IBM Supervised Machine Learning Regression

Course Final Project:
Major Projects Cost Estimating with Regression

By Javier A. Jaime-Serrano

June 17, 2021

Abstract

For this Project we used the Major Project dataset from the province of Alberta in Canada [1], it contains more than 700 currently active projects on the province, this Data set if filtered for Projects valued at $5 million or greater.

This Dataset contain a lot of valuable information on the Major Projects.
The Estimated Cost, Sector & Type and the Location data will be used.

From this dataset, we will like to find the features that can predict the cost of the project for a given project type in a sector and in a region.

Objectives

From the Major Projects in the dataset, we want to be able to estimate the cost of similar projects, with a rough order of magnitude (ROOM).

We will choose the type of projects were we have enough data within a valid range, so we can estimate the cost of similar projects, with a rough order of magnitude (ROOM) in a matter of seconds, instead of using current techniques.

This quick ROOM Cost Estimate can be used as benchmark in multiple industries.

Data Cleaning

In order to prepare and clean the dataset:

- We drop the projects where there is no estimated cost.
- Made some assumptions about schedule completion and status.
- Drop not required columns and renamed the remaining columns.
- Made corrections on project types and sectors.

A problem encountered was how to extract the location coordinates (Longitude & Latitude) from a GeoJASON column. The problem was solved with Python code that loop over all rows and extract the start (first) locations by a type condition.

Data Cleaning

Figure 1. Cleaned Dataset

ProjectId	Project Name	Estimated Cost (millions)	Municipality	Forecasted Completion	Sector	Type	Stage	Developer	Start Latitude	Start Longitude
7	StoneGate Landing	3000.0	Calgary	2021	Mixed-Use	Mixed-Use	Started	WAM Development Group / AIMCo	51.172501	-113.975800
11	Shepard Station Suburban Office Campus Building 1	22.0	Calgary	2020	Commercial	Office: Low-Rise	Started	Shepard Development Corp.	50.931721	-113.970596
22	Barron Building Renovation	100.0	Calgary	2021	Residential	Apartment: Mid-Rise	Proposed	Strategic Group	51.046070	-114.076614
26	Quarry Crossing II Office Building	72.8	Calgary	2027	Commercial	Office: Low-Rise	Proposed	Remington Development Corp.	50.966900	-114.002899
32	Nolan Hill TownHomes	5.0	Calgary	2027	Residential	Townhouses	Proposed	Jayman Modus	51.162041	-114.160912

Data Exploration

We explored the data, first with descriptive statistics and bar charts (see Figure 2).

Second, we used box plots for the cost estimate ranges by type (see Figure 3).

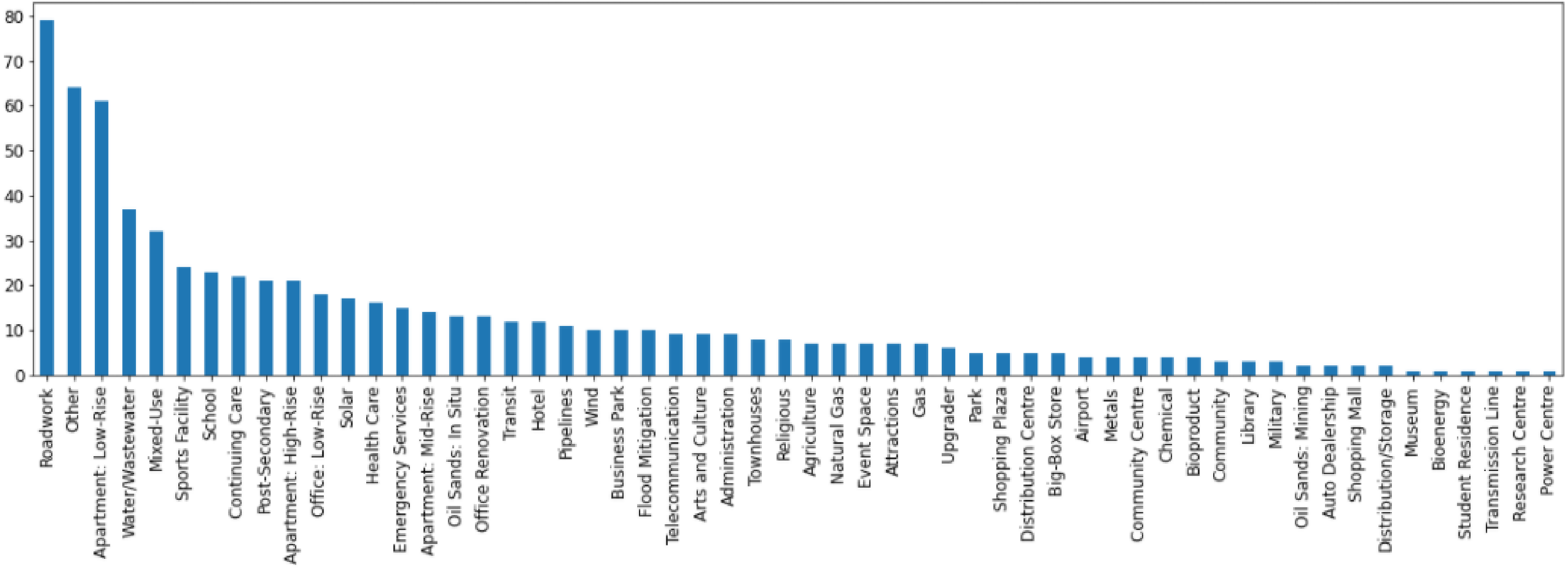
Third, we used folium library to create a map using latitude and longitude values (See Figure 4).

We also used KMeans [2] to Cluster the projects by geographical region using the location coordinates (See Figure 5).

And then we will be able to try the different Regressors from Scikit-learn [2] to estimate the (ROOM) Cost and evaluate the Results.

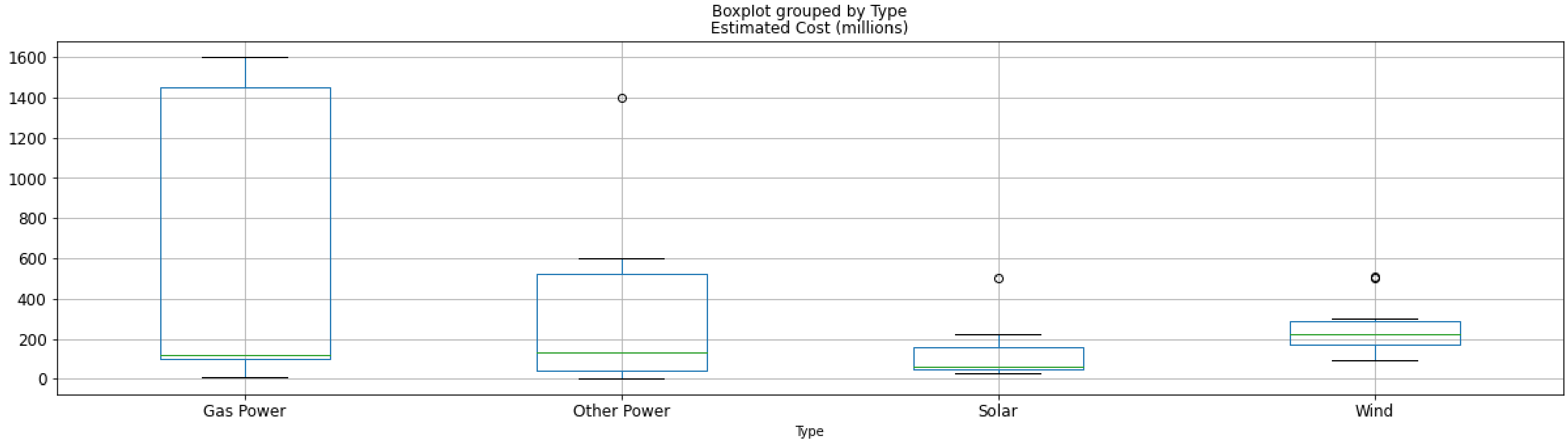
Data Exploration

Figure 2. Project types bar chart



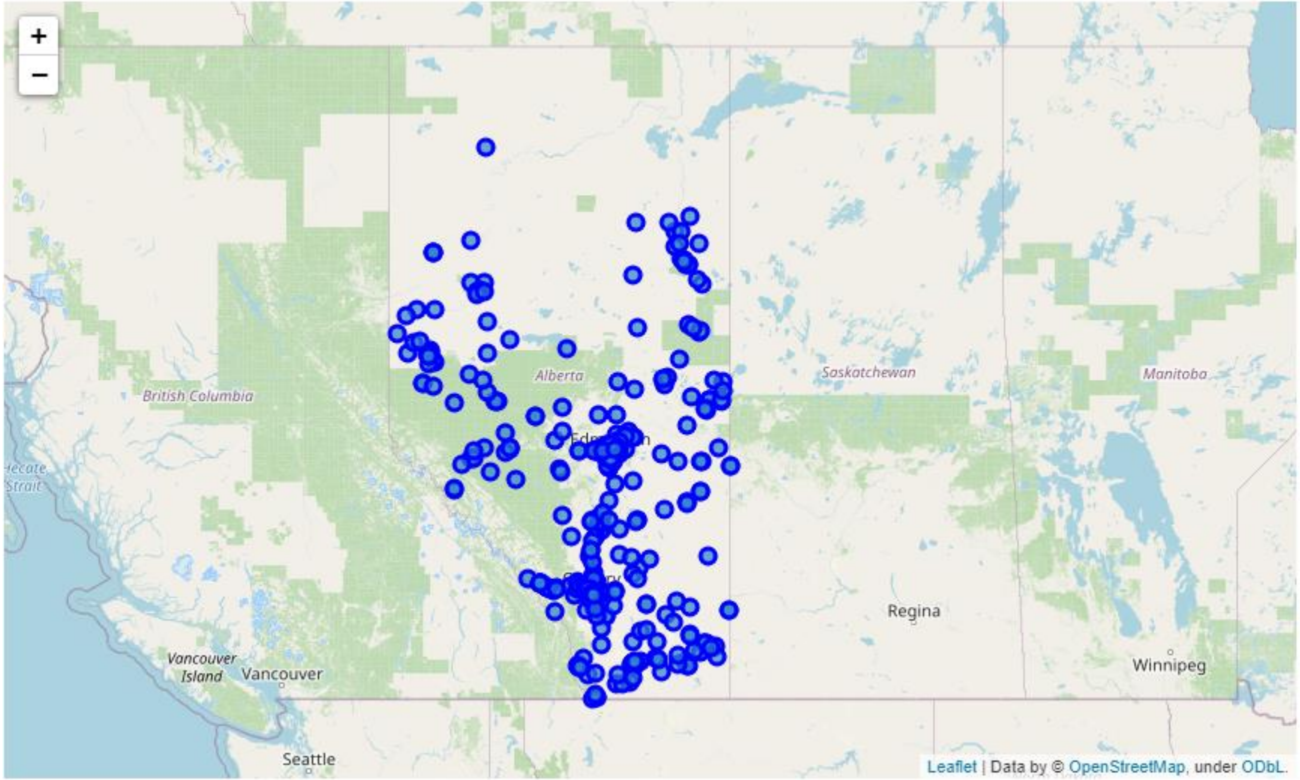
Data Exploration

Figure 3. Box Plot Estimated Cost for Power Sector



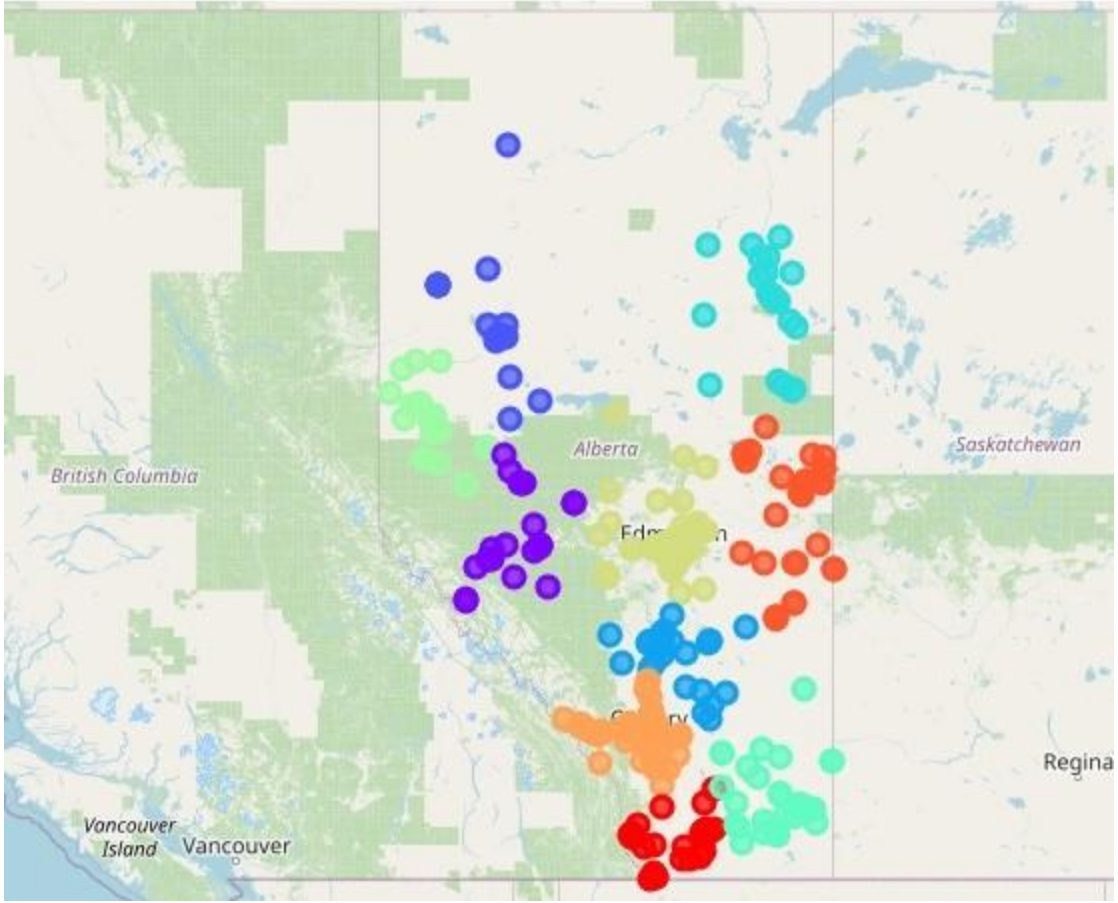
Data Exploration

Figure 4.
Projects Map



Data Exploration

Figure 5. Projects Map KMeans Clusters by Region with Project type



Key Findings and Insights

After corrections, we ended with 58 unique types in 9 sectors (see figure 6).

For the feature engineering, we extracted size and capacity data by type from the dropped project details column, adding units and cleaning it manually in excel.

The Estimated Cost was set as the Target variable in millions (removing 3 zeros).

The datasets were merged back, and one hot encoding was used again to transform the categorical features into numerical dummy features, resulting in 72 Columns.

Findings

Figure 6. Unique Sectors and Types

Sector	Type	Project Name
Commercial	Business Park	10
	Distribution Centre	5
	Office Renovation	13
	Offices	18
	Other Commercial	4
Industrial	Agriculture	7
	Bioproduct	4
	Chemical	4
	Metals	4
	Other Industrial	14
	Telecommunication	9
Infrastructure	Airport	4
	Flood Mitigation	10
	Other Infrastructure	16
	Roadwork	79
	Transit	12
	Water/Wastewater	37
Institutional	Administration	9
	Continuing Care	22
	Emergency Services	15
	Health Care	16
	Library	3
	Military	3
	Other Institutional	7
	Post-Secondary	21
	Religious	8
	School	23
Oil and Gas	Distribution/Storage	2
	Gas	7
	Oil Sands: In Situ	13
	Oil Sands: Mining	2
	Other Oil and Gas	2
	Pipelines	11
	Upgrader	6
Power	Gas Power	7
	Other Power	8
	Solar	17
	Wind	10
Residential	Apartment: High-Rise	18
	Apartment: Low-Rise	61
	Apartment: Mid-Rise	14
	Community	6
	Other Residential	4
	Townhouses	8
Retail	Auto Dealership	2
	Big-Box Store	5
	Mixed-Use	33
	Other Retail	4
	Shopping Mall	2
	Shopping Plaza	5
Tourism	Arts and Culture	9
	Attractions	8
	Community Centre	4
	Event Space	7
	Hotel	12
	Other Tourism	9
	Park	5
	Sports Facility	24

Hypothesis Testing

We want to be able to estimate the cost of similar projects, with a rough order of magnitude (ROOM). In order to test the predictive features with correlation, we need first to pick a sector, for this case we choose the power sector (see figure 7).
Is Size/Capacity in Megawatts is correlated with the Estimated Cost?

- Null Hypotesis (H0): Size/Capacity is not correlated with Estimated Cost
- Alternative Hypotesis (H1): Size/Capacity is correlated with Estimated Cost

From SciPy, Pearson Correlation Coeficient: `0.9676147594295178`
Two-tailed p-value: `6.958427885850458e-19`

Conclusion: There is a relationship between Capacity and Cost (see figure 8).

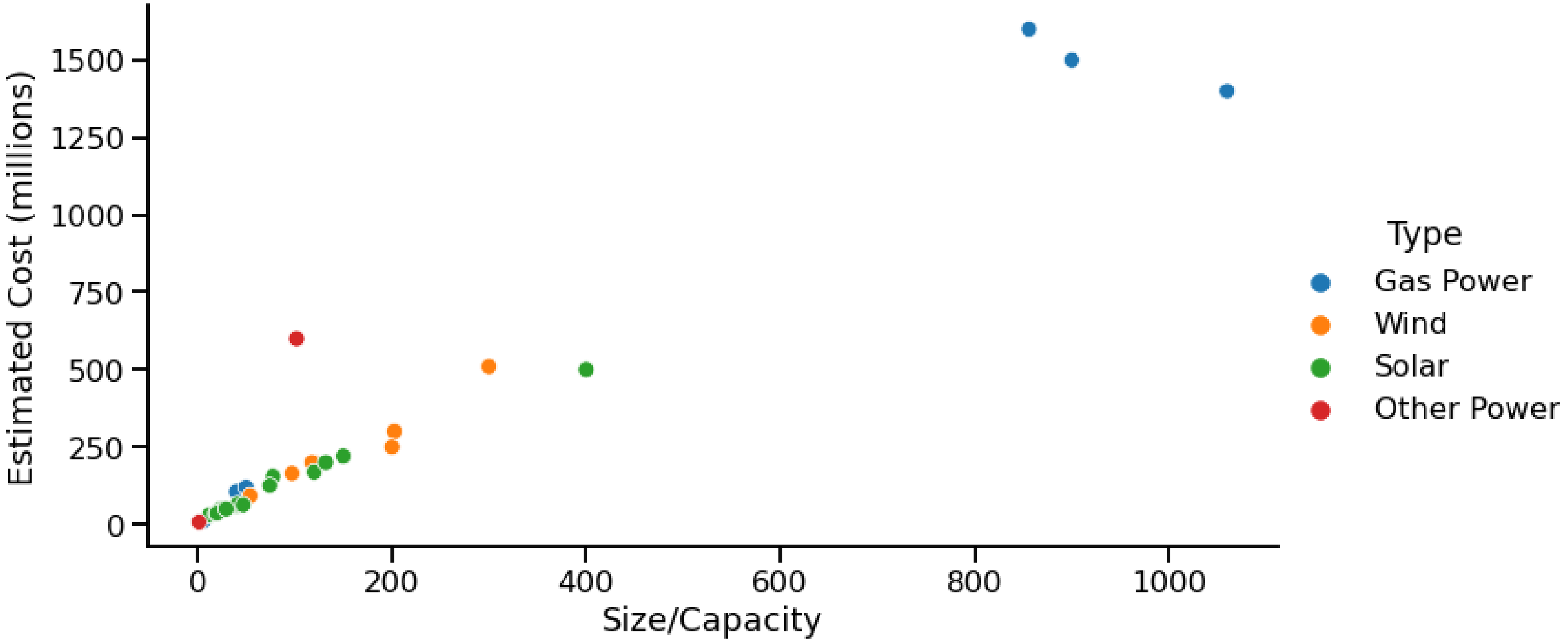
Hypothesis Testing

Figure 7. Power Sector Data

	Project Name	Estimated Cost (millions)	Type	Start Latitude	Start Longitude	Size/Capacity	$1000/Capacity
ProjectId							
642	Sundance 7 Gas-Fired Power Plant	1600.0	Gas Power	53.547501	-114.445000	856.0	1869.158879
644	Genesee Generating Station Units 4 and 5 Project	1400.0	Gas Power	53.547501	-114.445000	1060.0	1320.754717
649	Peace Butte Wind Farm	200.0	Wind	49.896000	-110.856003	120.0	1666.666667
653	Harvest Operations Gas Fired Power Plant	10.0	Gas Power	52.514198	-111.926003	5.6	1785.714286
2086	Vulcan Solar Project	155.0	Solar	50.093319	-112.848129	77.5	2000.000000

Hypothesis Testing

Figure 8. Estimated Cost Vs. Size/Capacity by Project Type



Results

We did the split of one hot encoded data, with 70% for training and the remaining 30% for Testing. And we run a simple linear regression with poor results.
We transformed the data with the Standar Scaler and fit the linear regression model again, with a considerable improvement on the predictions.
We added Polynomial Features to the one hot encoded data, fit the linear regression model again and obtained even worse results.

We tried a simpler model with the Capacity and Location Coordinates from the Power sector data only, from the Hypothesis Testing (See Figure 7), and after training the model with the spilt data, we obtained considerable better Results. We applied the Standard Scaler and added Polynomial Features (See Table 1).

Results

Table 1. r2 Score

r2 Score	One Hot Encoded	Not Encoded
LR Simple	**0.3562**	**0.6745**
LR with StandardScaler	**0.6477**	**0.7983**
LR with Poly Features	**0.3473**	**0.8769**

Conclusions

Using the Major Project dataset from the province of Alberta, after preparation and cleaning, we were able to extract valuable but limited information.

After exploring the dataset, we were able to do some feature engineering and filter the data set to extract valid information. We also performed significance testing to prove the correlation between the Capacity in the Power Sector and the Cost.

We tried two different approaches to linear regression: the one hot encoding of all the features and a simpler model with the selected features from the hypotesis testing with added Polynomial Features.

The Linear Regression Model with 2nd Degree Polynomal Features (without one hot encoding) obtained the best results using the r2 Score (See Table 1).

Future Work

Most of the projects have not enough data for the engineered features of size and/or capacity, so we only were able to predict the cost in one sector.

The polynomial features added complexity to the model, but without enough data in one sector, we couldn't obtain better results with higher degrees.

Further work is required to search for capacity data from other sources (company websites, industry associations, etc.).

With an augmented data set, we will be able to cross validate and fine tune the model with the optimum parameters.

References

[1] Alberta Major Projects: https://majorprojects.alberta.ca/

[2] Scikit-learn library: https://scikit-learn.org/stable/

[3] Jupiter Notebook: https://github.com/javier-jaime/IBM-Machine-Learning-Capstone/