**Predicting Electric Vehicle Registrations in Ontario by Region**

**EVON**

Geoff Krause gkrause@dal.ca | Hariharasudhan Giridharan [hr925260@dal.ca](mailto:hr925260@dal.ca) | Ye Wang [yz644140@dal.ca](mailto:Wang-yz644140@dal.ca) | Xuelian Xi xl884022@dal.ca)

Faculty of Computer Science

CSCI 6409 – Winter 2024

Dr. Evangelos E. Milios

January 31, 2024

# **1 Introduction**

The electric vehicle (EV) market in Canada has been experiencing significant growth and change. As of the first nine months of 2023, out of 1,286,951 vehicles registered in Canada, 132,783 were either battery electric or plug-in hybrids. This trend is a part of Canada's broader efforts to combat climate change, with the government setting ambitious targets like phasing out gas vehicles by 2035 and mandating that EVs constitute 20% of auto sales by 2026. This regulatory push is seen as essential to achieve a critical mass of investment in EV supply chains and infrastructure, which in turn would bring about efficiencies of scale and make EVs more economically viable (Cousin B, 2023).

Statistics from 2021 further illustrate this upward trajectory. EV registrations in Canada increased from under 20,000 in 2017 to over 86,000 in 2021. By the first quarter of 2022, 7.7% of all vehicles registered were electric, and over 90% of these were in Quebec, British Columbia, and Ontario. Interestingly, registrations of non-electric vehicles fell by 23% from 2017 to 2021. This growing interest in EVs is also reflected in consumer attitudes, with 71% of Canadians considering an electric vehicle for their next purchase and 79% requiring a minimum range of 400 kilometers per charge (Nicole B, 2024).

# **1.1 Background**

The transition towards electric vehicles (EVs) presents a crucial step in addressing environmental concerns and promoting sustainable transportation. Organizations and businesses are increasingly interested in understanding and predicting EV adoption patterns. This interest stems from the need to develop effective policies and strategies that encourage EV usage and to anticipate the infrastructural demands associated with a growing EV market. Recent studies have shown that various demographic, technical, economic, and behavioral factors significantly influence EV adoption. For instance, (Egbue et al., 2017) highlighted the importance of demographic determinants and behavioral attitudes in influencing individual adoption decisions (Egbue et al., 2017). Similarly, (Chen et al., 2020) emphasized the interconnected influence of socio-demographics and behavioral factors on EV adoption interest (Chen et al., 2020)

# **1.2 Objective**

The primary objective is to leverage these insights to develop a robust predictive model that can accurately forecast EV adoption trends based on demographic and population statistics. This model aims to assist organizations in strategizing their approach towards promoting EVs and preparing for the infrastructural demands they entail.

# **1.3 Expected Results**

The implementation of a regression-based predictive model for EV adoption is expected to yield several benefits. These include better-informed policy decisions, optimized investment in EV charging infrastructure, and targeted marketing strategies. The model's predictions can guide businesses in aligning their services with the anticipated demand, thus facilitating a smoother transition towards a sustainable transportation future. Studies like Wang and Hewitt, 2019) have demonstrated the effectiveness of using logistic regression analysis to understand the relationship between EV adoption and factors like household income and technology usage (Wang & Hewitt, 2019).

# **2 Data**

The final dataset used for this project will be derived by combining data obtained from a variety of sources, describing the number of registered electric vehicles (EVs) in the province of Ontario (our target variable), broken down by geographical area, and other characteristics of these areas and their populations. Key data preparation steps will involve joining together disparate datasets using geographic identifiers, filtering the appropriate characteristics from these datasets for use as features, and possibly engineering features by converting certain continuous variables to categorical ones.

## **2.1 Dataset sources**

The target feature, the number of EVs registered per region, is obtained from the *Electric Vehicles in Ontario – By Forward Sortation Area - Q4 2023* dataset, produced by the Ontario Ministry of Transportation, and released in January 2024 as a simple CSV file. This was obtained from the Ontario Data Catalogue (<https://data.ontario.ca/dataset/electric-vehicles-in-ontario-by-forward-sortation-area>).

Population statistics regarding these same areas were obtained from the *2021 Census Profile* data from Statistics Canada. We downloaded the comprehensive data file in CSV format, selecting the “Forward sortation areas” option, to correspond to the EV data. Note that this file contains data for all of Canada, in a one-row-per statistic and area format, so it may require significant processing. The main CSV is supplemented with a data dictionary and a CSV with the starting row numbers for each geographical area.

Statistics Canada also provides data on commuting details, broken down by geographical region, in the *Main mode of commuting* dataset. This is derived from the long form census and is similarly available in CSV format.

Lastly, Statistics Canada also provides boundary files for FSAs, indicating their location, shape, and size. These ultimately originate with Canada Post, which operates the postal code system.

## **2.2 Dataset details**

The *Electric Vehicles* dataset consists of a single entry for each of the 550 Forward Sortation Areas (FSAs) in the province of Ontario, which correspond to the geographic region represented by the first three characters in the Canadian postal code system. Postal codes are more stable than census identifiers, which may change with each subsequent census; individuals are more likely to be able to identify their postal code, as it is part of their address, leading to it being more commonly available.

In addition to the FSA, a categorical variable, the *Electric Vehicles* dataset contains the total number of electric vehicles registered in the corresponding area, as well as subsets of this for battery-based and hybrid vehicles. All three values are continuous.

The *Census Profile* dataset contains entries for all 1,646 FSAs in Canada, each of which has one row for each of the 2,631 characteristics measured in the *Profile*, for a total of 4,330,627 rows/entries. Each entry contains several identifiers for the dataset, FSA, and characteristic, as well as the frequency and proportion of the characteristic among the area’s population, both in total and broken into male/female columns. The (categorical) FSA identifiers will be used to link this dataset to our target data, while the characteristics of interest (e.g., age, income, employment, housing) will be transposed into columns. The characteristics listed in this dataset are numerical and continuous.

In the *Main mode of commuting* dataset, each row in this dataset represents a mode of commuting within a geographical subdivision, with 11 such rows for each area. Columns for each of these characteristics include the total number of commuters by type, counts of commuters in various commute-time groupings, and the average commute time. Data is available regarding various levels of census subdivisions, from the country as a whole, down to the 5,124 Census Subdivisions; however, the data is not available at the FSA level. We may choose to select the larger divisions, Census Agglomerations (CAs) and Census Metropolitan Areas (CMAs) and assign these values to the smaller FSAs contained within them. Once again, the characteristics are largely continuous.

The FSA boundary file can be used to extract information regarding the total area and latitude of each of the 550 areas we are examining. These values are inherently continuous, though we may choose to re-engineer them into categorical variables representing region (e.g. “Greater Toronto Area”, “Northern Ontario”) or community type (e.g., “urban”, “suburban”, “rural”).

# **3 Data Science Problem**

The primary data science problem to be addressed by this project is the prediction of electric vehicle (EV) uptake by drivers within a given region, based on the specific characteristics of the region and its population, including population size and density; age, income, and education; commuting times and modes; and the north-south location (latitude) of the area. This is, in essence, a regression problem, with the continuous target variable being the number of EVs registered within a given region, normalized over the population. We plan to further explore the correlations of the features used in making these predictions, in order to better understand the geographic, economic, social, and transportation-related determinants of EV usage and their interplay. We may also explore variations in battery-based and hybrid EV numbers. Any discrepancies between our predictions and actual EV usage may be instructive in uncovering other factors affecting uptake of this technology and could be worth examining on a regional basis.

# **4 Functionality**

## **4.1 Minimum Functionality**

* Data Processing and Cleaning: Enables collection, integration and cleaning of data from various data sources.
* Descriptive Statistics Analysis: Provides basic data descriptive statistics such as mean, median, standard deviation, etc. to provide an initial understanding of the characteristics of the data set.
* Preliminary Trend Analysis: Implements preliminary trend analysis to identify and characterize trends in the number of EV registrations over time.

## **4.2 Expected Functionality**

* Multivariate Regression Analysis: Develop multivariate regression models to analyze the impact of different factors (e.g., demographics, geographic location, economic indicators) on EV registrations.
* Correlation Analysis: Perform correlation analysis between variables to identify the main factors influencing EV registrations.
* Model Evaluation: Cross-validate predictive models to ensure accuracy and reliability of predictions.

## **4.3 Bonus Functionality**

* Advanced Prediction Techniques: Explore and implement more advanced prediction techniques, such as machine learning algorithms, to improve prediction accuracy.
* Factor Importance Analysis: Evaluate the importance of different characteristic variables in EV registration prediction models.
* Geospatial Analysis: Perform spatial analysis using Geographic Information System (GIS) techniques to identify the impact of geographic location on EV registration.

# **5 Project Organization**

* Data preprocessing module: Responsible for collecting, cleaning, and integrating data from multiple data sources, including EV registration data, demographics, etc. Performs feature engineering, such as feature selection and transformation, to enhance the predictive power of the model.
* Statistical Analysis Module: Implement descriptive statistical analyses and correlation analyses to provide an initial understanding of the underlying characteristics of the dataset. Apply cross-validation methods to ensure the accuracy and generalization ability of the selected model across different datasets.
* Predictive Modeling Module: Develop multiple linear regression models for analyzing the relationship between multiple independent variables and EV registrations. Implement more complex models such as decision tree regression and random forest regression to handle nonlinear relationships and complex data structures. Optimize model performance using hyperparameter tuning techniques.
* Model Evaluation and Optimization Module: Evaluate model performance and make necessary adjustments and optimizations to improve prediction accuracy. Regularly review model results and adjust feature engineering and model parameters to cope with data changes.
* Report Generation Module: Generate reports containing key analytical results and inferences for evaluation and decision-making by project teams and stakeholders. Present machine learning model results in an easy-to-understand format to ensure transparency and interpretability.

The entire project will be developed using a modular and iterative approach to ensure flexibility and adaptability in the processing of data and analysis. Each module will be led by a dedicated team member, with regular meetings to ensure overall project consistency and progress.

**5.1 Possible Models:**

1, Multivariable linear regression

<https://en.wikipedia.org/wiki/Linear_regression>

2, Decision tree regression

<https://en.wikipedia.org/wiki/Decision_tree_learning>

3, Random forest regression

<https://en.wikipedia.org/wiki/Random_forest>

Note: In machine learning, decision trees/random forest are often used in classification problems, but they can also be adapted for regression problems, where the output of continuous values is predicted instead of category labels. This is called decision tree /random forest regression.

**5.2 Possible Measures:**

1. Feature engineering

Applicability: In the EV registration prediction project, feature engineering can be used to select the most influential variables, such as the income level and population density of the region. In addition, new, more informative features can be created by combining or transforming existing features, such as converting raw demographic data into metrics with more predictive value.

2. Cross-validation

Applicability: Cross-validation can be used to assess the performance of a selected model (e.g., multiple linear regression, decision trees, or random forests) in EV registration forecasting. This helps to ensure that the model makes accurate predictions across different regions and conditions.

3. Hyperparametric tuning

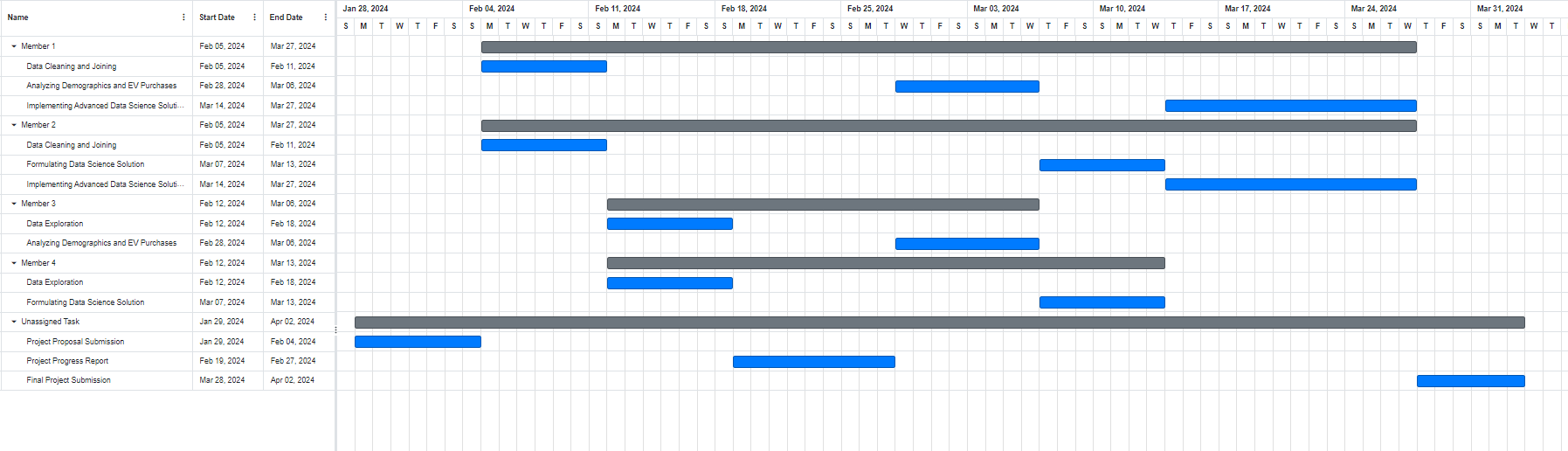
Applicability: For complex models (e.g., random forests), prediction accuracy can be improved by tuning hyperparameters (e.g., number of trees, depth of trees, etc.). Methods such as grid search or random search can be used to find the optimal combination of hyperparameters.

# **6 Project Timeline**

The project timeline is illustrated below with the help of a Gantt chart (Hariharasudhan, 2024). The unassigned tasks are meant to be covered by all the group members. Each task has its own timeline set with a starting and an ending date for each team member.

Important Deadlines to meet are as follows:

* Project proposal submitted – January 30
* Proposal presentation – Week of January 30
* Progress report – February 27
* Final implementation, presentation, & report due – First or second week in April



**References:**

Egbue, O., Long, S., & Samaranayake, V. A. (2017). Mass deployment of sustainable transportation: evaluation of factors that influence electric vehicle adoption. Clean Technologies and Environmental Policy, 19, 1927-1939.

<https://link.springer.com/article/10.1007/s10098-017-1375-4>

Chen, C. F., de Rubens, G. Z., Noel, L., Kester, J., & Sovacool, B. K. (2020). Assessing the socio-demographic, technical, economic and behavioral factors of Nordic electric vehicle adoption and the influence of vehicle-to-grid preferences. Renewable and Sustainable Energy Reviews, 121, 109692.

<https://www.sciencedirect.com/science/article/abs/pii/S1364032119308974>

Wang, Y., & Hewitt, E. L. (2019, October). Plug-in Electric Vehicle (PEV) Adoption in US Transport for Policy. In 2019 International Energy and Sustainability Conference (IESC) (pp. 1-21). IEEE.

<https://ieeexplore.ieee.org/abstract/document/8976770>

Cousins,B.( 2023, December). What does 2024 have in store for the EV industry? Bloomberg BNN.

<https://www.bnnbloomberg.ca/what-does-2024-have-in-store-for-the-ev-industry-1.2014925>

Blair, N. (2024, January). Electric Vehicle Adoption Statistics in Canada. Made in CA. <https://madeinca.ca/electric-vehicle-adoption-statistics-canada/>

Hariharasudhan. (2024, January). Gantt Chart. <https://drive.google.com/file/d/1fhqYzruDUoFbEK2s_pyFu3tl9PXkhwf5/view?usp=sharing>