Report

Project: E-cars Taxation Analysis

Asta Björk
Tuula Jakobsson Peralta
Matias Brax

Course: Introduction to Data Science 2024

Table of contents

Introduction **3**
Data collection & preprocessing **3**
Exploratory data analysis & visualizations **4**
Learning task & approach **5**
Results **5**
Conclusion & discussion **6**

Introduction

The motivation for the project topic originated from an interest in studying the relationship between car taxation and the number of electric cars in Finland. Especially in light of climate change and sustainable growth, we are concerned about how the planned increase in the vehicle tax for electric cars from 2025 would affect the willingness to buy electric cars in the future.

The vehicle tax base is, and has been, diverse over the years. The elements related to vehicle taxation are for example price, usage, emissions and driving power of the vehicle, both directly and via value added taxes. Since 1994, the most significant tax rates related to vehicles are vehicle tax and car tax. The latter is usually paid only once, when the car is registered or commissioned for the first time in Finland. Due to its repetitive nature, we included the first mentioned in our study. Vehicle tax is collected annually from registered and operational vehicles. The amount of vehicle tax is based on the driving force (such as petrol, diesel, electric or hybrid) and carbon dioxide emissions.

The historical data of the review include the vehicle tax revenue collected by the state each year, the number of gasoline cars and, in the shorter term, the number of electric cars as well. The review period of this data is the years 1990 – 2023. We start forecasting the annual registrations of electric cars from 2025. The forecast has been corrected with the cost of living index in our model. By adding other factors affecting the number of electric cars as variables, the predictive model could become very complex.

Data collection & preprocessing

For our data collection, the initial focus was on gathering essential data related to the number of newly registered cars and the total amounts of car tax collected. As the project evolved, we recognized the need to account for inflation and other economic factors, leading us to include the cost of living index. This would enhance our analysis and predictions by adding an inflationary context. We relied on the public APIs of Statistics Finland, known for its comprehensive datasets. The platform also provided convenient JSON queries, allowing us to retrieve the required data efficiently.

Before starting the actual data collection, we conducted preliminary research on the availability and scope of Statistics Finland's data. We found that the earliest records for newly registered cars dated back to 1990. As a result, we set the historical range for our analysis from 1990 to 2023, ensuring consistency across all data queries. We chose Jupyter Notebook as the working environment to streamline our analysis and integrate other parts of our project seamlessly.

For the newly registered cars, we filtered the JSON queries to extract data specifically for gasoline and electric vehicles. In terms of amounts of tax data, we focused on car and vehicle taxes, filtering for both gasoline vehicle taxes and e-car taxes in Finland. Since this data was reported annually, the volume was relatively small. The cost of living index, however, was collected as monthly data, which provided a more granular view of economic conditions during the study period.

Exploratory data analysis & visualizations

At the beginning of the data analysis we used non-graphical methods by saving the data in JSON format. After exploration of the structure of the JSON, we acquired the right columns to and constructed dataframes utilizing pythons library *pandas* to explore the data. Figure 1. combines two JSON-arrays of data and is printed to non-graphical terminal for analysis.

	Year	Gasoline-cars	E-cars
0	1990	131830	0
1	1991	87997	2
2	1992	65095	1
3	1993	51465	0
4	1994	63455	0
5	1995	74386	0
6	1996	82878	2
7	1997	89225	0
8	1998	106025	0
9	1999	114295	3
10	2000	108273	0
11	2001	91412	0

Figure 1. Combined car amounts in a dataframe

After confirming correctness of the dataframe we made some graphical visualizations to find some insights about the structure of the data. Figure 2. shows the trend of newly registered gasoline and e-cars. There is clear trend of downslide around 1993 and 2008 in gasoline cars, and at the current time. E-car amount has been grown fast since 2021, almost meeting the gasoline car amounts. However, the gathered tax amount of combined vehicle & motor vehicle tax has been grown steadily to 2018, and is now slowly decreasing.

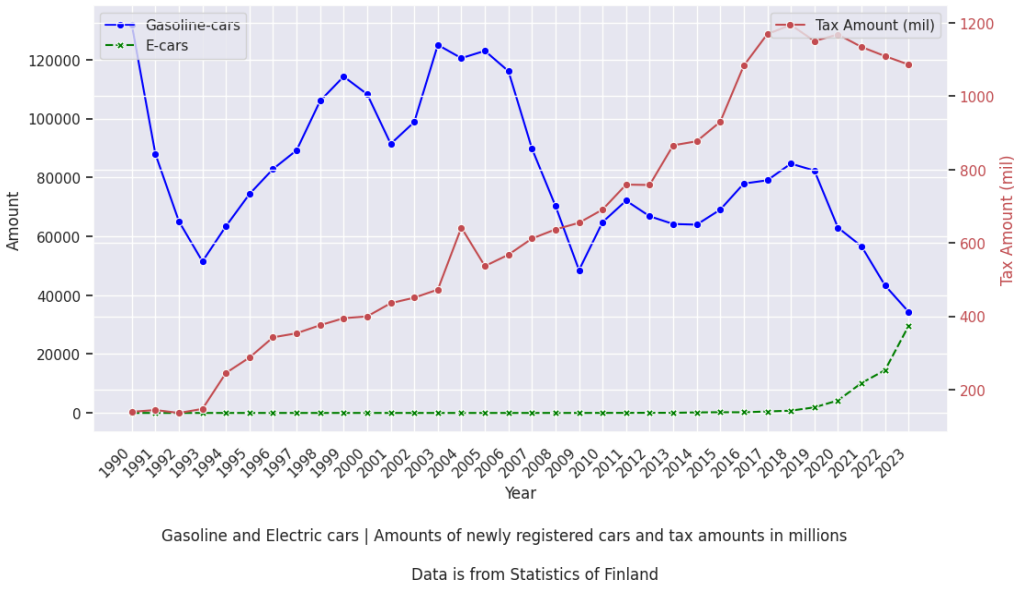

Image 2. Visualization of registered gasoline and electric cars and vehicle tax in Finland 1990-2023

Learning task & approach

The primary goal is to predict future e-car sales based on historical data. The key problem can be defined as the number of e-cars sold annually as the target variable and the historical data on tax rates for combustion engine cars, e-cars, historical sales of both types of cars, and, in the extended model, the cost of living in Finland as the independent variables. The task is to forecast the change in e-car sales after the tax-free period ends in 2025, when new tax rates may influence consumer choices. By analyzing this relationship, the model aims to provide insights into future sales trends based on changing tax (and economic conditions).

We applied two types of regression models to predict the number of e-cars sold annually from 2025 to 2034. Linear regression model assumes a linear relationship between input features and the target variable (e-car sales). Polynomial regression model explores potential non-linear relationships by fitting higher-degree polynomials to the input features. More complex models like Polynomial Regression could also be explored as the relationship between tax rates and sales proved to be non-linear. The data was split into training and test sets to validate model performance. Recent data may be given higher importance through weighting, as it could better capture future trends and improve the prediction accuracy. This approach aims to build an interpretable model that can accurately predict the shift in e-car sales based on changes in tax policy.

Both models (linear regression model and polynomial regression model) were evaluated based on three key metrics: Mean Squared Error (MSE), which measures the average squared difference between predicted and actual values (lower MSE indicates better model performance), the R^2 Score, which indicates the proportion of variance in the target variable explained by the model (R^2 values range from negative infinity to 1, where values close to 1 suggest good model fit), the accuracy, which measures the proportion of correctly predicted outcomes, though it's less relevant for regression models since the target variable is continuous.

Results

The initial evaluation metrics for both models were as follows: Linear Regression Model: MSE: 21,678,381.39, R^2: -721.05, Accuracy: 0.00. Polynomial Regression Model: MSE: 14,980,727.59, R^2: -497.59, Accuracy: 0.00

Both models have extremely high MSE values, suggesting that the predicted values are far off from the actual sales values. The polynomial model performs slightly better than the linear model but still shows poor prediction accuracy. The R^2 scores for both models are highly negative, which is a strong indicator of poor model fit. Negative R^2 values imply that the model performs worse than a simple horizontal line (mean prediction), meaning it fails to capture the underlying relationship between tax rates and e-car sales. Accuracy in these regression models is 0, which indicates that the models do not make any exact predictions matching the actual data. Since regression predictions are continuous, this metric is less meaningful in this context.

The model only considers tax rates and past car sales, which may not fully capture all relevant factors influencing e-car purchases, such as consumer preferences, technological advancements, and broader economic conditions. The linear regression model may be too simplistic to capture the complex,

non-linear relationships between tax rates and e-car sales. Even though the polynomial regression attempts to account for non-linearity, it still struggles to capture these relationships effectively.

To improve the model's predictive performance, cost of living data was introduced as an additional independent variable. The cost of living can be a significant economic factor affecting consumer decisions, particularly large purchases like cars. After incorporating cost of living, both models show a decrease in MSE. The linear regression model's MSE decreased from approximately 21 million to 15 million, while the polynomial regression model's MSE decreased significantly from around 15 million to 3 million. This indicates that adding the cost of living as an input feature helped improve the accuracy of the predictions, particularly for the polynomial model:

Linear Regression Model with Cost of Living: MSE: 15,133,227.22, R^2: -503.05, Accuracy: 0.00, Polynomial Regression Model with Cost of Living: MSE: 3,029,910.00, R^2: -99.92, Accuracy: 0.00.

While there is an improvement in the R^2 scores, both models still have negative R^2 values, indicating that they continue to perform worse than a baseline mean prediction. However, the polynomial regression model shows a marked improvement in R^2, moving from -497.59 to -99.92, suggesting it is getting closer to explaining some of the variance, though it remains far from optimal. The accuracy remains at 0.00 for both models. As mentioned earlier, this metric is less relevant for regression tasks since it measures exact matches between predicted and actual values, which is not expected for continuous variables like car sales.

The addition of cost of living as an input variable has clearly improved the performance of both models, particularly the polynomial regression model. The MSE and R^2 values show notable improvement, but the overall performance is still suboptimal, with negative R^2 values indicating that the model does not adequately explain the variance in e-car sales. Therefore, the model's performance improved when cost of living data was included, highlighting the importance of considering broader economic factors. However, the continued poor R^2 scores and MSE values suggest that the model still struggles to fully capture the relationships between taxes, cost of living, and e-car sales.

Some suggestions for potential improvements: Future models could benefit from the inclusion of more relevant factors, such as government subsidies, technological advancements, or changing consumer preferences toward electric vehicles. Also, more sophisticated models, such as Random Forests, Gradient Boosting, or Neural Networks, could be explored to capture complex, non-linear relationships more effectively. In addition, creating new features that combine or transform existing variables (e.g., interaction terms between taxes and cost of living) could improve the model's ability to capture subtle patterns in the data. While the current model shows some improvement with the addition of cost of living data, there remains significant room for enhancement to accurately predict future e-car sales.

Conclusion & discussion

The observation that both Linear Regression and Polynomial Regression models predict that more e-cars will be sold with rising costs of living and higher taxes can seem counterintuitive. However, several factors may explain this outcome.

Economic Theory of Demand: As the cost of living rises, consumers' disposable income might also increase. If consumers perceive e-cars as premium or environmentally friendly options, they may be inclined to invest in e-cars despite higher costs. Higher taxes may correlate with better public services,

infrastructure, or income levels in some contexts, making consumers more likely to afford and choose e-cars.

Government Incentives: Higher taxes on gasoline or combustion engine vehicles may be accompanied by subsidies or tax breaks for purchasing electric vehicles (tax incentives for e-cars). Thus, while taxes are higher, incentives could make e-cars more appealing and financially accessible. Rising taxes can be used to fund public infrastructure, including charging stations for e-cars, making them more practical and convenient for consumers.

Consumer Awareness and Preference: As awareness of climate change and environmental issues increases, more consumers may prefer e-cars regardless of rising costs. Higher living costs might motivate consumers to seek more sustainable transport options, especially in urban areas. Over time, consumers might recognize that e-cars could lead to lower operational costs, making them an attractive alternative despite higher upfront costs.

Market Dynamics: As demand for e-cars rises, manufacturers may respond by increasing production and lowering prices due to economies of scale. This could result in higher sales volumes even in challenging economic conditions. The growth of the e-car market often brings a broader range of models at various price points, catering to different consumer segments. As more affordable options enter the market, sales may increase, even with higher taxes.

Behavioral Economics: Higher taxes on gasoline cars could act as a nudge, encouraging consumers to consider e-cars as a viable alternative. This behavioral shift may not be fully captured in a traditional demand model but can influence consumer choices. As more people adopt e-cars, social norms may shift, making e-cars more desirable and socially acceptable. This could lead to increased sales even in a higher-cost environment.

When interpreting the model's results, it’s crucial to consider the following: First, the relationships between features (e.g., how taxes interact with cost of living) can lead to complex outcomes. For instance, the impact of taxes on gasoline might be significant enough to drive consumers toward e-cars, even as costs rise. Second, the model captures historical data and trends, but it may not fully account for future shifts in consumer behavior or policy changes. Models are simplifications of reality, and unexpected factors can influence outcomes.

In summary, while the model suggests that higher costs of living and taxes may lead to more e-car sales, the explanation lies in various economic, social, and behavioral factors. The model reflects a complex interaction of consumer preferences, government policy, and market dynamics that can drive the transition towards electric vehicles, even under challenging economic conditions. The two models do not adequately explain the variance in e-car sales. The addition of cost of living as an input variable has clearly improved the performance of both regression models, particularly the polynomial regression model. Therefore, the fact that the model’s performance improved when cost of living data was included, highlights the importance of considering broader economic factors. However, the continued poor R^2 scores and MSE values suggest that the model still struggles to fully capture the relationships between taxes, cost of living, and e-car sales.