Using IRS Data to Predict Electric Vehicle Count

Sri Lakshmi Chundru, Collin Real, Joaquin Ramirez, Seth Harris

## 1 Abstract

This study investigates the use of advanced machine learning models to predict vehicle count by zip code, utilizing a dataset enriched with socio-economic, demographic, and regional features. The primary aim is to accurately estimate vehicle counts, aiding in efficient resource distribution and urban planning. We implemented and evaluated several models, including Decision Tree, Random Forest, XGBoost, and HistGradientBoosting, focusing on their accuracy, interpretability, and computational efficiency.

Each model was trained and validated using cross-validation and a hold-out test set, with performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared used to assess their effectiveness. Among the models, the Random Forest model demonstrated superior performance, offering the best balance of predictive accuracy and model robustness. It effectively captured the complex interactions between the features and the target variable, outperforming the other models in terms of both accuracy and generalization.

The results highlight the Random Forest model’s effectiveness in handling diverse and complex datasets, making it a valuable tool for predicting vehicle counts in different zip codes. This study provides critical insights into the practical applications of machine learning in urban planning and resource management, particularly in the context of transportation and infrastructure development.

## 2 Introduction

The increasing adoption of electric vehicles (EVs) is a pivotal shift in the transportation sector, driven by growing environmental concerns, advancements in technology, and supportive government policies. Understanding the factors that influence the distribution of EVs is essential for developing targeted policies and infrastructure, such as charging stations and incentives. However, EV adoption is not uniform across regions; it varies significantly based on socio-economic and demographic factors. This study aims to explore these variations by predicting vehicle counts, particularly EVs, at the zip code level using a comprehensive dataset that combines electric vehicle registration data (n.d.-a) with individual income tax statistics from the Internal Revenue Service (IRS) (n.d.-b).

The IRS data offers detailed insights into various financial attributes of taxpayers by zip code, including income levels, deductions, and credits. By integrating this data with EV registration data, we can analyze the relationship between financial capacity, tax-related benefits, and vehicle ownership patterns. Such an analysis is crucial for identifying the socio-economic determinants of EV adoption, which can inform policymakers and urban planners.

The main objectives of this study are twofold: first, to develop predictive models that accurately estimate vehicle counts by zip code, and second, to identify and quantify the impact of key socio-economic factors on these counts. We employed a variety of machine learning models, including Decision Tree, Random Forest, XGBoost, and HistGradientBoosting, to achieve these objectives. These models were chosen for their ability to handle complex, structured data and provide interpretable results.

The findings of this study are expected to contribute valuable insights into the socio-economic factors driving EV adoption. By identifying the key predictors of vehicle counts, we can better understand the barriers to and facilitators of EV adoption in different communities. This knowledge is crucial for designing effective policies and interventions that promote sustainable transportation solutions. Furthermore, the study’s methodology, which includes rigorous data cleaning, feature transformation, and model validation, provides a robust framework for future research in this area.

## 3 Methods

In this study, we employed a variety of machine learning models to predict vehicle counts by zip code using socio-economic and demographic features extracted from IRS tax data. The data preparation, feature transformation, and model training processes are detailed below.

#### 3.0.1 Data Preparation

The dataset was compiled by merging electric vehicle registration data with IRS tax data, focusing specifically on zip codes within the state of Washington. This geographical focus provided a diverse socio-economic landscape, enhancing the robustness of our analysis. The dataset included features such as ‘Net premium tax credit amount,’ ‘Home mortgage interest paid amount,’ ‘Real estate taxes amount,’ and several others related to income, tax credits, and deductions.

To prepare the dataset for analysis, we first addressed data quality issues. We standardized column names using regular expressions, ensuring consistency across the merged dataset. This step was essential for preventing errors during data processing and analysis. We also applied data filtering to include only the relevant zip codes from Washington state, further refining our dataset for localized analysis.

#### 3.0.2 Feature Transformation

Given that many of the predictor variables exhibited skewed distributions, we applied a log transformation to normalize these features. This transformation was particularly necessary for financial features, such as ‘Salaries and wages amount’ and ‘Home mortgage interest paid amount,’ which spanned several orders of magnitude. The transformation formula used was log⁡(x+1)log(x+1), which effectively handled zero values and stabilized variance across the dataset. Normalizing the data in this manner helped improve the performance and interpretability of our machine learning models by reducing the impact of outliers and facilitating a more uniform distribution of feature values.

#### 3.0.3 Model Development

We implemented four different machine learning models: Decision Tree, Random Forest, XGBoost, and HistGradientBoosting. Each model was selected for its unique strengths in handling structured data and complex relationships between features.

#### 3.0.4 Model Training and Validation

Each model was trained on the processed dataset, with training and validation sets created using cross-validation techniques. Cross-validation helped ensure that the models generalize well to unseen data, preventing overfitting. We evaluated model performance using metrics such as Root Mean Squared Error (RMSE) and R-squared, which provided insights into the accuracy and explanatory power of each model.

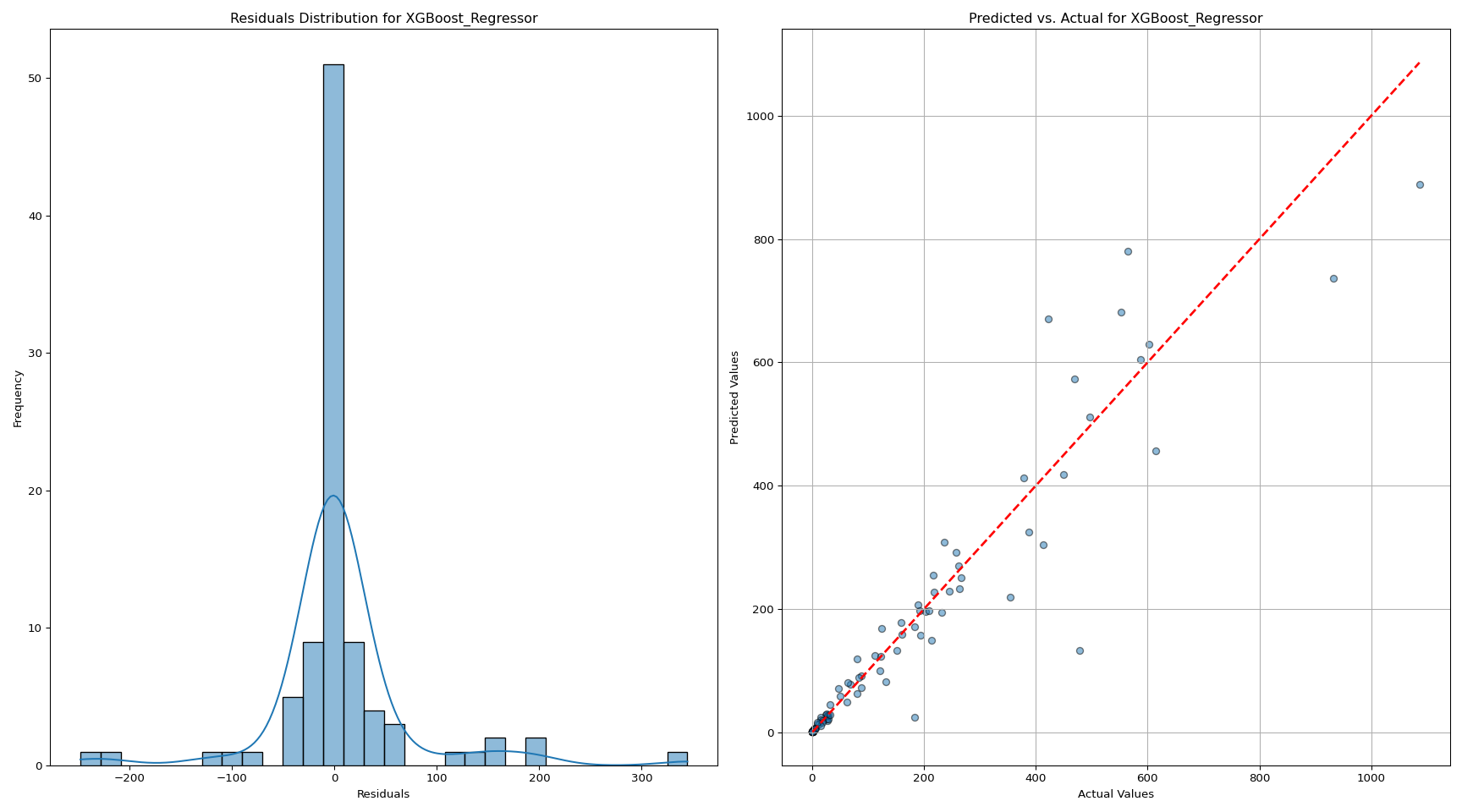
Overall, the methods employed in this study—from data preparation and feature transformation to model development and validation—were designed to rigorously test and compare the predictive capabilities of multiple machine learning models. These methods provided a comprehensive framework for analyzing the socio-economic factors influencing vehicle counts by zip code, with a particular emphasis on the adoption of electric vehicles.

## 4 Results

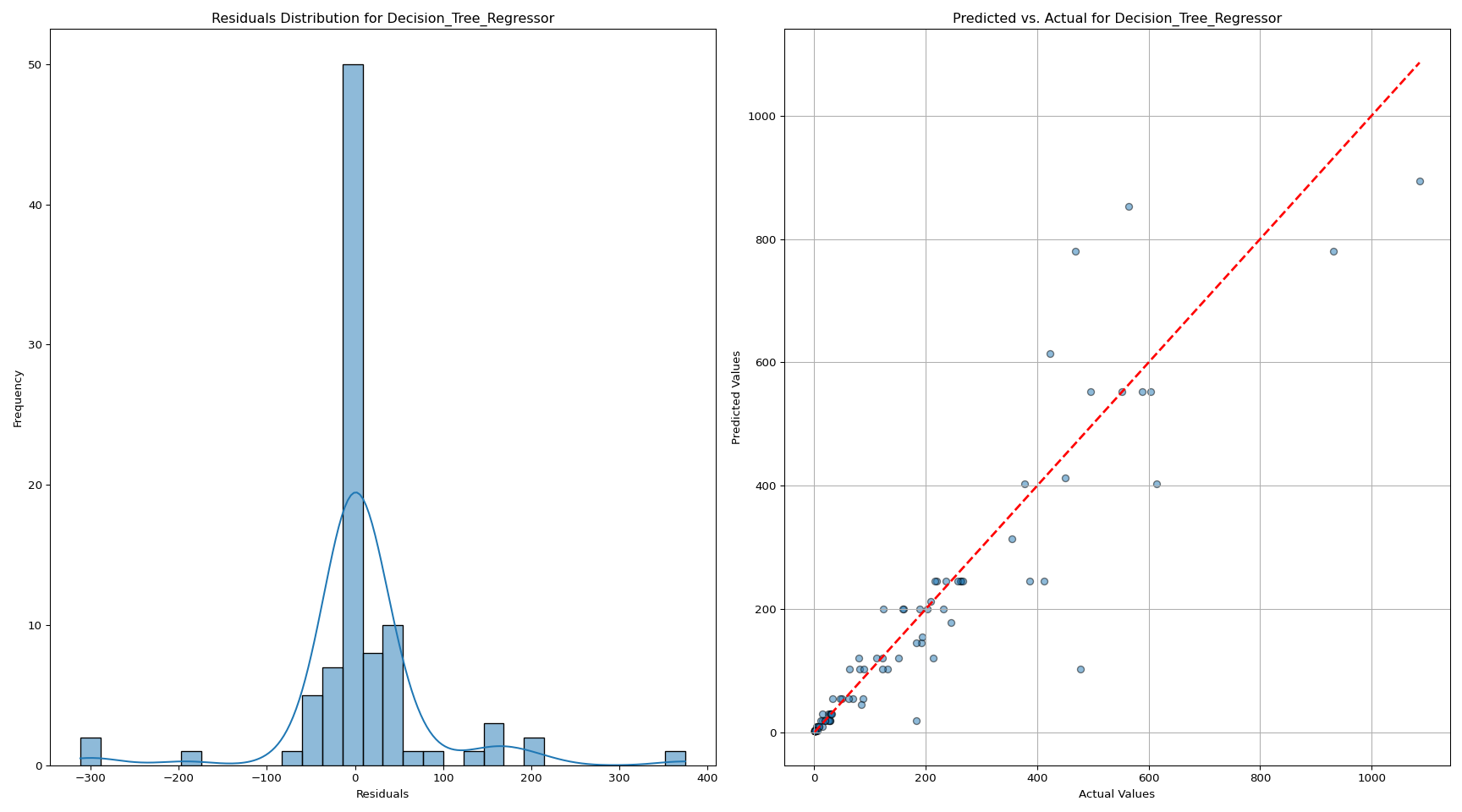
This section presents the comparative analysis of various machine learning models used to predict vehicle count by zip code. The models evaluated include XGBoost, Decision Tree, HistGradientBoosting, and Random Forest. The primary evaluation metrics considered were Root Mean Squared Error (RMSE) and R-squared, which provide insights into the accuracy and explanatory power of the models.

### 4.1 Model Performance Comparison

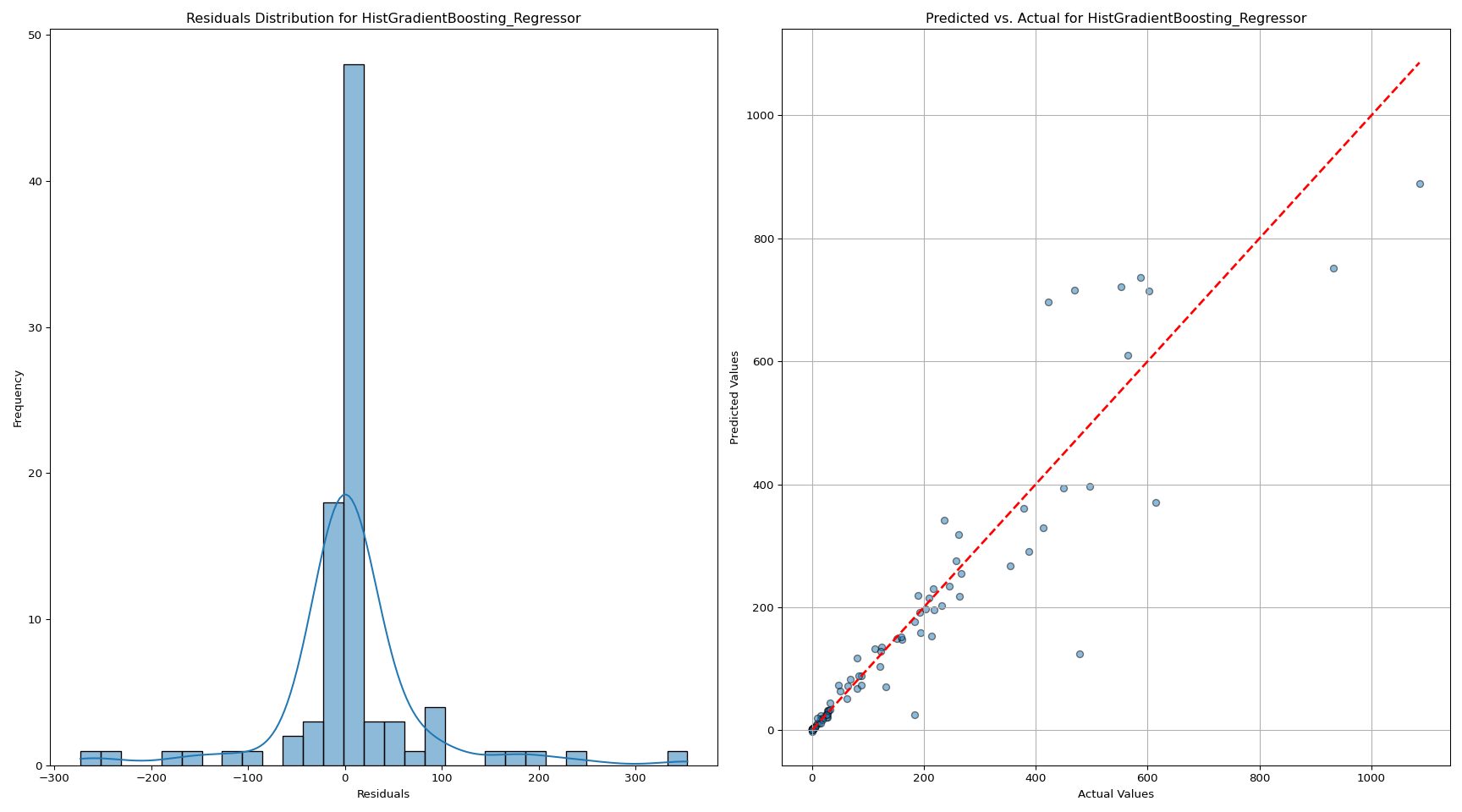
#### 4.1.1 XGBoost

* RMSE: 71.27
* R-squared: 0.91
* XGBoost demonstrated strong predictive capabilities, achieving a relatively low RMSE of 71.27. The model’s R-squared value of 0.91 indicates that it explains 91% of the variance in the vehicle count data, suggesting good model fit and predictive accuracy. 

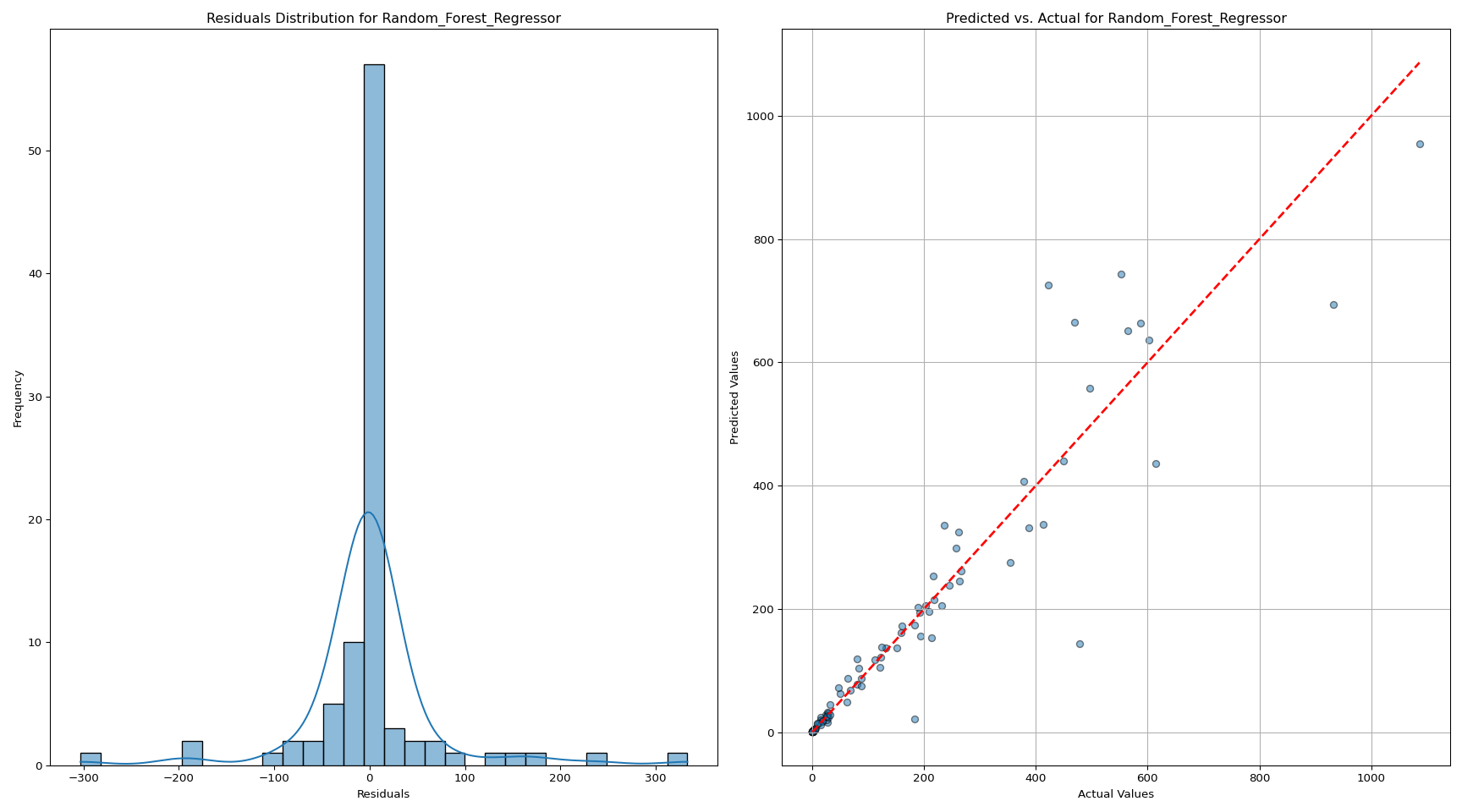
#### 4.1.2 Decision Tree

* RMSE: 100.77
* R-squared: 0.83
* The Decision Tree model had the highest RMSE among the evaluated models at 100.77, reflecting less precise predictions. Its R-squared value of 0.83 indicates that it explains 83% of the variance, making it the least accurate model in this comparison. 

#### 4.1.3 HistGradientBoosting

* RMSE: 83.42
* R-squared: 0.88
* The HistGradientBoosting model offered a moderate performance with an RMSE of 83.42 and an R-squared value of 0.88. While better than the Decision Tree model, it still lagged behind XGBoost and Random Forest in terms of prediction accuracy and variance explanation. 

#### 4.1.4 Random Forest

* RMSE: 62.16
* R-squared: 0.935
* The Random Forest model outperformed all other models, achieving the lowest RMSE of 62.16 and the highest R-squared value of 0.935. These metrics indicate that the Random Forest model not only provided the most accurate predictions but also explained 93.5% of the variance in vehicle counts, making it the most reliable model in this analysis. 

### 4.2 Best Model Selection

Based on the comparative analysis of the models, the Random Forest model emerged as the best performing model. It achieved the lowest RMSE and the highest R-squared, demonstrating superior accuracy and generalization capability compared to the other models. The Random Forest model’s ensemble nature allows it to effectively capture complex patterns in the data, reduce overfitting through bootstrapping, and handle interactions between predictor variables more adeptly. These attributes collectively contributed to its superior performance, making it the most suitable choice for predicting vehicle counts by zip code in this study.

The excellent performance of the Random Forest model suggests it is well-suited for applications requiring high prediction accuracy and robustness, particularly in urban planning and resource allocation scenarios where accurate vehicle count predictions are crucial.

## 5 Discussion

The objective of this study was to develop machine learning models capable of predicting vehicle counts by zip code, using various socio-economic and demographic features derived from IRS tax data. The analysis identified the Random Forest model as the most accurate and robust among the evaluated models, including XGBoost, Decision Tree, and HistGradientBoosting. In addition to the model performance, a detailed examination of the top features contributing to the model predictions provides valuable insights into the socio-economic factors influencing vehicle counts, particularly for electric vehicles (EVs).

### 5.1 Key Feature Analysis

The top 10 features contributing to the model predictions include:

Net Premium Tax Credit Amount: This feature represents the tax credit amount received by individuals under the Affordable Care Act. A higher tax credit may indicate lower income levels, suggesting a potential relationship between income and the ability to purchase vehicles, including EVs.  
  
Home Mortgage Interest Paid Amount: This feature reflects the interest paid on home mortgages, often correlating with homeownership and wealth. Areas with higher mortgage interest payments may have a higher concentration of middle to upper-income households, which can influence vehicle ownership patterns, including the adoption of EVs.  
  
Real Estate Taxes Amount: Real estate taxes are indicative of property values and, indirectly, wealth. Higher real estate taxes often correspond to higher property values, which may correlate with a higher prevalence of EV ownership due to greater financial capacity.  
  
Self-Employed Health Insurance Deduction Amount: This deduction is available to self-employed individuals who pay their health insurance premiums. It reflects the presence of self-employed individuals in a region, who may have different financial and vehicle ownership patterns compared to salaried employees.  
  
Salaries and Wages Amount: This feature directly indicates income levels from employment. Higher aggregate salaries and wages in a zip code suggest greater disposable income, which can facilitate the purchase of vehicles, including newer and potentially more expensive EVs.  
  
Limited State and Local Taxes: This feature represents the capped amount of state and local taxes that can be deducted from federal taxes. The deduction cap, especially in high-tax states, can impact disposable income and financial planning, potentially influencing vehicle purchasing decisions.  
  
Home Mortgage from Personal Seller Amount: This feature indicates transactions involving home mortgages from personal sellers. While less common, these transactions can indicate unique financial arrangements in certain areas, which might correlate with specific economic profiles affecting vehicle ownership.  
  
Individual Retirement Arrangement (IRA) Payments Amount: Contributions to IRAs reflect financial planning and the presence of disposable income. Areas with higher IRA payments may have wealthier populations, potentially increasing the likelihood of EV adoption due to greater financial resources.  
  
Self-Employed (Keogh) Retirement Plans Amount: Similar to IRAs, Keogh plans are retirement savings plans for self-employed individuals. Contributions to these plans suggest higher income and financial planning, which can correlate with higher rates of vehicle ownership, including EVs.  
  
Additional Medicare Tax Amount: This tax applies to high-income earners and indicates a population segment with substantial earnings. Higher additional Medicare tax payments correlate with higher income levels, potentially leading to increased EV adoption due to greater financial capacity.

### 5.2 Implications and Socio-Economic Insights

The analysis of these top features reveals a consistent pattern: higher vehicle counts, particularly for EVs, are associated with higher income levels, wealth indicators (such as mortgage interest and real estate taxes), and financial planning (such as retirement contributions). These factors suggest that regions with greater financial capacity are more likely to adopt EVs, reflecting broader trends in consumer behavior and market accessibility.

The emphasis on features related to self-employment and additional taxes for high earners highlights the impact of income diversity on vehicle ownership patterns. The presence of deductions and credits for self-employed individuals suggests that entrepreneurial regions may exhibit unique vehicle ownership trends, possibly favoring flexibility and lower-cost vehicles or, conversely, reflecting higher disposable income that supports EV adoption.

## 6 Conclusion

The Random Forest model’s ability to incorporate and weigh these diverse socio-economic features underscores its robustness and suitability for this type of predictive analysis. The insights gained from the top features not only enhance our understanding of the factors driving vehicle ownership but also offer valuable guidance for policymakers and urban planners aiming to promote EV adoption. Future research could expand on these findings by incorporating additional data sources, such as environmental factors or public transportation availability, to provide a more comprehensive view of the determinants of vehicle ownership and EV adoption.

## 7 Citations

Jain, A. (n.d.). Electric vehicle population [Data set]. Kaggle. Retrieved June 6, 2024, from https://www.kaggle.com/datasets/jainaru/electric-vehicle-population

Internal Revenue Service. (n.d.). SOI tax stats - Individual income tax statistics - ZIP code data (SOI). Retrieved June 16, 2024, from https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi

(n.d.-b). Retrieved from <https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi>

(n.d.-a). Retrieved from <https://www.kaggle.com/datasets/jainaru/electric-vehicle-population>