IBM Supervised Machine Learning Classification

Course Final Project:
Major Projects Classification by Cost Range

By Javier A. Jaime-Serrano

June 16, 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.

We want to be able to classify projects by type, cost and capacity and compare them with similar projects from the same dataset.

We will use Decision Tree and Random Forest Classifiers from Scikit-learn [2].

Objectives

From the Major Projects in the dataset, we will like to know if the cost of the project is within a reliable range for the project type and size and/or capacity.

If the classifier results are projects “too cheap” or “too expensive”, or there is not enough data available in the data set, further analysis will be required.

For the type of projects that we have enough data within the range, we will be able to estimate the cost of similar projects, with a rough order of magnitude (ROOM).

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).

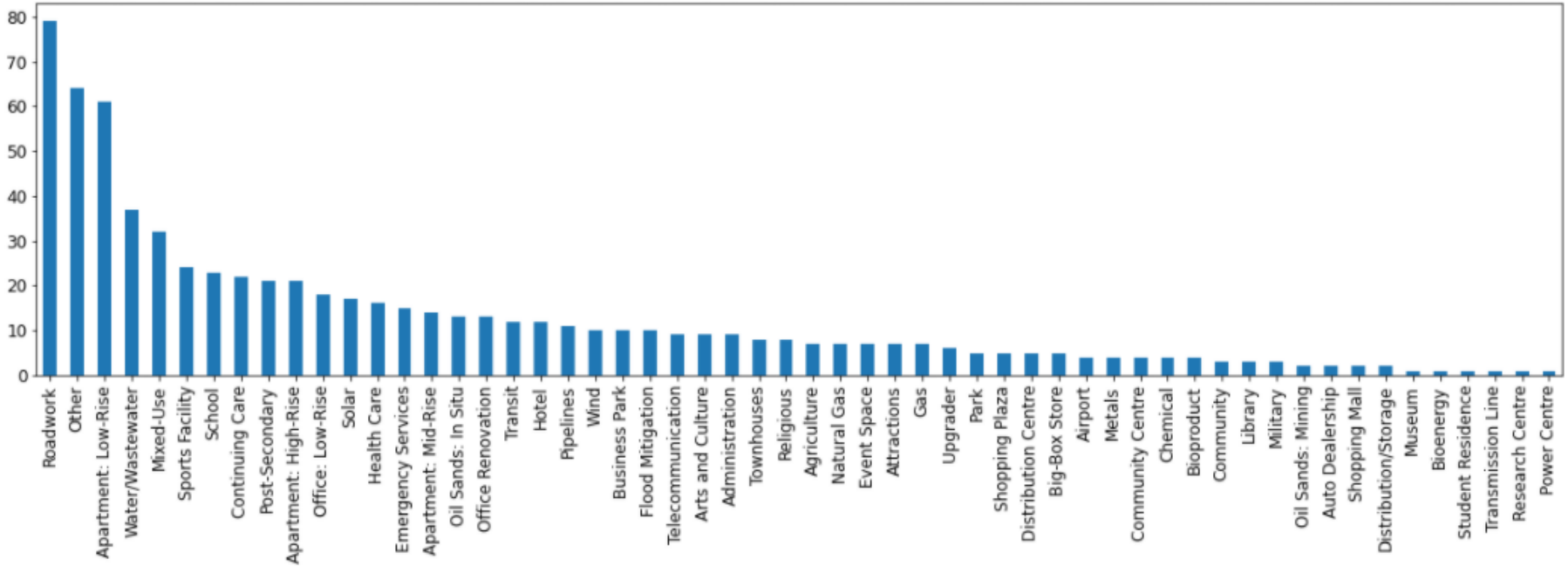
We also used KMeans [2] to Cluster the projects by geographical region using the location coordinates (See Figure 5).

And then we used the Classifiers from Scikit-learn [2] to predict one of 4 labels:

1. Too Cheap 2. In the Range. 3. Too Expensive 4. Not Enough Data

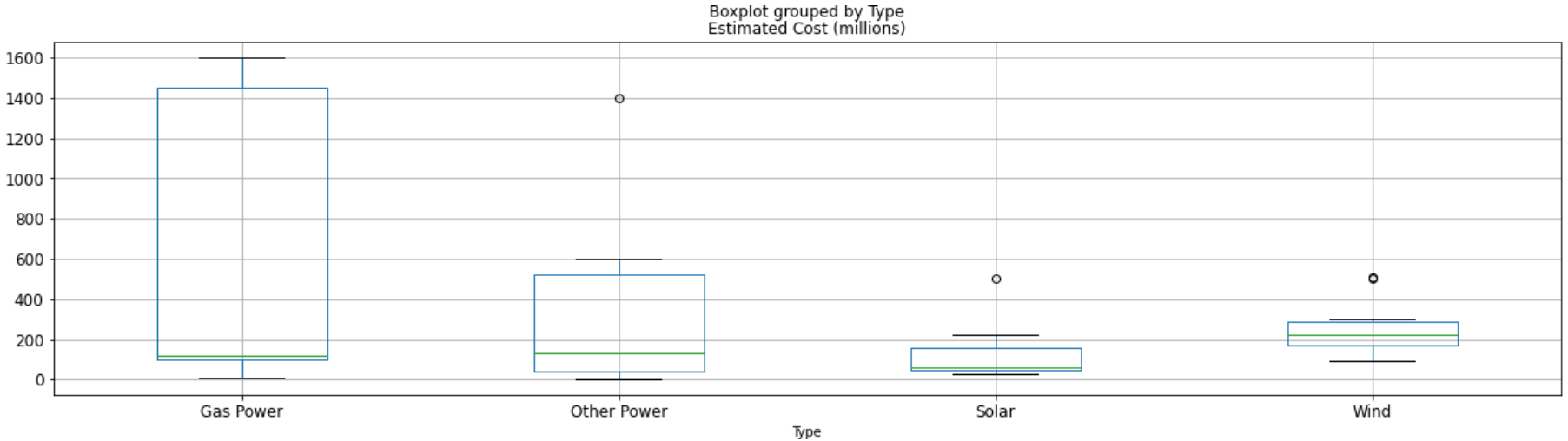
Data Exploration

Figure 2. Project types bar chart



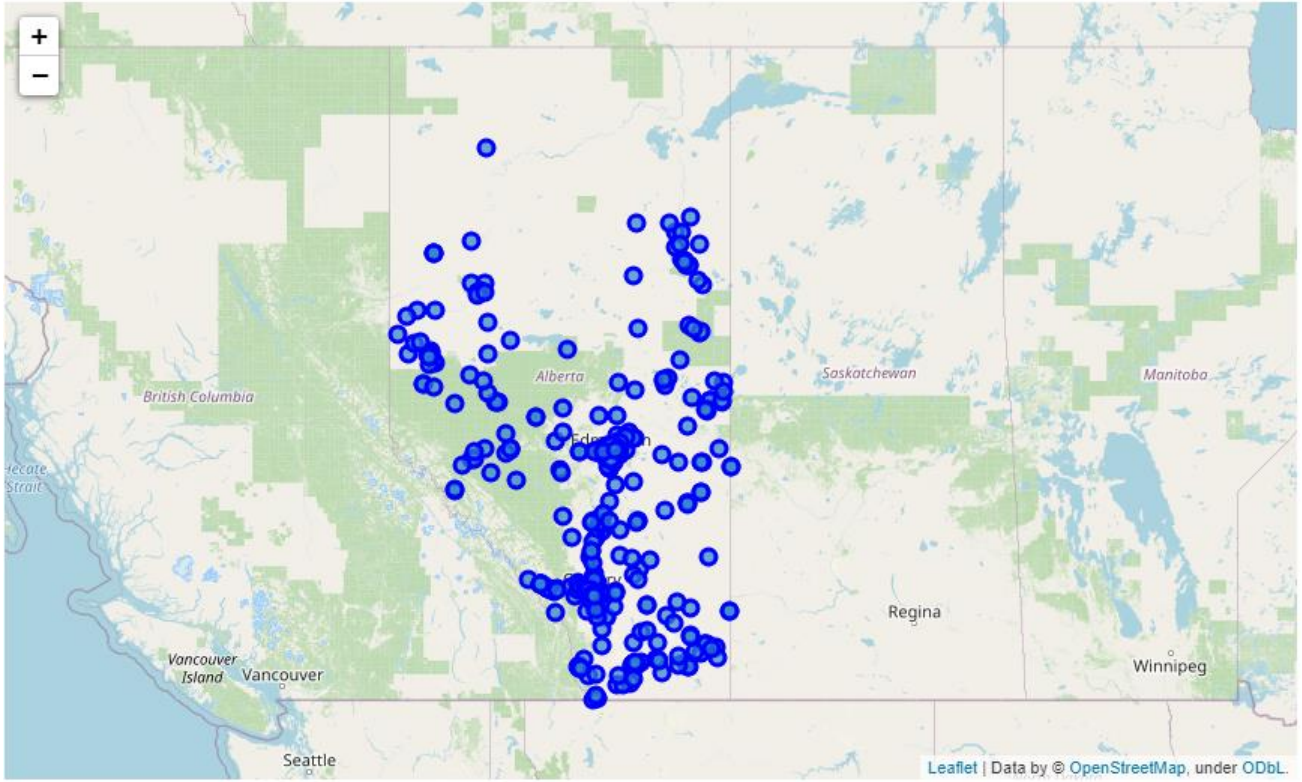
Data Exploration

Figure 3. Box Plot Estimated Cost for Power Sector



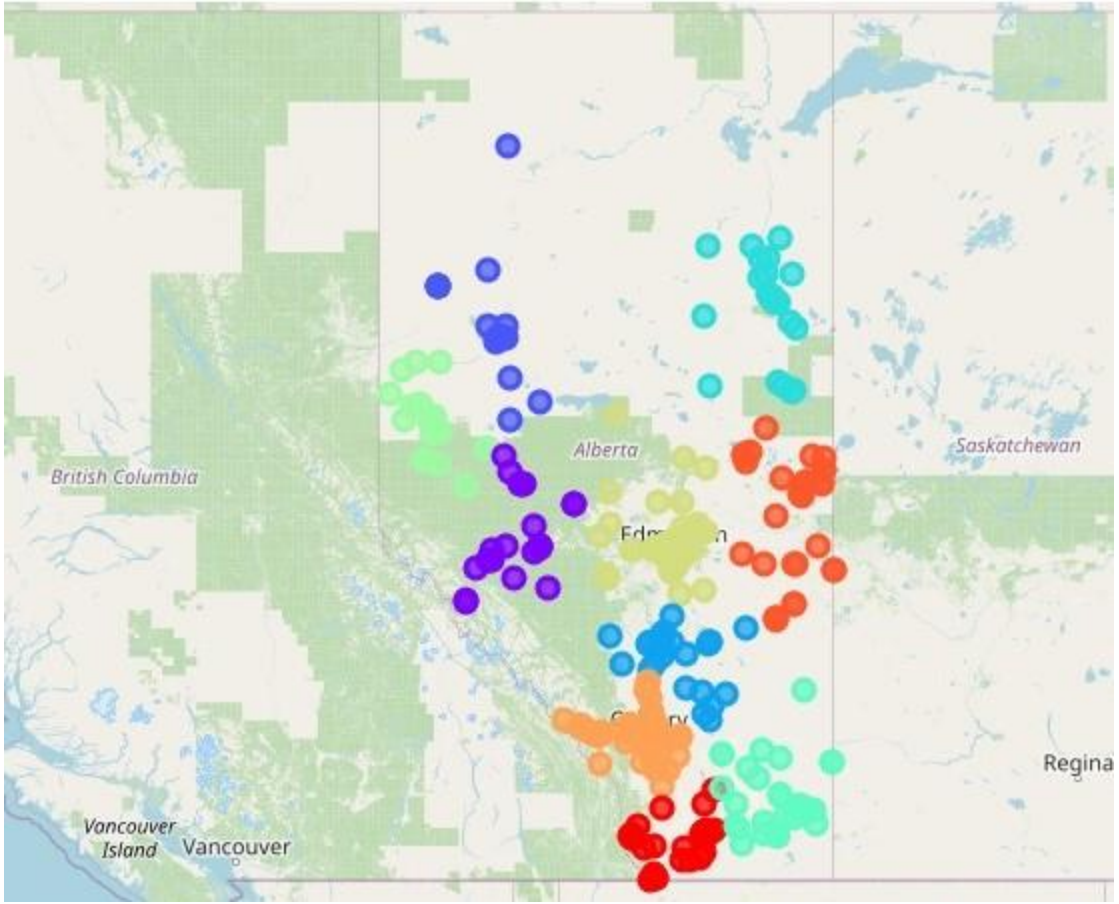
Data Exploration

Figure 4.
Projects Map



Data Exploration

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



Findings

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 target variable was set from a percentage range in relation to average cost, divided by capacity per type of projects, labeled with conditional formulas in excel.

The datasets were merged back, and one hot encoding was used again to transform the categorical features into numerical dummy features (89 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

Results

The Decision Tree Classifier was trained with 70% of the Dataset, reserving the remaining 30% for Testing.

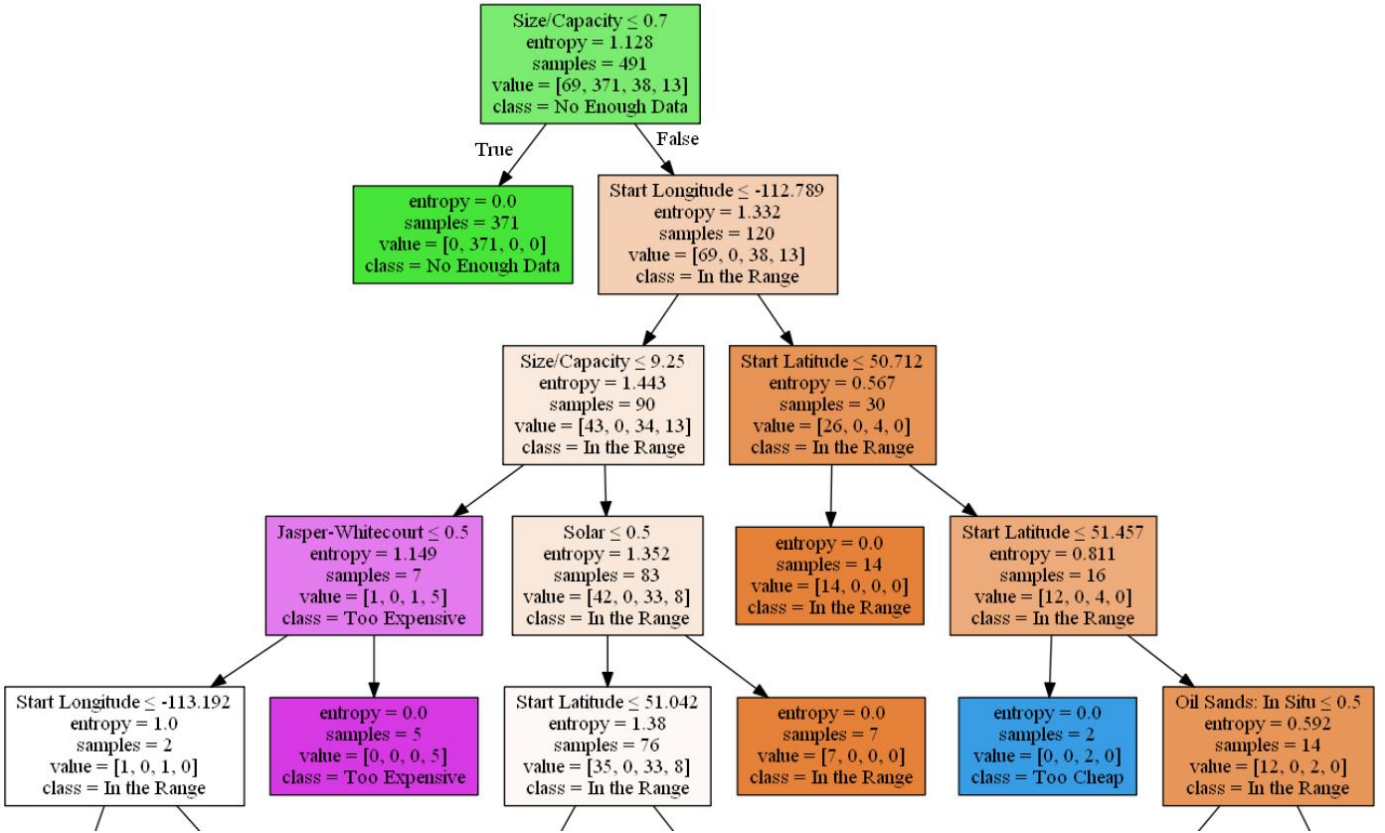
We used the Entropy (information gain) as the classifier criterion, and a maximum depth of 20 nodes (these are the 20 questions of the Project Name).

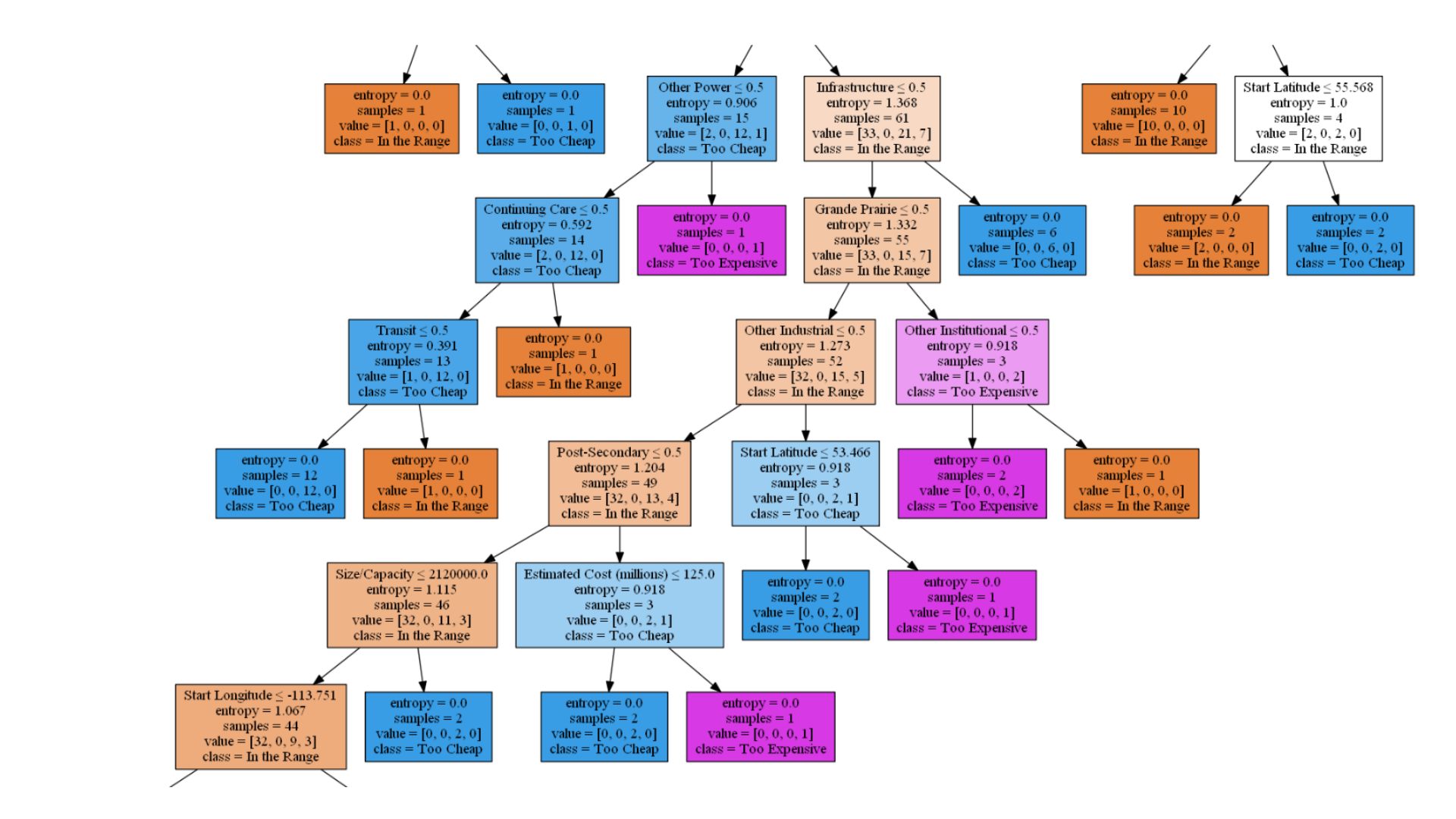
We used Grid Search to find the best hyperparameters of number of nodes and maximum depth of the tree.

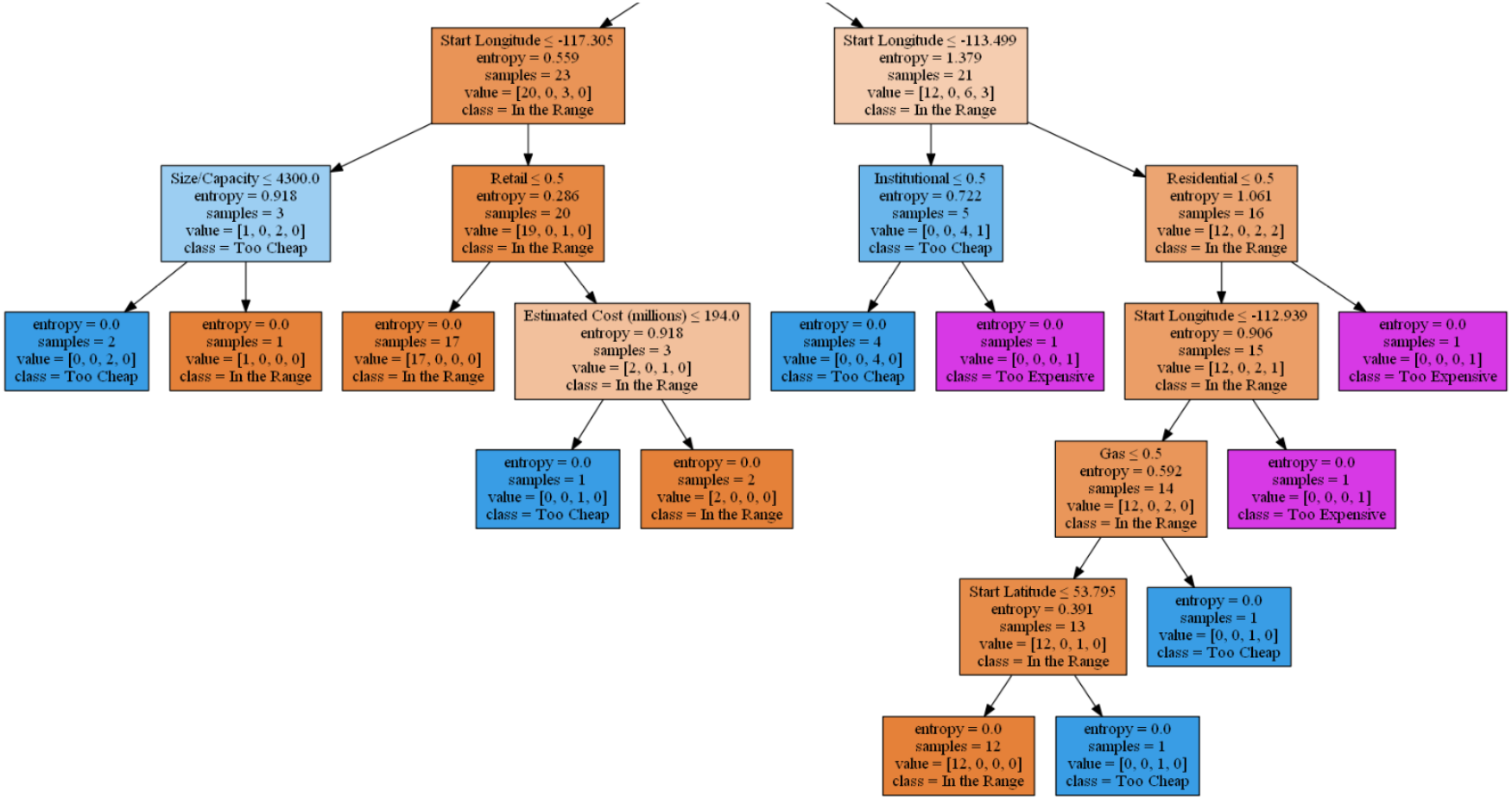
The Decision Tree Accuracy is 89% on the test set, the maximum number of nodes, depth is 16 and number of nodes is 67 (see figure 7).

Results

Figure 7.
Decision Tree




entropy = 0.0
samples = 1
value = [1, 0, 0, 0]
class = In the Range
entropy = 0.0
samples = 1
value = [0, 0, 1, 0]
class = Too Cheap
Other Power ≤ 0.5
entropy = 0.906
samples = 15
value = [2, 0, 12, 1]
class = Too Cheap
Infrastructure ≤ 0.5
entropy = 1.368
samples = 61
value = [33, 0, 21, 7]
class = In the Range
entropy = 0.0
samples = 10
value = [10, 0, 0, 0]
class = In the Range
Start Latitude ≤ 55.568
entropy = 1.0
samples = 4
value = [2, 0, 2, 0]
class = In the Range
Continuing Care ≤ 0.5
entropy = 0.592
samples = 14
value = [2, 0, 12, 0]
class = Too Cheap
entropy = 0.0
samples = 1
value = [0, 0, 0, 1]
class = Too Expensive
Grande Prairie ≤ 0.5
entropy = 1.332
samples = 55
value = [33, 0, 15, 7]
class = In the Range
entropy = 0.0
samples = 6
value = [0, 0, 6, 0]
class = Too Cheap
entropy = 0.0
samples = 2
value = [2, 0, 0, 0]
class = In the Range
entropy = 0.0
samples = 2
value = [0, 0, 2, 0]
class = Too Cheap
Transit ≤ 0.5
entropy = 0.391
samples = 13
value = [1, 0, 12, 0]
class = Too Cheap
entropy = 0.0
samples = 1
value = [1, 0, 0, 0]
class = In the Range
Other Industrial ≤ 0.5
entropy = 1.273
samples = 52
value = [32, 0, 15, 5]
class = In the Range
Other Institutional ≤ 0.5
entropy = 0.918
samples = 3
value = [1, 0, 0, 2]
class = Too Expensive
entropy = 0.0
samples = 12
value = [0, 0, 12, 0]
class = Too Cheap
entropy = 0.0
samples = 1
value = [1, 0, 0, 0]
class = In the Range
Post-Secondary ≤ 0.5
entropy = 1.204
samples = 49
value = [32, 0, 13, 4]
class = In the Range
Start Latitude ≤ 53.466
entropy = 0.918
samples = 3
value = [0, 0, 2, 1]
class = Too Cheap
entropy = 0.0
samples = 2
value = [0, 0, 0, 2]
class = Too Expensive
entropy = 0.0
samples = 1
value = [1, 0, 0, 0]
class = In the Range
Size/Capacity ≤ 2120000.0
entropy = 1.115
samples = 46
value = [32, 0, 11, 3]
class = In the Range
Estimated Cost (millions) ≤ 125.0
entropy = 0.918
samples = 3
value = [0, 0, 2, 1]
class = Too Cheap
entropy = 0.0
samples = 2
value = [0, 0, 2, 0]
class = Too Cheap
entropy = 0.0
samples = 1
value = [0, 0, 0, 1]
class = Too Expensive
Start Longitude ≤ -113.751
entropy = 1.067
samples = 44
value = [32, 0, 9, 3]
class = In the Range
entropy = 0.0
samples = 2
value = [0, 0, 2, 0]
class = Too Cheap
entropy = 0.0
samples = 2
value = [0, 0, 2, 0]
class = Too Cheap
entropy = 0.0
samples = 1
value = [0, 0, 0, 1]
class = Too Expensive


Start Longitude ≤ -117.305
entropy = 0.559
samples = 23
value = [20, 0, 3, 0]
class = In the Range
Start Longitude ≤ -113.499
entropy = 1.379
samples = 21
value = [12, 0, 6, 3]
class = In the Range
Size/Capacity ≤ 4300.0
entropy = 0.918
samples = 3
value = [1, 0, 2, 0]
class = Too Cheap
Retail ≤ 0.5
entropy = 0.286
samples = 20
value = [19, 0, 1, 0]
class = In the Range
Institutional ≤ 0.5
entropy = 0.722
samples = 5
value = [0, 0, 4, 1]
class = Too Cheap
Residential ≤ 0.5
entropy = 1.061
samples = 16
value = [12, 0, 2, 2]
class = In the Range
entropy = 0.0
samples = 2
value = [0, 0, 2, 0]
class = Too Cheap
entropy = 0.0
samples = 1
value = [1, 0, 0, 0]
class = In the Range
entropy = 0.0
samples = 17
value = [17, 0, 0, 0]
class = In the Range
Estimated Cost (millions) ≤ 194.0
entropy = 0.918
samples = 3
value = [2, 0, 1, 0]
class = In the Range
entropy = 0.0
samples = 4
value = [0, 0, 4, 0]
class = Too Cheap
entropy = 0.0
samples = 1
value = [0, 0, 0, 1]
class = Too Expensive
Start Longitude ≤ -112.939
entropy = 0.906
samples = 15
value = [12, 0, 2, 1]
class = In the Range
entropy = 0.0
samples = 1
value = [0, 0, 0, 1]
class = Too Expensive
entropy = 0.0
samples = 1
value = [0, 0, 1, 0]
class = Too Cheap
entropy = 0.0
samples = 2
value = [2, 0, 0, 0]
class = In the Range
Gas ≤ 0.5
entropy = 0.592
samples = 14
value = [12, 0, 2, 0]
class = In the Range
entropy = 0.0
samples = 1
value = [0, 0, 0, 1]
class = Too Expensive
Start Latitude ≤ 53.795
entropy = 0.391
samples = 13
value = [12, 0, 1, 0]
class = In the Range
entropy = 0.0
samples = 1
value = [0, 0, 1, 0]
class = Too Cheap
entropy = 0.0
samples = 12
value = [12, 0, 0, 0]
class = In the Range
entropy = 0.0
samples = 1
value = [0, 0, 1, 0]
class = Too Cheap

Limitations

The results of Decision Tree Classifier show the limitations of the Dataset, the first split take most of the projects out, classifying them as “Not Enough Data”.

Only 97 projects were classified as “In the Range”, and Only 3 types has more than 10 Projects in this valid range.

This work was done with only one Province Dataset, and other Canadian Provinces has Datasets available for Future work in similar formats.

Conclusions

We were able to classify project by type, cost and capacity with the use of a Decision Tree Classifier with 89% accuracy on the training set.

Most of the projects has not enough data for the engineered features of size and/or capacity. Further work is required to search for capacity data from other sources (company websites, industry associations, etc.).

We found additional uses for the classifier, as the detected outliers were in a couple of cases, big projects announcements (as publicity stunts) and later withdrawn. And projects that were too expensive were never completed.

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/