Springboard

Capstone 2
Final presentation

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

By Laurence Forget Brisson
Presented to Ricardo D. Alanis-Tamez

Sources:
https://www.readersdigest.ca/wp-content/uploads/2021/08/natural-disasters-in-canada-tornado.jpg
https://images.twnmm.com/c55i45ef3o2a/3WsJpNSAbkFqefOefM9Kuo/be63580aaa0a8279687e6eb7499ab7af/Wildfire-Noah-Berger-AP-Photo-Conversation.jpg
https://d3d0lqu00lnqvz.cloudfront.net/media/media/86edab5a-1476-4145-911f-6e0a94bb25cc.jpg

Datasets

2 datasets were used to address the question:

Canadian Disaster Database – Dataset

	EVENT CATEGORY	EVENT GROUP	EVENT SUBGROUP	EVENT TYPE	PLACE	EVENT START DATE	COMMENTS	FATALITIES	INJURED / INFECTED	EVACUATED	...	EVENT END DATE	FEDERAL DFAA PAYMENTS
0	Disaster	Natural	Meteorological - Hydrological	Flood	Eastern Canada	4/18/2019 12:00:00 AM	Extensive flooding in April and May was experi...	NaN	NaN	NaN	...	NaN	NaN
1	The most severe flooding took place in Quebec ...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN		NaN	NaN
2	States of emergency extended across the 3 prov...	1	NaN	10000									

Operating and financial statistics for major Canadian airlines, monthly

	REF_DATE	GEO	DGUID	Airports	Class of operation	Peak hour and peak day of movements	UOM	UOM_ID	SCALAR_FACTOR	SCALAR_ID	VECTOR	COORDINATE
0	1997-01	Canada	2016A000011124	Total, all airports	Total, itinerant and local movements	Number of movements	Number	223	units	0	v41126217	1.1.1.2
1	1997-01	Canada	2016A000011124	Total, all airports	Itinerant movements	Number of movements	Number	223	units	0	v41126218	1.1.2.2
2	1997-01	Canada	2016A000011124	Total, all airports	Local movements	Number of movements	Number	223	units	0	v41126219	1.1.3.2
3	1997-01	Canada	2016A000011124	Total, all airports	Civil local movements	Number of movements	Number	223	units	0	v41126220	1.1.4.2
4	1997-01	Canada	2016A000011124	Total, all airports	Military local movements	Number of movements	Number	223	units	0	v41126221	1.1.5.2

Data wrangling

- ✓ Missing values
- ✓ Date and time value formatting
- ✓ Feature selection
- ✓ Timeframe adjustment between both datasets (2001 to 2018)
- ✓ Preparation of 3 distinct datasets to explore in EDA:

    1. disaster: Natural disasters recorded in Canada between 2001 and 2018.
    2. airline_total: Monthly total aerial movements for Canada between 2001 and 2018.
    3. airline_local: Monthly aerial movements recorded for each airport of Canada between 2001 and 2018.

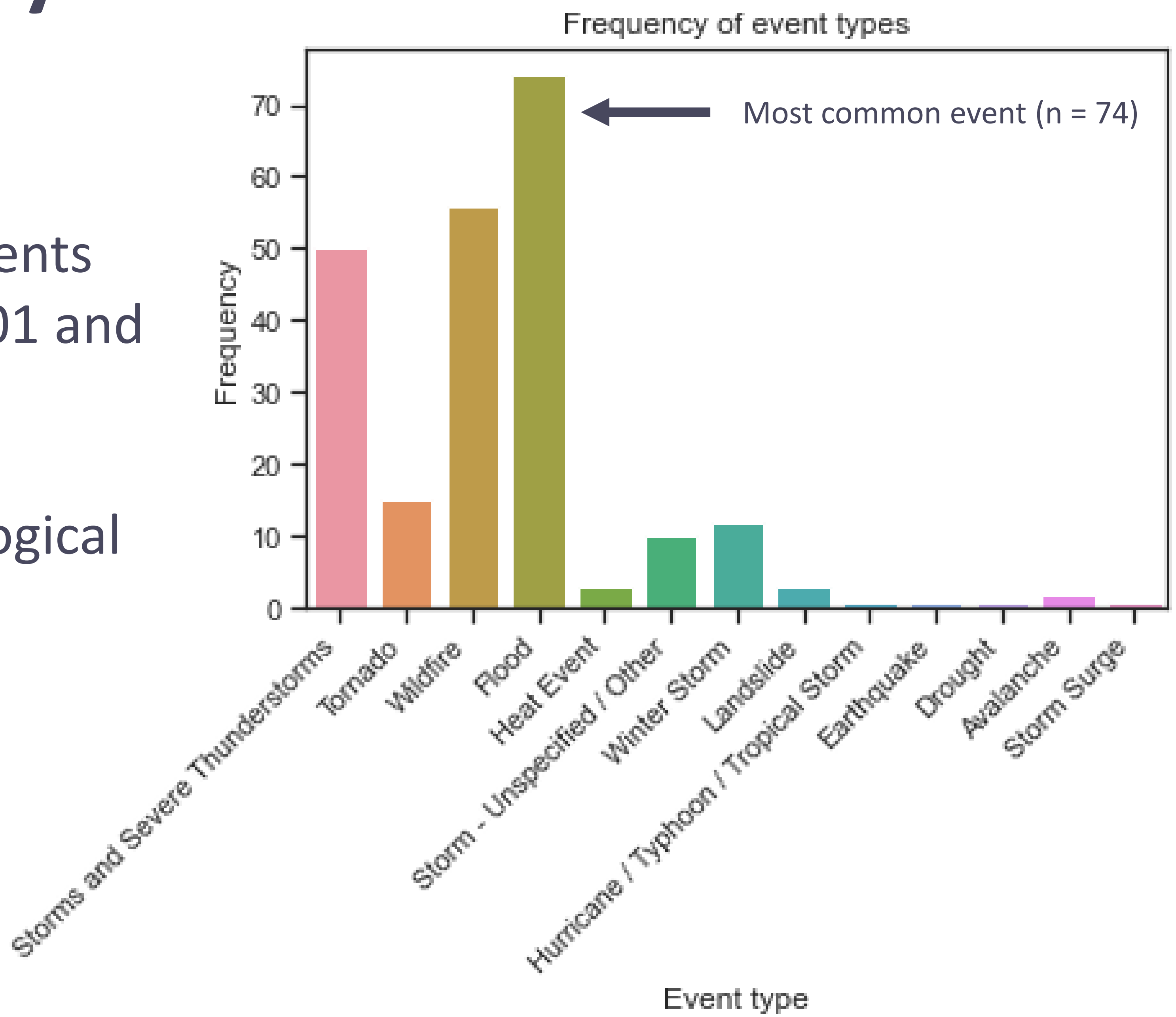
```
Entrée [12]:  # Isolating the timeframe of interest (between 04-2000 to 04-2019)
              date_range = (disaster['EVENT_START_DATE'] > '2001-1-1') & (disaster['EVENT_START_DATE'] <= '2018-12-31')
              disaster = disaster.loc[date_range]

Entrée [13]:  # Counting event by type.
              disaster.count()
```

Exploratory data analysis

229 environmental disaster events recorded in Canada between 2001 and 2018.

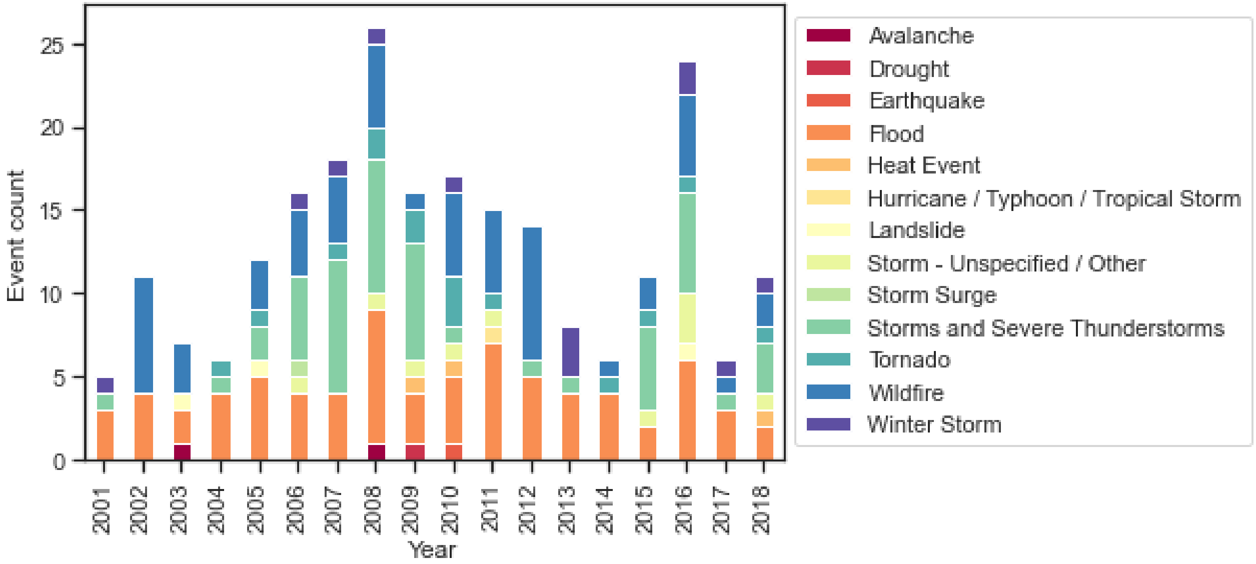
98% = meteorological or hydrological
2% = geological



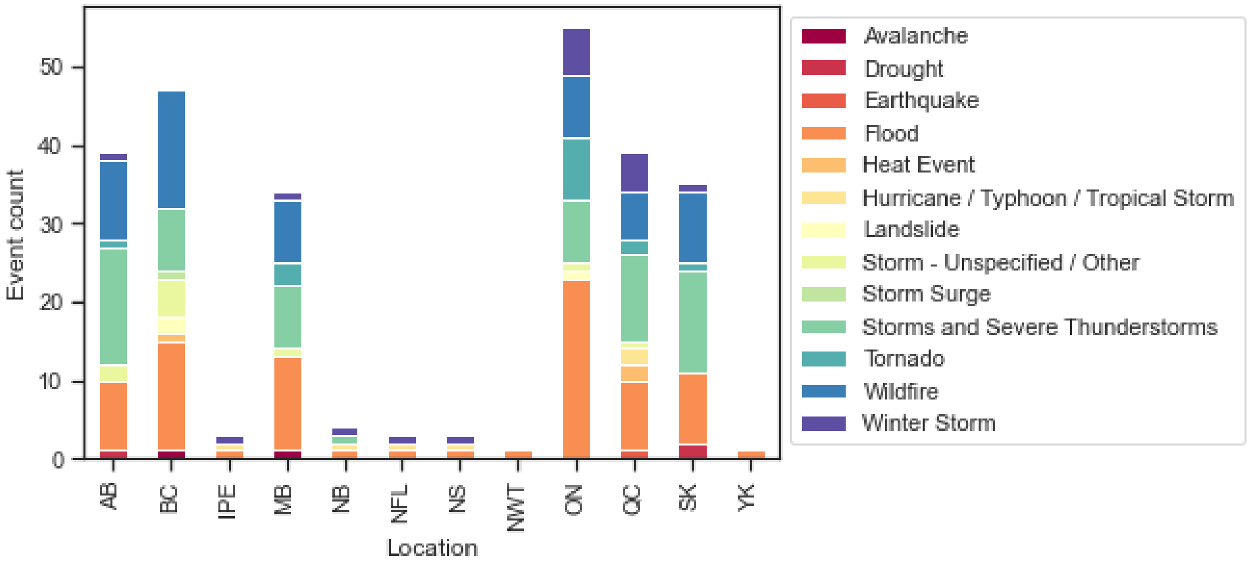
Exploratory data analysis

Distribution of natural disaster events:

Per year



By provinces and 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

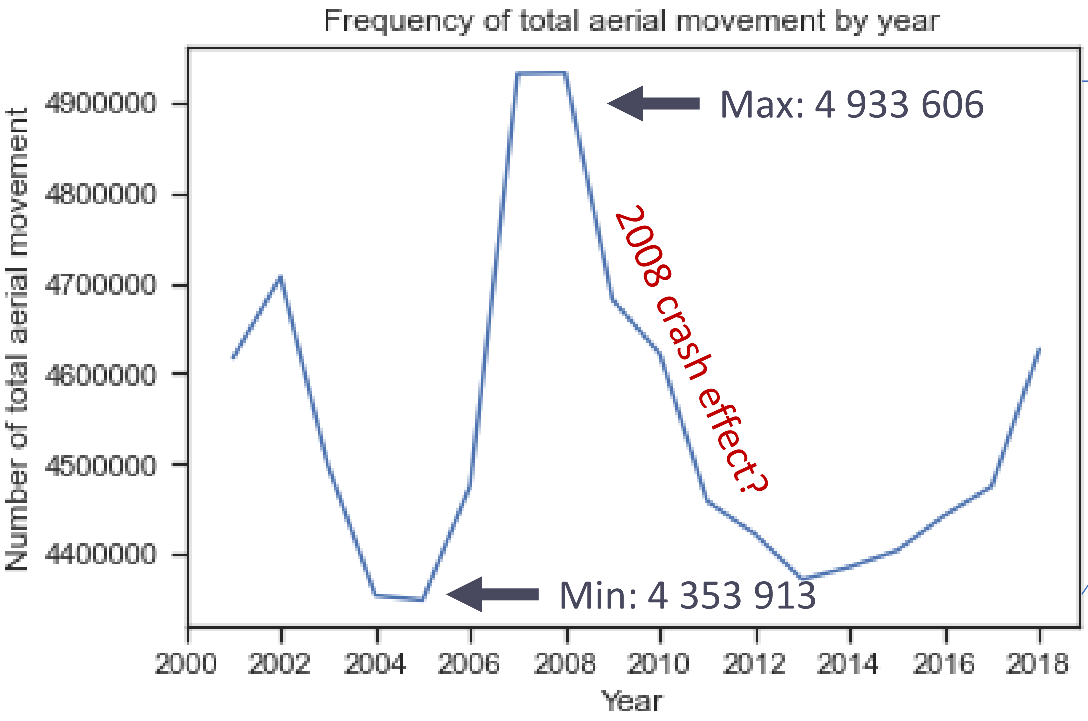
Some insights:

- Most events are happening in the most populated provinces.
  - Possible bias in data logging.
- Flooding happened every year.

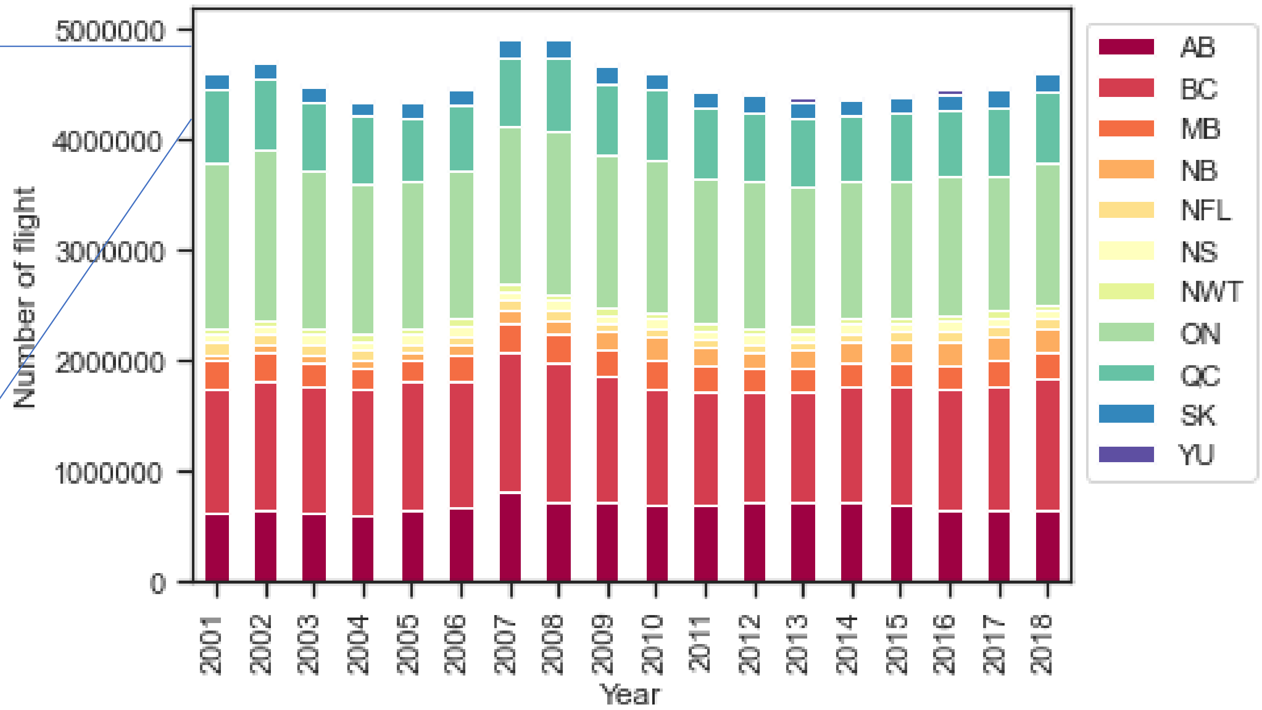
Exploratory data analysis

Aerial traffic at the country scale:

Total per year



By provinces and territories, per year

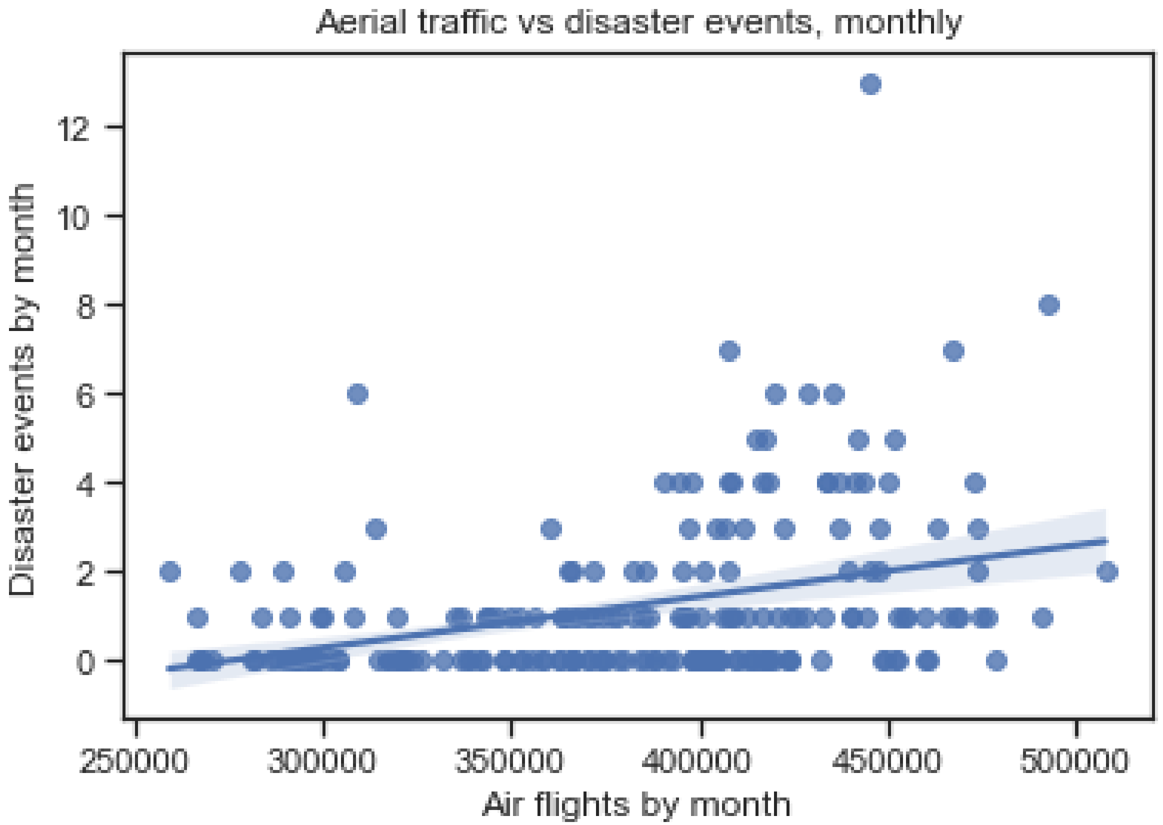


Some insights:

- ❖ Ontario and British Columbia are the principal contributors to aerial traffic.
- ❖ Flights connecting remote places are under evaluated (e.g. Nunavut)

Exploratory data analysis

Merging the datasets to analyse the number of flights as well as the occurrence of each type of environemental disaster events **monthly, for each province or territory**.



Some insights:

- ❖ Very weak correlation coefficient: ~0.365.
- ❖ The highest number of lights seems to coincide with the highest number of environmental disasters.
- ❖ It is well possible that the number of natural disasters is not the most important factor nor even a major one influencing the number of flights.
    - ➢ There is probably a confounding effect that is not taken into account (e.g. economic growth).

Pre-processing and training data development

- ✓ Dummy variables were created for the column ‘prov_ter’ (categorical values).
- ✓ Date values (year and month) were subdivised in two columns and transformed into float.
- ✓ Every other values were transformed into float.
- ✓ Since there is only one numerical quantification variable in the dataset (aerial traffic value) no normalization step was performed.
- ✓ Data was split into training and testing sets according to an 80-20 ratio.

```
df['year'] = df['year'].astype(float)
df['month'] = df['month'].astype(float)
```

```
Entrée [10]:
# Changing the datatype of VALUE to float.
df['VALUE'] = df['VALUE'].astype(float)
```

```
Entrée [11]:
# Visualizing the new version of the dataset
df.head()
```

Modeling

The goal is to predict the aerial traffic (a continuous variable), considering the occurrence of environmental disaster events.

2 variations of 2 different model types were tested.

To simulate the relation between one or multiple independent variables and the target (dependent) variable to assess trends and generate estimation or prevision.

- Linear regression
  - Default
  - Ordinary least square (OSL)
- Random forest regression
  - Default
  - Parameter tuned

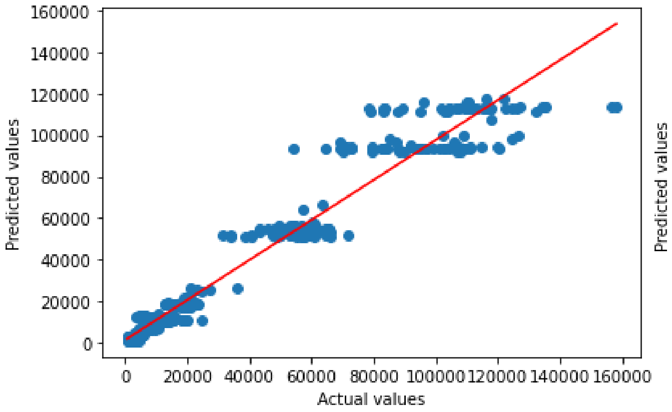
Performs multiple decision trees, each making a prediction to improve accuracy and minimize over-fitting.

Modeling

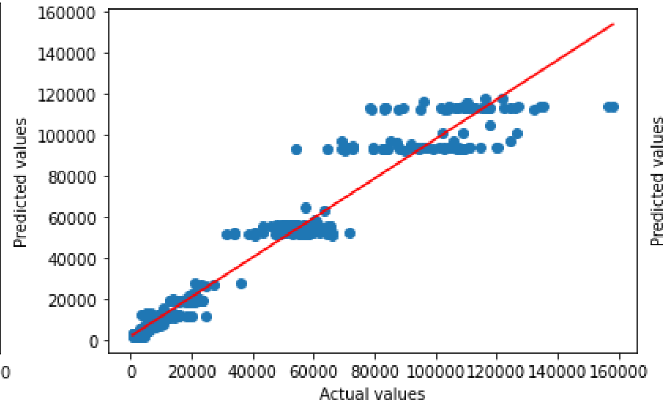
Model performance comparison

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%

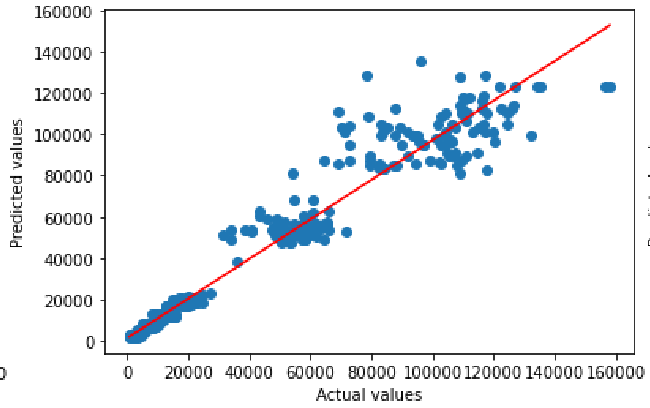
Linear regression



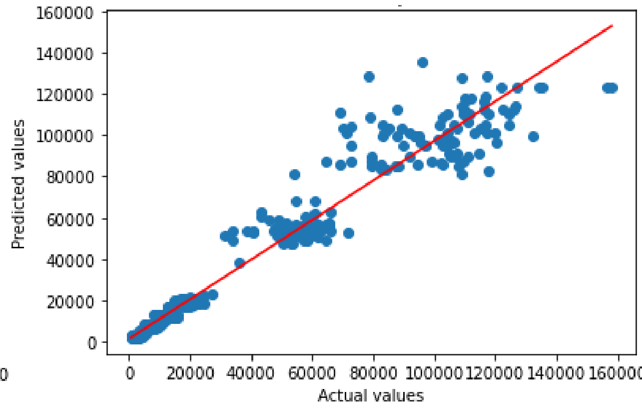
OLS Linear regression



Default random forest



Tuned random forest



Very similar

Very similar

Conclusion

- In this case where the dataset is of reasonable size, 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, and both linear regression models could be acceptable.

Best model:

Parameter tuned radom forest regressor

To look further

To improve and better understand the question of the impact of environmental disaster events on aerial traffic, there is a need to:

Obtain more detailed datasets.

- Have amplitude recorded for each disaster event.
- Have an increased granularity for the aerial traffic dataset, for instance recording the number of flights weekly.

Look for a supplementary dataset documenting cancelled or postponed flights.

Yet, a better dataset would only be useful if environmental disaster do actually have an impact of the target variable. This is not what is suggested by the modeling presented in this report.

Github links

Jupyter notebooks documenting each step of the data science method.

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_PreprocessingTraining_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