Classification and Trends: Distribution of Electric Vehicle Types in Washington State

Name: Semiu Kolapo
Student Number: 501145293
Supervisor: Tamer Abdou
Date: April 1st 2024

Ryerson University
The Chang School of Continuing Education

Why Electric Vehicle?

- Government of Canada committed to achieve 100% zero-emission vehicle sales by 2035
- Washington State's progressive policies
- Tesla

Using ML to identify trends and distribution

TESLA

Ryerson University

The Chang School of Continuing Education

Predicting Electric Vehicles

- Can we predict type of Electric Vehicle?
- Can we predict the distribution of BEVs and PHEVs?
- Which ML algorithms can yield the best and most accurate results?
- Are there parameters within the Dataset that can give insights on the distribution of BEVs and PHEVs?
- How confident are we in our findings?

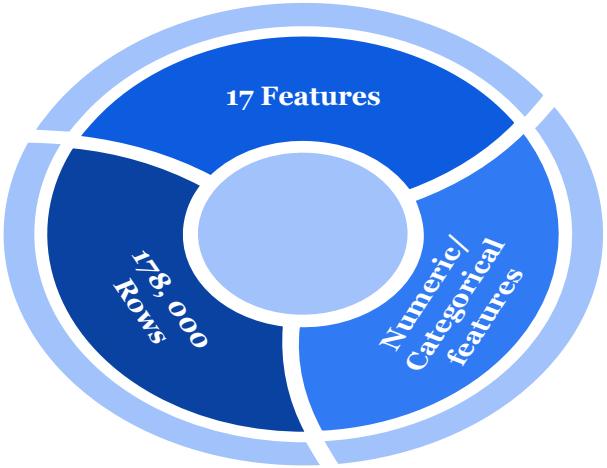
Ryerson University
The Chang School of Continuing Education

Initial Dataset

- Electric Vehicles in Ontario – By Forward Sortation Area
  - Total EVs by Forward Sortation Area (FSA)
- Insufficient Information
- High Bias
- Feature Engineering Constraints
- Lack of Diversity

Selected Dataset

link



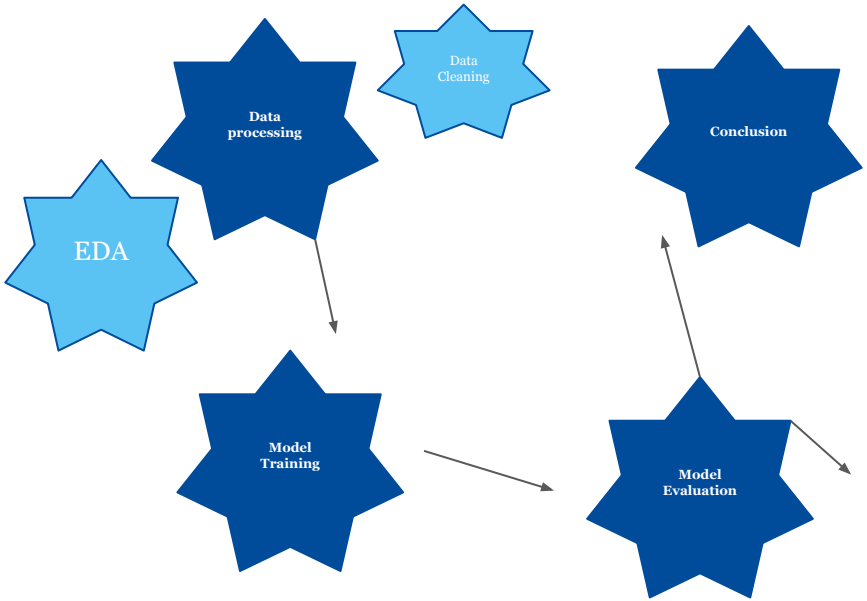
Electric Vehicle Population Data

VIN	County	City	State	Postal Code	Model Year	Make	Model	Electric Vehicle Type
Electric Range	Base MSRP	Legislative District	DOL Vehicle ID	Vehicle Location	Electric Utility	2020 Census Tract	CAFV	

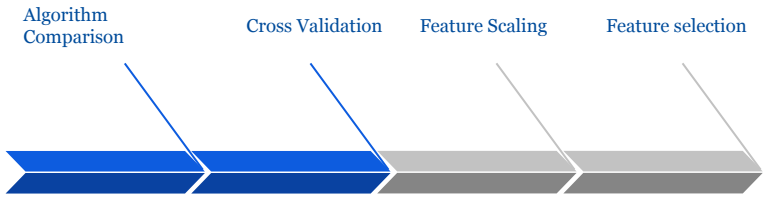
Ryerson University

The Chang School of Continuing Education

Approach

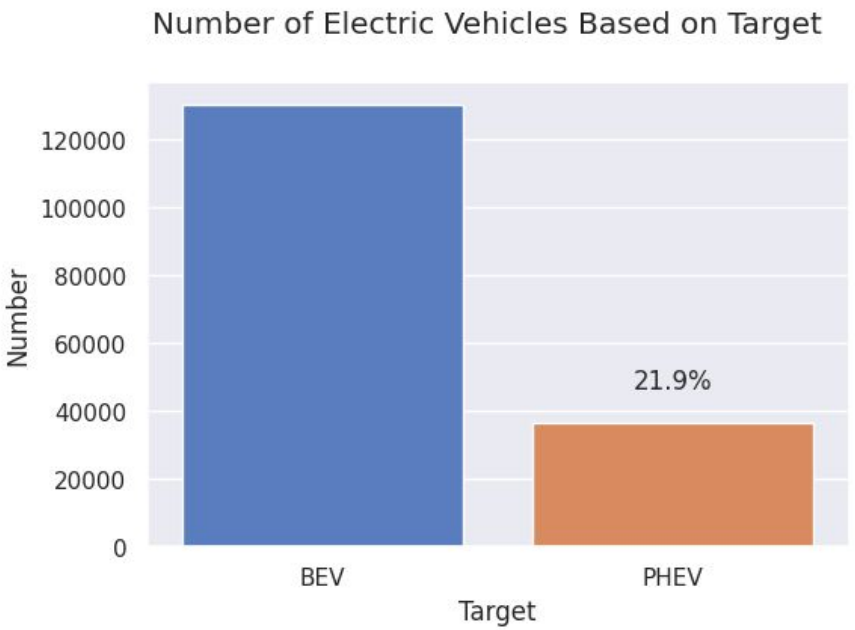


Naive Bayes
Logistic Regression
Random Forest
XGBoost

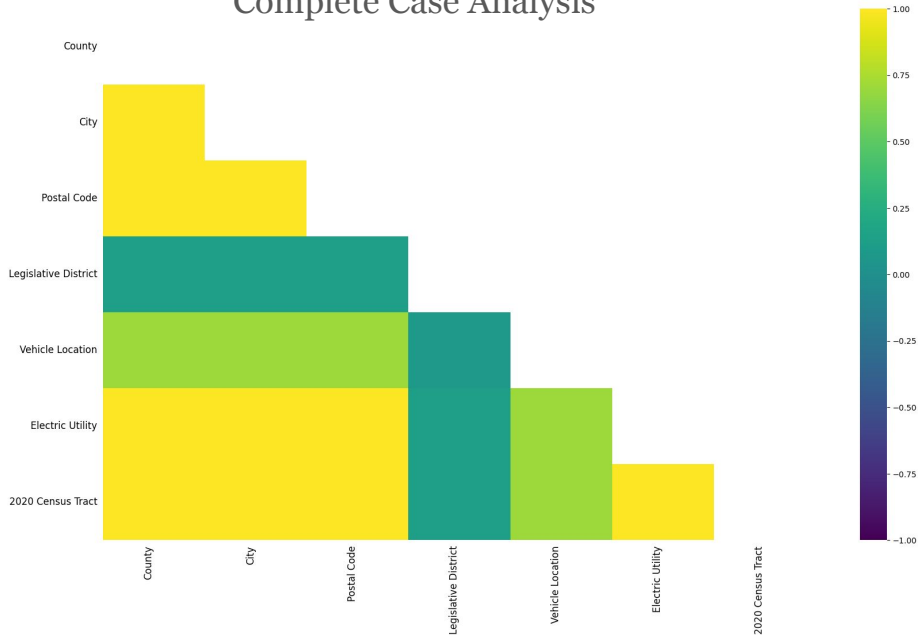


Ryerson University
The Chang School of Continuing Education

EDA



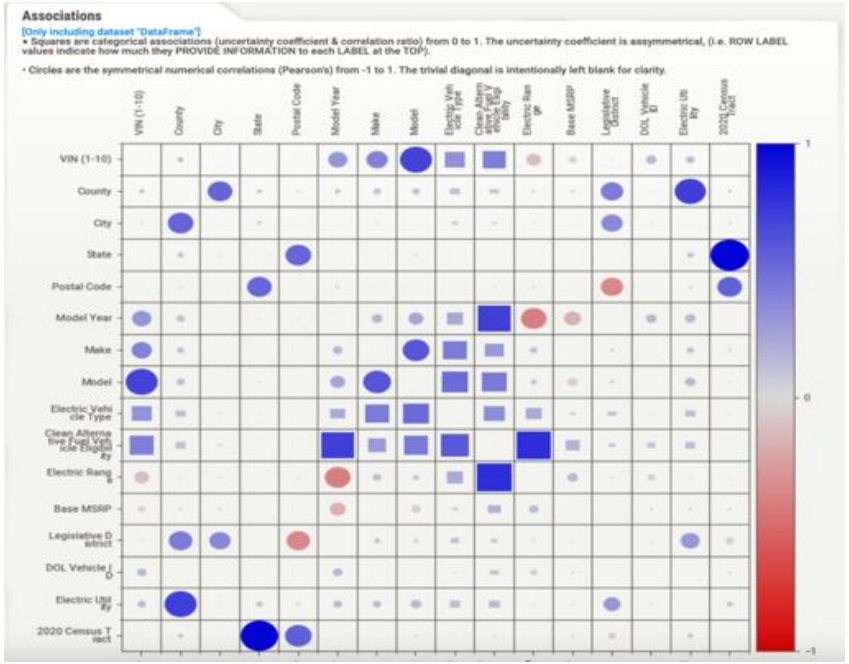
Handling Missing Values with Complete Case Analysis

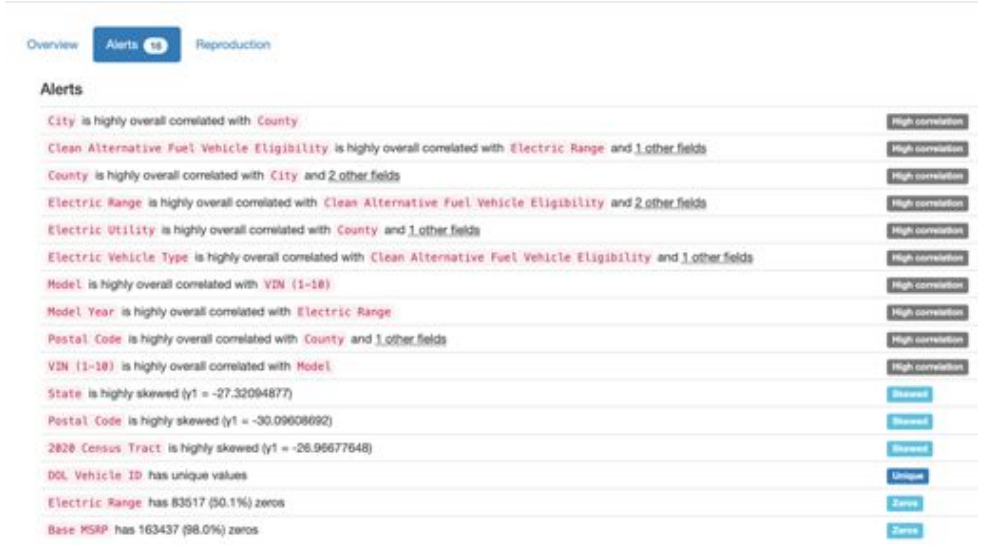


Ryerson University

The Chang School of Continuing Education

EDA using Sweetviz and Pandas Profiling Report

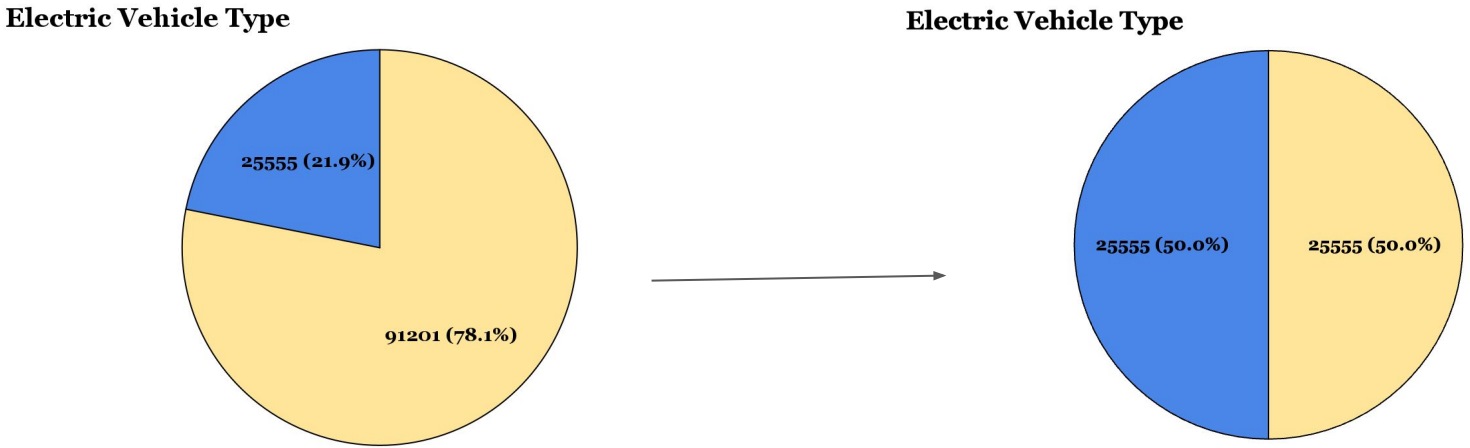




Ryerson University

The Chang School of Continuing Education

Class Imbalance



NearMiss approach is utilized due to its capacity to focus on selecting relevant majority class samples near the decision boundary, effectively reducing class imbalance while preserving vital information.

Initial Result

	Model	Accuracy	ROC-AUC
0	Random Forest	0.998801	0.999994
1	XGBoost	0.998941	0.999983
2	Naive Bayes	0.780491	0.727559
3	Logistic Regression	0.516097	0.514732

Feature Importance

	Feature	Importance
0	Electric Range	0.453681
1	Clean Alternative Fuel Vehicle Eligibility	0.196643
2	Model	0.160677
3	Make	0.052106
4	VIN (1-10)	0.040262

	Feature	Importance
0	State	3.712071e-08
1	Electric Utility	9.238881e-05
2	Base MSRP	3.686047e-04
3	Legislative District	7.104148e-04
4	Latitude	8.808175e-04

Final Result

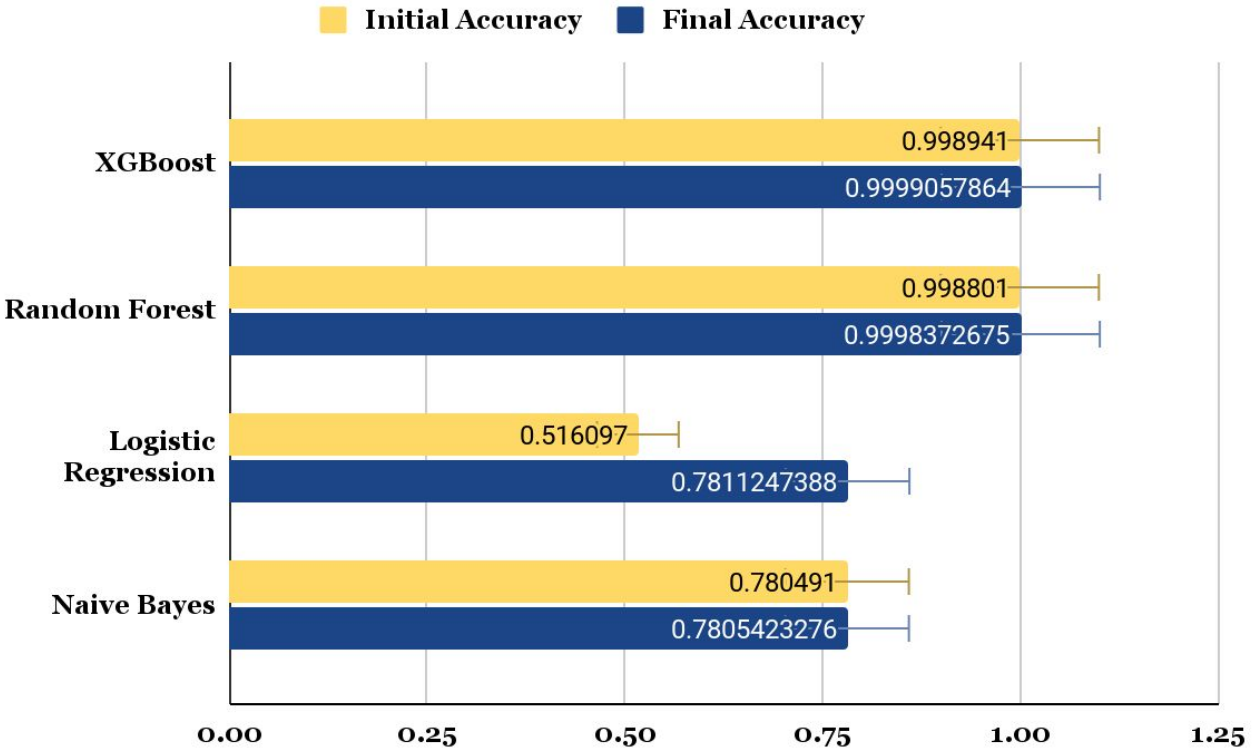
Cross Validation

5-fold cross-validation to evaluate the performance of all the machine learning algorithms.

Using 5-fold cross-validation strikes a balance between accuracy and computational efficiency, ensuring that the model has enough data for training and evaluation while also generalizing well to different subsets of the dataset.

	Model	Accuracy	Accuracy w/ C-valid	ROC-AUC	ROC-AUC w/C-valid
0	XGBoost	0.9913867183596794	0.9999057864263935	0.994322349043133	0.9998692774101849
1	Random Forest	0.9820739822938108	0.9998372674637706	0.9884270113907274	0.9998676679820891
2	Logistic Regression	0.7811307180399288	0.781124738771455	0.5	0.5
3	Naive Bayes	0.7804112792022223	0.7805423275891603	0.5006567895225318	0.5006975238197879

Model Evaluation



Ryerson University

Education

Interpretation

- After feature selection (drop), feature scaling XGBoost performed with the greatest accuracy of 0.99139

- After Cross Validation XGBoost accuracy increased to 0.9999059

Conclusion

- Can we predict the type of Electric Vehicle?
- BEVs vs PHEVs
- Which Machine Learning yield the most efficient and accurate result?
- Does Cross Validation within ML yield improve results?
- What can we do to improve our findings?
- Electric Vehicle distribution by country?
- Most common make?

- Using Machine Learning it is possible to predict the type of Electric Vehicle and its distribution
- XGBoost and Random Forest are the best performed among the ML algorithms that were selected.
- C-Validation increased all the ML algorithms
- Hypertuning, feature selection, handling categorical features
- King County stands out as the epicenter of electric vehicle adoption
- Tesla - its impact in future production

Questions?