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

As electric vehicles (EVs) become increasingly prevalent, understanding their distribution and associated factors at a granular level, such as by zip code, is crucial for infrastructure planning and policy development. In this study, we sought to predict vehicle counts, with a particular focus on EVs, across various zip codes using a comprehensive dataset combining electric vehicle registration data with IRS tax data. This integration provides a robust foundation for analyzing socio-economic and demographic influences on vehicle distribution.

The dataset was curated by merging electric vehicle registration records, which provide detailed counts of EVs by zip code, with IRS tax data, offering insights into the income levels and financial characteristics of the residents in each zip code. This integration was facilitated by matching records based on zip codes, creating a unified dataset that captures a wide array of features, including income levels, tax filings, and vehicle counts.

Before analysis, significant data cleaning was necessary to ensure the dataset’s accuracy and consistency. The initial data contained various inconsistencies, particularly in column names and formats. We employed regular expressions, a powerful tool for text manipulation, to standardize column names, correcting typographical errors, and ensuring uniformity in naming conventions. This step was crucial for seamless data manipulation and analysis, allowing for accurate merging and comparison of data across different sources.

Furthermore, to focus our analysis on a specific geographic region and manage data complexity, we filtered the dataset to include only records from the state of Washington. This state was chosen due to its diverse socio-economic landscape and significant adoption of electric vehicles, providing a rich dataset for analysis. The filtering process involved selecting rows where the state code matched Washington’s, ensuring that the analysis remained relevant to the regional context.

By combining these datasets and refining them through meticulous cleaning and filtering, we prepared a comprehensive and reliable dataset. This dataset serves as the foundation for training machine learning models aimed at predicting vehicle counts by zip code, with a particular emphasis on understanding the factors influencing the distribution of electric vehicles.

## 3 Methods

To address skewness and non-normal distributions in our predictor variables, we applied a log transformation to each feature in the dataset. This transformation helps to stabilize variance, minimize the impact of outliers, and create a more normally distributed set of predictor variables. Specifically, we used the natural logarithm function, applying the transformation as log⁡(x+1)log(x+1) to handle zero values gracefully. This preprocessing step was crucial for enhancing the performance and interpretability of our machine learning models.

### 3.1 Model Building

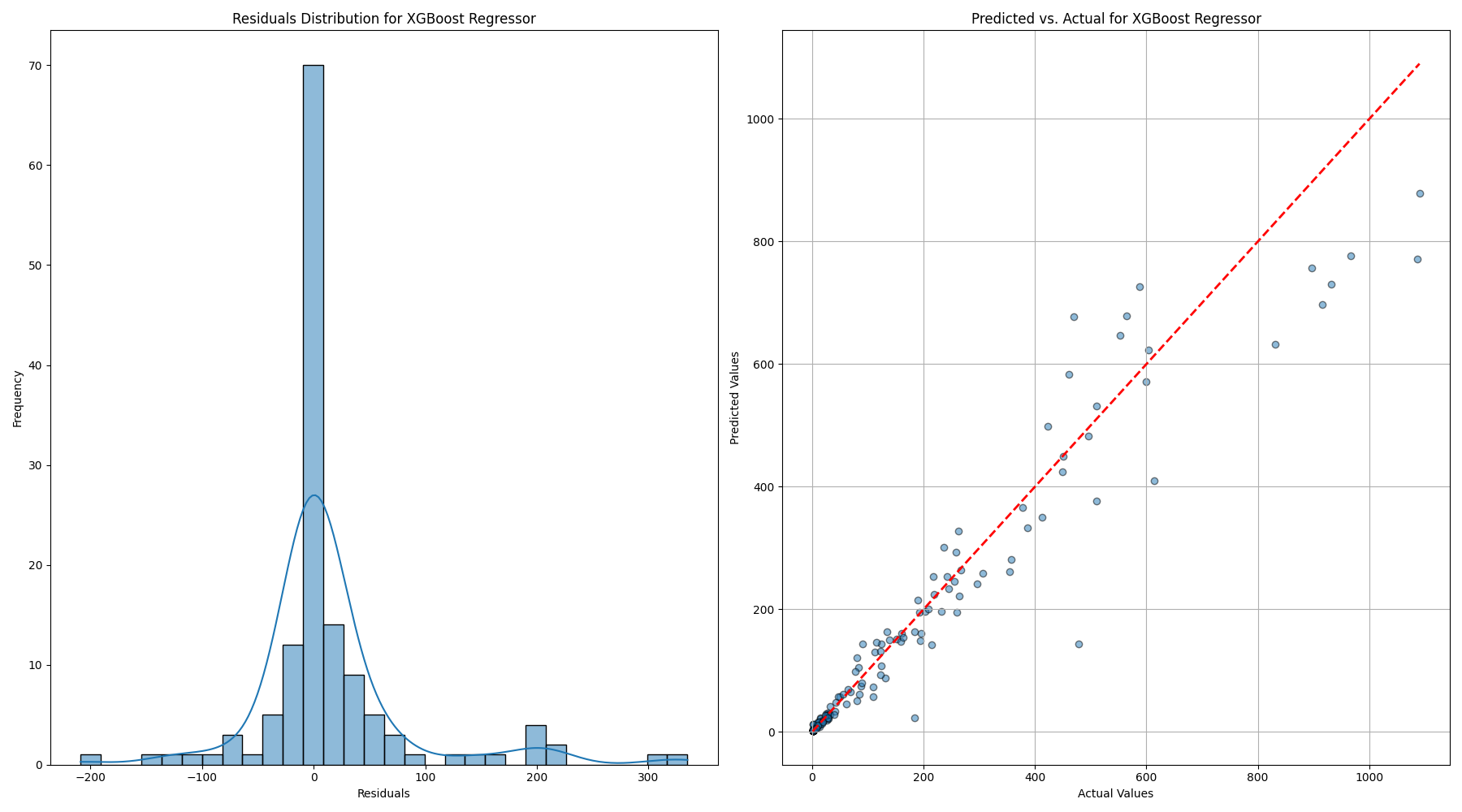
1. Decision Tree Regressor The Decision Tree Regressor is a simple yet powerful model that splits the data based on feature values, recursively partitioning the data into subsets with similar outcomes. For this model, we set parameters such as the maximum depth of the tree and the minimum number of samples required to split a node. We performed hyperparameter tuning to find the optimal tree depth and the minimum number of samples per leaf, balancing model complexity and overfitting.
2. Random Forest Regressor The Random Forest Regressor builds upon the Decision Tree model by constructing a multitude of decision trees during training and outputting the mean prediction of the individual trees. This ensemble method improves predictive accuracy and controls overfitting. We specified key parameters such as the number of trees in the forest (n\_estimators), the maximum depth of each tree, and the maximum number of features considered for splitting at each node. We also used bootstrap sampling, allowing each tree to be trained on a random subset of the data, enhancing model robustness.
3. XGBoost Regressor XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting designed for speed and performance. It sequentially builds decision trees, where each subsequent tree aims to correct the errors of the previous ones. For XGBoost, we tuned hyperparameters such as the learning rate (eta), the maximum depth of trees (max\_depth), the number of boosting rounds (n\_estimators), and regularization parameters (lambda and alpha) to prevent overfitting. Additionally, we used subsampling of rows (subsample) and columns (colsample\_bytree) to introduce variability in the model, further preventing overfitting.
4. HistGradientBoosting Regressor The HistGradientBoosting Regressor, part of the scikit-learn library, is a fast and scalable implementation of gradient boosting using histogram-based binning. This method is particularly efficient for large datasets. We configured parameters including the maximum number of iterations (max\_iter), learning rate, maximum depth of the trees, and the number of bins (max\_bins) for histogram binning. The model also included support for early stopping, which halts training when the validation loss stops improving, thus preventing overfitting and reducing training time.

## 4 Results

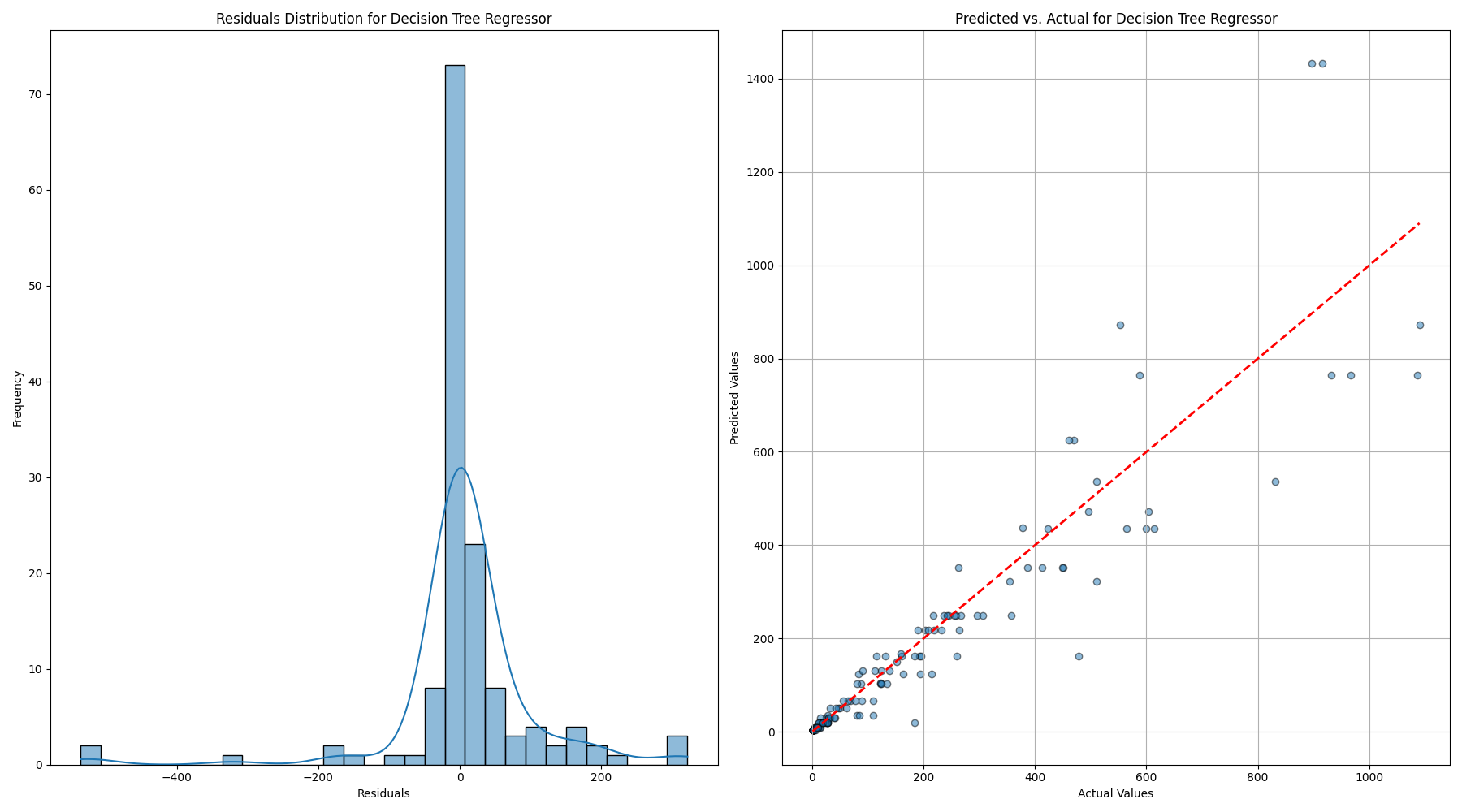
This section presents the comparative analysis of various machine learning models used to predict vehicle count by zip code. The models evaluated include XGBoost, Decision Tree, HistGradientBoosting, and Random Forest. The primary evaluation metrics considered were Root Mean Squared Error (RMSE) and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

### 4.1 Model Performance Comparison

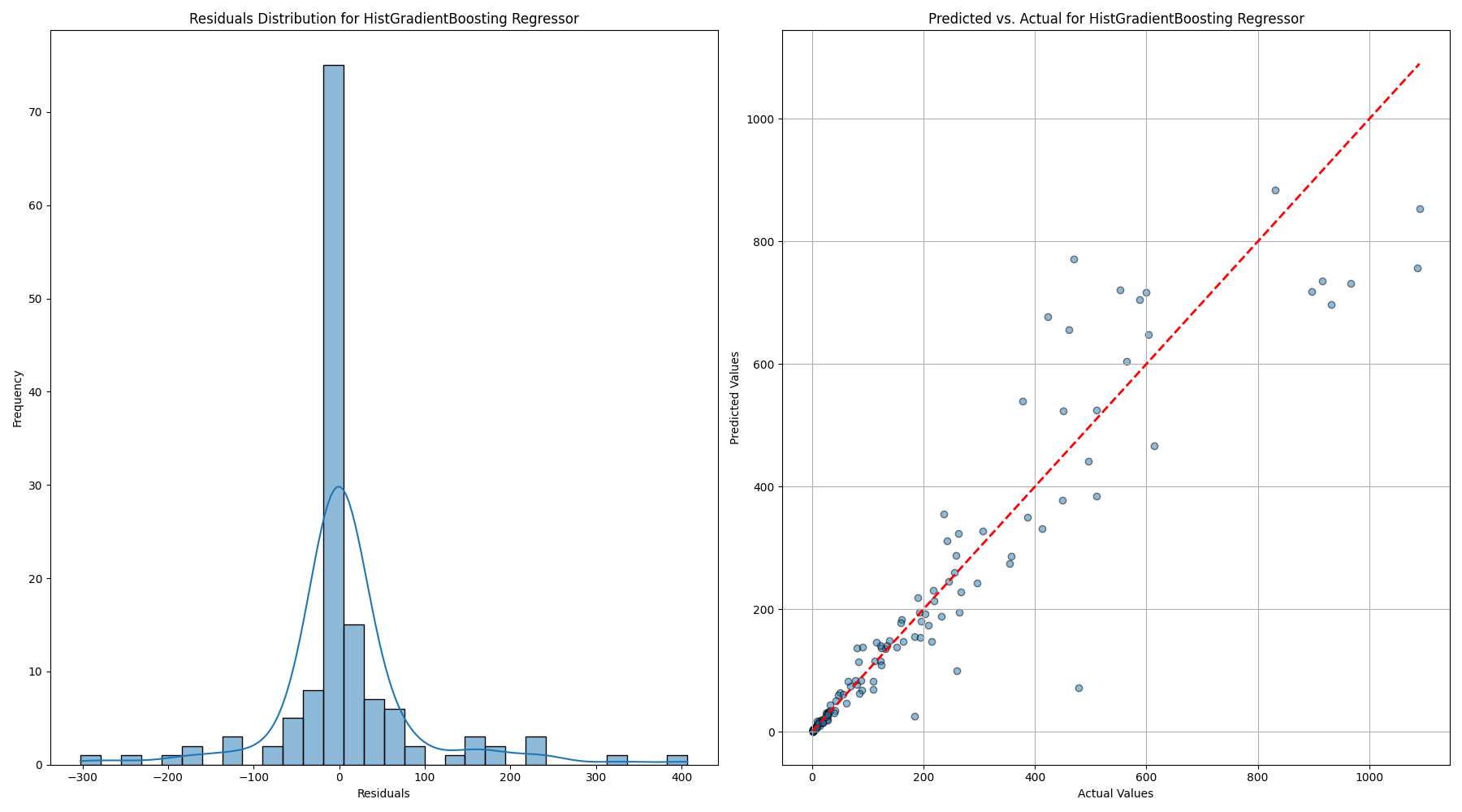
#### 4.1.1 XGBoost

* 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. 

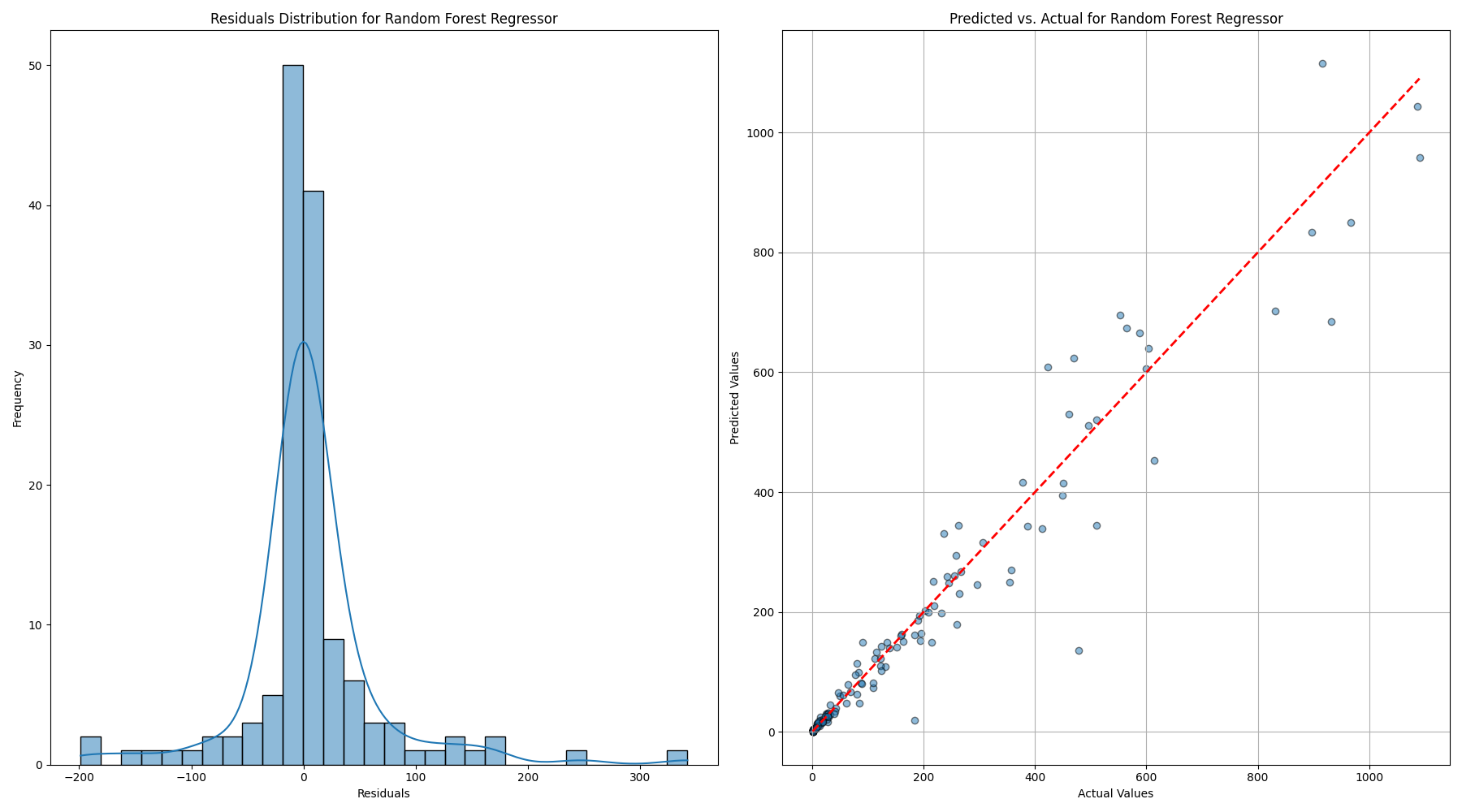
#### 4.1.2 Decision Tree

* 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