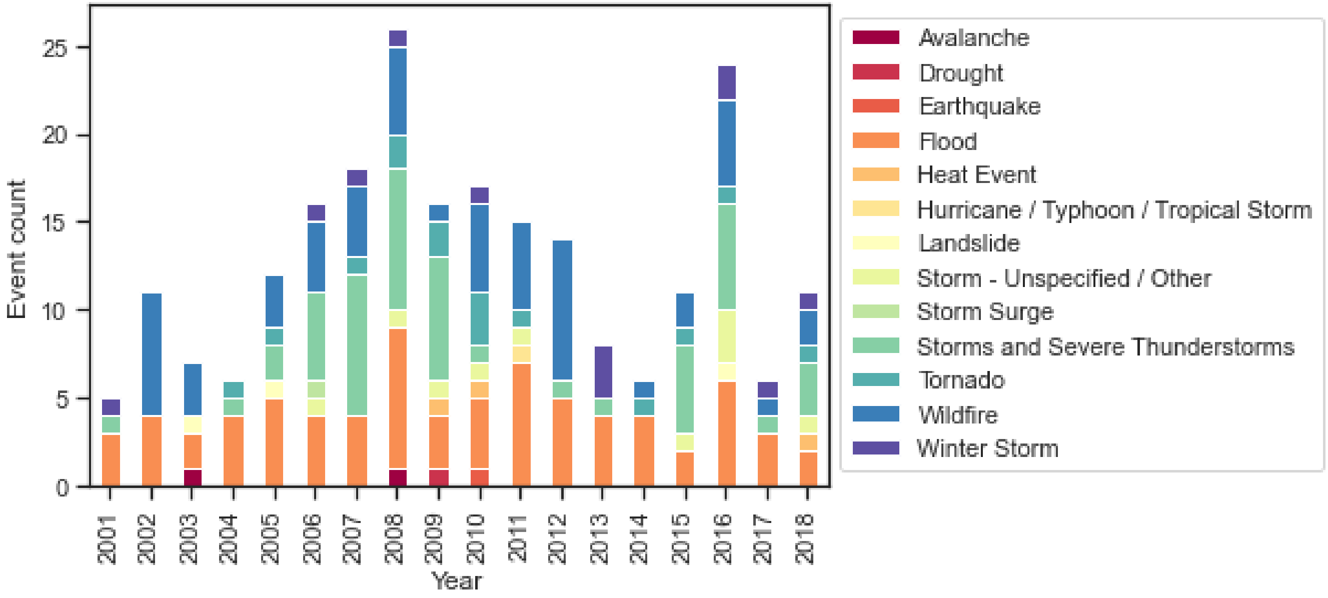

Figure 2 : Occurrence of each environmental disaster event type per year, between 2001 and 2018.

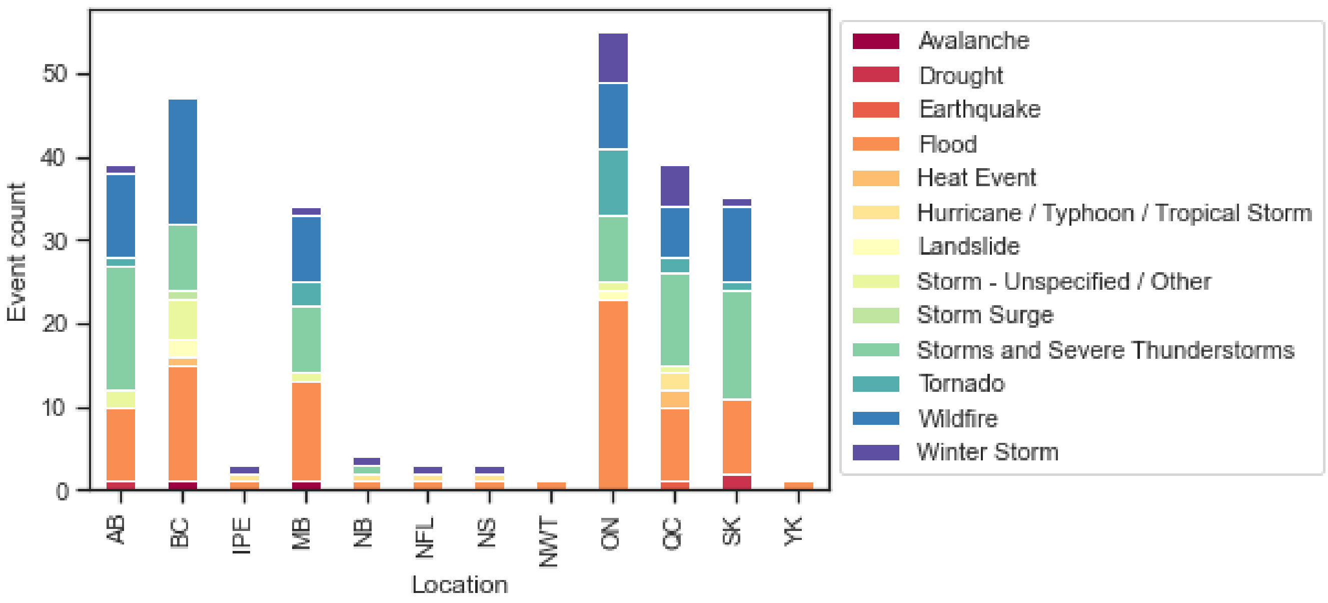

Figure 3 : Occurrence of each environmental disaster event type per provinces or territories. AB: Alberta, BC: British Columbia, IPE: Prince Edward Island, MB: Manitoba, NB: New Brunswick, NFL: Newfoundland and Labrador, NS: Nova Scotia, NWT: Northwest Territories, ON: Ontario, QC: Quebec, SK: Saskatchewan, YK: Yukon.

We can see that most event are happening in the most populated provinces, and this may be due to a bias in data logging. For instance, it is quite hard to believed that no environmental disaster ever happened in Nunavut.

Considering aerial traffic at the country scale, we can see in figure 4 that there is a major decrease in the number of air flight between 2001 (the first year considered) and 2006. There is then a peak in the number of flights between 2007 and 2008, and then a subsequent decrease. Between 2001 and 2018, the minimum number of aerial movements is 4 353 913 (in 2004), while the maximum is 4 933 606 (in 2008).

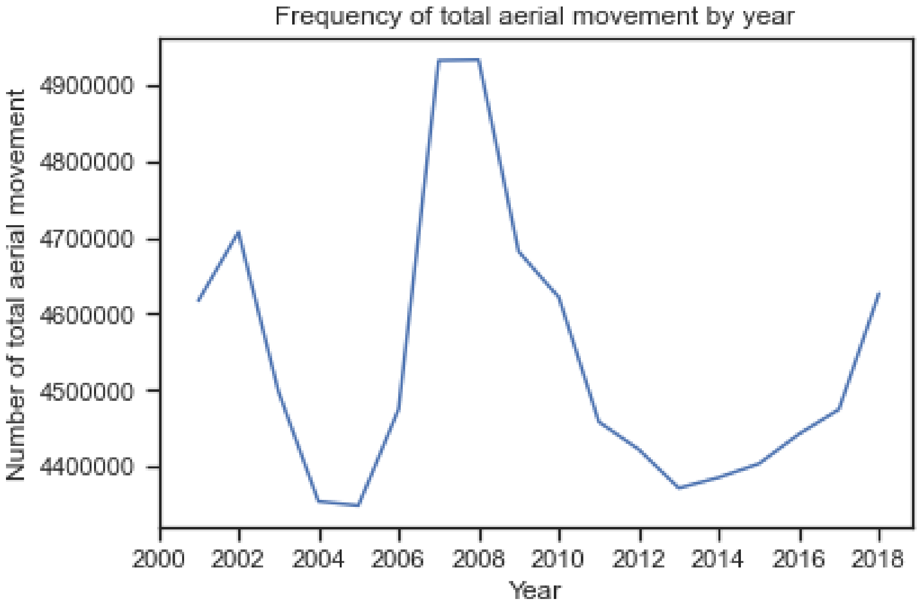

Figure 4 : Number of aerial movements per year in Canada.

By looking at the distribution of flights yearly, per provinces or territories (figure 5), we can see that Ontario and British Columbia are the principal aerial traffic contributors. It must be noted that flights connecting remote places are under evaluated in this dataset. This could be related to the fact that these flights are combining passenger and merchandise transportation and hence are not recorded in this dataset, and this is probably why there is no aerial traffic associated to Nunavut.

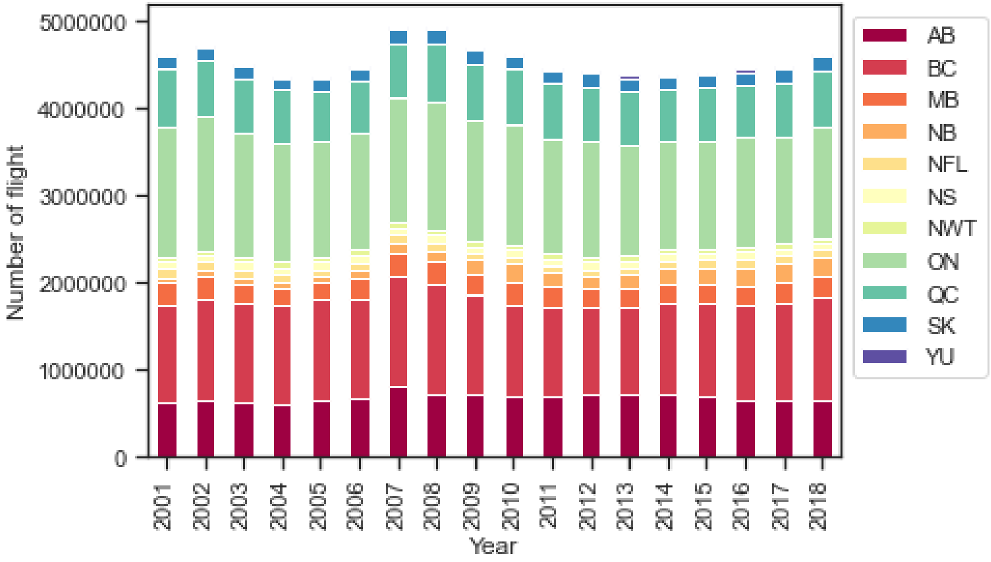

Figure 5 : Aerial traffic per year, by provinces or territories.

Both datasets were merged in order to have, monthly, for each province or territory, the number of flights as well as the occurrence of each type of environmental disaster events in the same table. Basic analysis was then carried out to visualise (figure 6) the relationship between the total number of flight and the sum of all disaster events monthly.

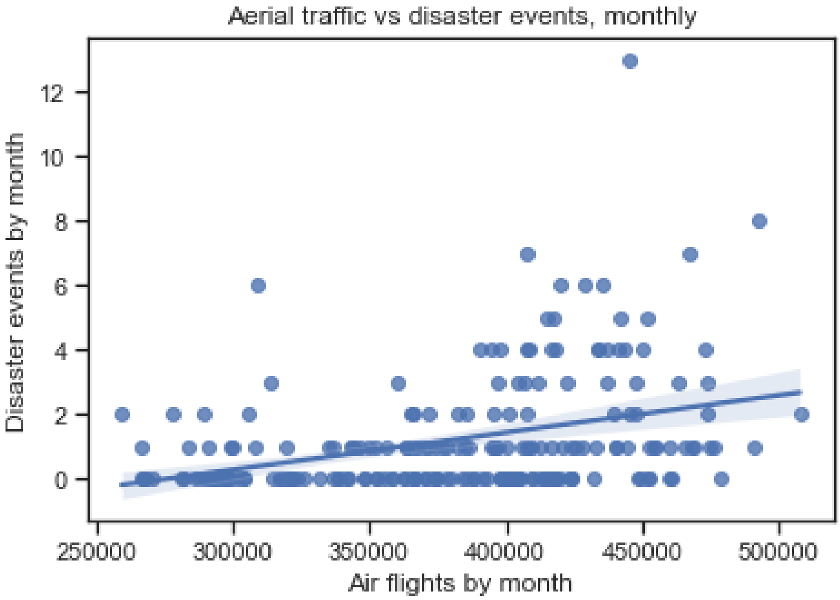

Figure 6 : Relation between aerial traffic and the number of disaster event, monthly.

It can be noticed that while the correlation coefficient is quite weak (~0.365), contrary to one might expect, the number of flights seams to be increasing with the number of environmental disaster events happening. However, it is very well possible that this last variable is not the most important one nor even a major one influencing the number of flights and maybe there is a confounding effect that is not taken into account (e.g. economic growth).

Pre-processing and training data development

In the column "prov_ter" is enclosed provinces and territories information, which are categorical values. As model only consider numerical values, dummy variables were used to translate this data, allowing his use. In this same perspective, dates (year and month) were transformed into float, as well as all other variables.

Since there is only one numerical quantification variable in the dataset, which correspond to aerial traffic value, no normalization step was performed.

Modeling

Modeling is used in this project to predict the aerial traffic (a continuous variable) through time considering the occurrence of environmental disaster events. Two types of models have been tested on this dataset, which are 1) linear regression and 2) random forest. While a linear regression model allows to simulate the relation between one or multiple independent variables and the target (dependent) variable to assess trends and generate estimation or prevision, random forest performs multiple decision trees, each making a prediction (classification or regression), in order to improve accuracy and minimize over-fitting. In all cases, data were split into training and testing set according to an 80-20 ratio.

Linear regression

Basic linear regression

In the case of the basic linear regression modeling, the R-squared coefficient (R^2 score), measuring the rightness of the fit, can be averaging to 0.95 and indicates that the data are well honored by the model. The Root Mean Squared Error (RMSE) indicates how spread out the data is around the best fit. In the present case, the RSME is quite high (8370). The overall model accuracy is 77.68% while the average error is 4847. The metrics are displayed in table 2. The relation between the actual and the predicted values (figure 7) show a good fit for the lower values, but an increase in step in the higher range, engendering a large spread in the predicted values.

Ordinary Least Squares (OLS) linear regression

For OSL linear regression, the R^2 score and the RMSE are equivalent to the ones of the basic linear regression (table 2), indicating that both models are doing similarly in terms of prediction. There is a slight decrease in average error and a small increase in model accuracy. The high similarities between the fits can be considered by comparing figures 7 and 8.

Random forest

Default random forest regressor

Default random forest regression performed without any parameter tuning has an equivalent R^2 score as the two last presented models, while the RSME value is now a little higher (table 2). There is a considerable increase in the average error value, as well as for the model accuracy. It can be noticed that the relation between the actual and the predicted values is better than for both version of linear regression in the lower range, and the prediction is now more gradual and less in step in for the higher values (figure 9).

Parameter tuned random forest regressor

K-fold cross-validation and random grid search was used to identify the best parameters which could improve the result of random forest regression modeling. The parameters to be tuned are: “n_estimators”, “max_features”, “max_depth”, “min_samples_split” and “min_samples_leaf”. Best values for the parameters, indicated by “search.best_params” are presented in table 1.

It has to be noted that there is no improvement in the R^2 score with parameter tuning. The RSME value is also exactly the same as for the basic random forest regressor. Yet there is a decrease in the average error value and a small increase (0.07%) for the model accuracy. However, there is no visible change in the fit (figure 10).

Table 1 : Best parameters for random forest regression.

Best parameters	
n_estimators	600
min_samples_split	10
min_samples_leaf	2
max_features	auto
max_depth	100
Boostrap	True

Model’s summary

Between the different tested models, both random forest regressors have the best results, according to accuracy and average error. In this case where the dataset is of reasonable size, the implementation and running time do not proscribe the use of parameter tuned random forest to benefit from the slight improvement in accuracy and average error. Yet if the conditions were to change, the use of default random forest could be also appropriate.

Table 2 : Model performance comparison

Model	R^2 score	RMSE	Average error	Accuracy
Linear regression	0.953	8370	4847°	77.68%
OLS linear regression	0.946	8379	4810°	77.72%
Default random forest regressor	0.952	8465	4499°	82.91%
Parameters tuned random forest regressor	0.952	8465	4429°	82.98%

With the aim of improving and deepening the question of the impact of environmental disaster events on aerial traffic, a more detailed dataset should be used. For each disaster event, amplitude should be recorded. For the aerial traffic dataset, an increased granularity could help, for instance by having a number of flights weekly. A supplementary dataset documenting cancelled or postponed flights could also be included to have a better understanding of the impact. Yet, a better dataset would be useful only in the case where environmental disaster do actually have an impact of the target variable. This is not what is implied by the modeling presented in this report.

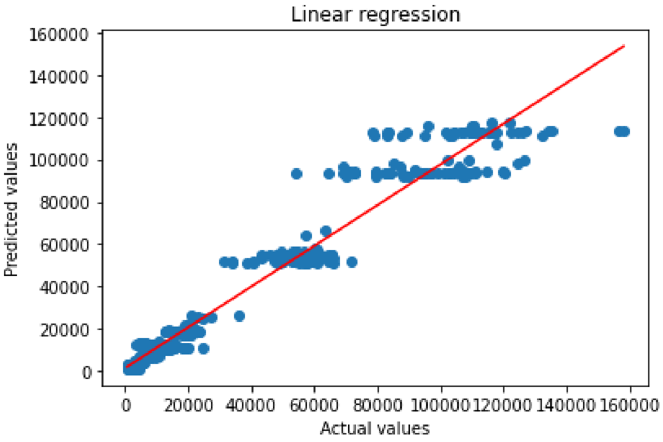

Figure 7 : Linear regression modeling.

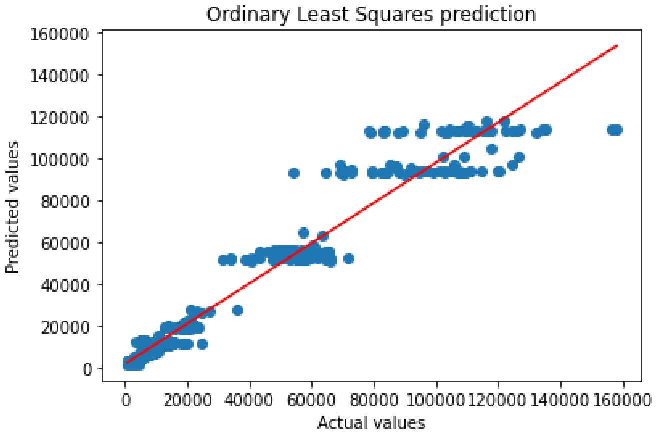

Figure 8 : Ordinary least squares prediction modeling.

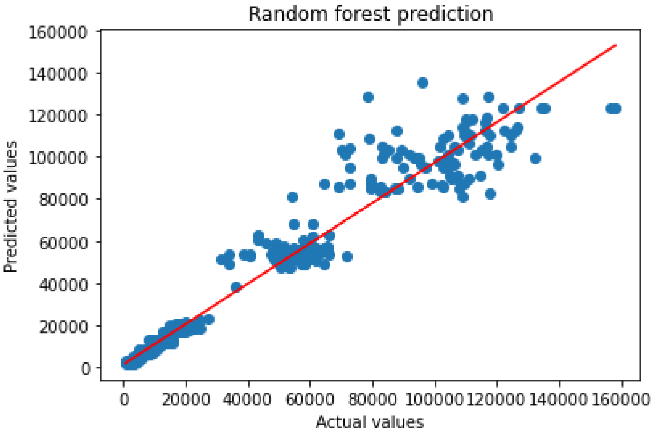

Figure 9 : Basic: random forest prediction modeling

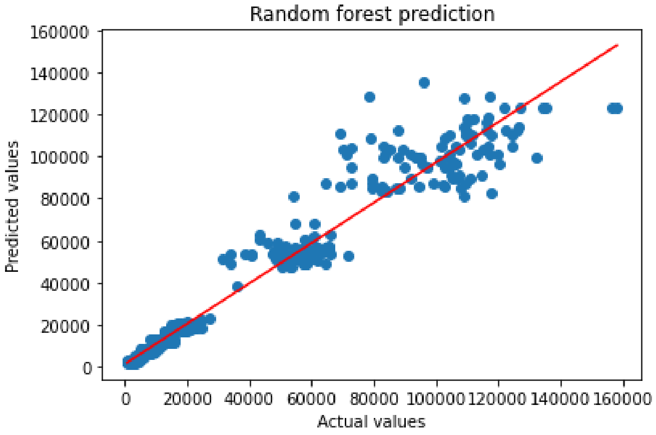

Figure 10 : Parameter tuned random forest modeling.

Github links

Data Wrangling

https://github.com/LaurenceFB/Capstone2/blob/main/Capstone2_DataWrangling_LForgetBrisson.ipynb

Exploratory Data Analysis

https://github.com/LaurenceFB/Capstone2/blob/main/Capstone2_EDA_LForgetBrisson.ipynb

Pre-processing and Training Data Development

https://github.com/LaurenceFB/Capstone2/blob/main/Capstone2_Pre-processingTraining_LForgetBrisson.ipynb

Modeling

https://github.com/LaurenceFB/Capstone2/blob/main/Capstone2_Modeling_LForgetBrisson.ipynb

Metrics file

https://github.com/LaurenceFB/Capstone2/blob/main/Captone2_metrics.txt