Project 1: Tesla Supercharging Stations Location Prediction

Author: Venkat Jagadeesh Jampani

DSC680 Applied Data Science

3/31/2024

Table of Contents

Business Problem

Certainly, when individuals are considering purchasing an Electric Vehicle (EV), there are several factors that commonly influence their decision-making process.

Range Anxiety: One of the primary concerns for potential EV buyers is the range of the vehicle on a single charge. They want to ensure that the EV's range meets their daily commuting needs and provides sufficient mileage for longer trips without causing range anxiety.

Charging Infrastructure: Access to a reliable and convenient charging infrastructure is crucial for EV owners. Buyers often consider the availability of charging stations, both public and private, along their regular routes and at their residence or workplace.

Cost of Ownership: The total cost of ownership, including the initial purchase price, ongoing maintenance, and operating costs, is a significant consideration for EV buyers. Many individuals compare the cost of EVs to traditional internal combustion engine vehicles to assess long-term savings, including fuel and maintenance expenses.

Government Incentives and Rebates: Government incentives, tax credits, and rebates for purchasing EVs can significantly influence buyers' decisions. Potential buyers often consider these incentives to offset the upfront cost of purchasing an EV and make it more financially appealing.

Environmental Impact: Concerns about environmental sustainability and reducing carbon emissions are driving factors for many EV buyers. They may prioritize purchasing an EV to minimize their carbon footprint and contribute to combating climate change.

Vehicle Performance and Features: EV buyers also consider factors such as vehicle performance, driving experience, interior comfort, technology features, and overall design aesthetics when evaluating different models.

Battery Technology and Lifespan: The quality and durability of the vehicle's battery pack are essential considerations for EV buyers. They want assurance that the battery technology is reliable, long-lasting, and capable of maintaining performance over the vehicle's lifespan.

Resale Value: Potential EV buyers may also consider the vehicle's resale value and depreciation rate over time. Factors such as brand reputation, battery warranty, and market demand for used EVs can impact resale value considerations.

Brand Reputation and Trust: The reputation and reliability of the EV manufacturer are crucial factors for buyers. They often research reviews, ratings, and customer feedback to assess the brand's reputation for quality, customer service, and innovation in EV technology.

Perceived Social Status and Image: For some buyers, owning an EV may be associated with status, environmental consciousness, and technological advancement. The perceived social status and image associated with driving an EV may influence their purchasing decision.

Some of the main factors that are mainly considered for are below:

- Brand
- Features & Technology
- Charging Stations
- Superior Design
- Environmentally Conscious

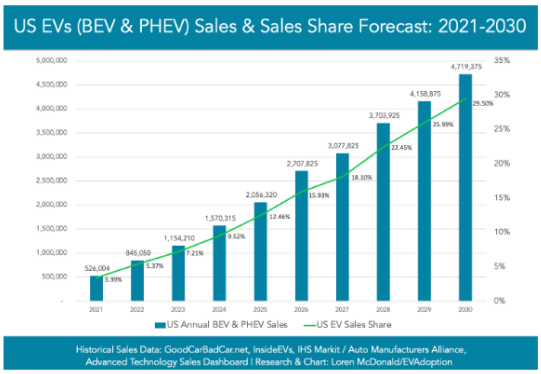
- Performance & Range

Of all the various brands in the current market, Tesla is clearly the market leader not just because of the brand value and design but mostly because of the reliable Tesla supercharging stations. The public charging stations had their share of problems as they sometimes don't work which is quite frustrating to the EV car owners. There are recent developments and collaboration between EV automakers to utilize a Tesla supercharging station which stands out to be the **best and reliable,** so far with its large network.

One of the key features for prospective EV car buyers is **charging stations**. As a part of this project, I will be exploring more on this aspect where it might be useful for upcoming EV buyers to make purchase based on the charging stations and their possible expansion pattern/prediction. The focus of this model is limited to United States of America.

Below is our latest long-term forecast for new electric vehicle (BEV and PHEV) sales in the US through 2030. EV sales should grow to reach approximately 29.5% of all new car sales in 2030 from an expect roughly 3.4% in 2021.

This would also see sales increase to 4.7 million from a little more than 500,000 in 2021.



Background/History

For any new or prospective EV car buyer, there are still few concerns before taking their decision to go ahead and buy one. While the factors like environmental consciousness, brand, features, technology, Government Incentives and Rebates, Resale Value**,** design encourage them in the first place, the **network of charging locations** is a major factor. Based on car sales data, Tesla is clearly the market leader in EV market and perhaps the popular one compared to all EV's(RIVIAN, LUCID, FISKER etc) available in the USA market. They have sold more cars than any other EV car maker. Based on surveys, EV car owners and prospective buyers consider not just the brand, features, design or tech but most importantly the Tesla's supercharging stations network. Tesla's supercharging stations are the most reliable charging stations while the public charging stations often cause problems for the owners and not reliable most of the times. Therefore, it would be nice to build a prediction model for Tesla's supercharging locations expansion for prospective EV car buyers to confidently go ahead and make their decision.

Data Explanation

The datasets that I used for this project are extracted from Kaggle website.

1. https://www.kaggle.com/datasets/omarsobhy14/supercharge-locations
2. https://www.kaggle.com/datasets/richardg9/tesla-car-sales-quaterly

The Tesla supercharge locations dataset is a treasure trove of information. It contains geographical coordinates, amenities, and other details for each supercharge location to analyze the data, discover optimal routes and uncover patterns for electrifying adventures.

This dataset has around 6000 records covering worldwide and about 2200+ for USA. Below are the various attributes of the dataset.

1. Supercharger: This feature represents the name or identifier of the Tesla Supercharger location. It helps identify and distinguish each Supercharger station in the dataset.
2. Street Address: This feature contains the specific street address where the Supercharger station is located. It provides the physical location information for each station.
3. **City: **This feature represents the city where the Supercharger station is situated. It helps identify the geographical location of each station.
4. State: This feature indicates the state or province where the Supercharger station is located. It provides additional regional information about each station's location.
5. **Zip: **This feature represents the postal code or ZIP code associated with the Supercharger station's address. It helps identify the precise location within a city or region.
6. Country: This feature indicates the country where the Supercharger station is situated. It provides information about the specific country in which each station is located.

7. Stalls: This feature represents the number of charging stalls available at the Supercharger station. It indicates the capacity of the station to accommodate multiple vehicles simultaneously.
8. kW: This feature represents the power capacity or kilowatt rating of the Supercharger station. It indicates the charging speed or power output available at each station.
9. GPS: This feature provides the GPS coordinates (latitude and longitude) of the Supercharger station. It offers precise location information for mapping and navigation purposes.
10. Elev(m): This feature represents the elevation or altitude of the Supercharger station above sea level. It provides information about the station's height relative to the surrounding area.
11. Open Date: This feature indicates the date when the Supercharger station was opened or made available for public use. It provides information about the timeline of station deployment and expansion.

We will be also using the Tesla quarterly car sales from 2013-2022 and few other datasets for supplemental datasets.

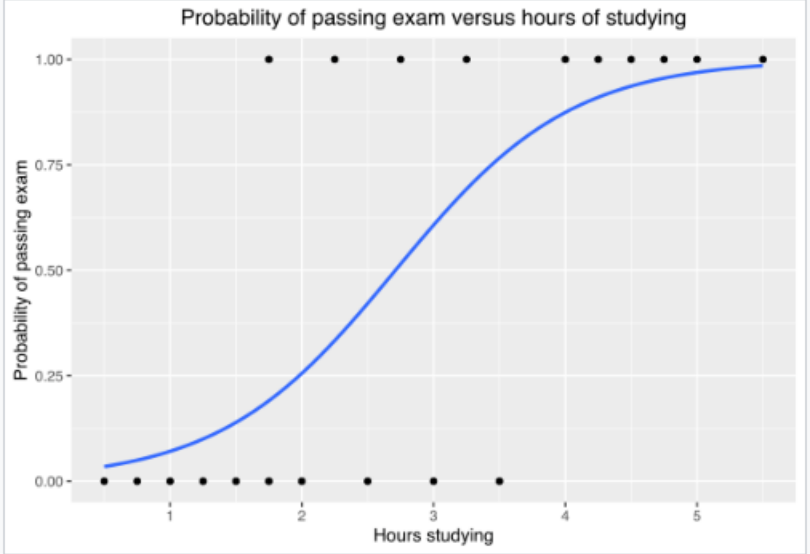
Methods

The following algorithms or model techniques will be utilized on the dataset to determine which features are related to our target variable “Purchase” (likely to purchase).

1. Logistic Regression
2. Random Forest
3. Decision Tree

Logistic regression is a statistical analysis method used to predict a binary outcome such as yes or no based on prior observation of the data set. Here, “Purchase” feature present in the dataset has only binary values and will be used as target for the model. This model falls under supervised learning as the data is well labelled and has a target variable, a column in the data representing values to predict from other columns in the data. Under supervised learning, this dataset falls under classification model as it reads the input and generates an output that classifies the input into two categories: one having purchase as “Yes” and “No”. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes.

Logistic regression is a statistical method used for modeling the probability of a binary outcome based on one or more predictor variables. Unlike linear regression, which predicts continuous outcomes, logistic regression is specifically designed for classification tasks where the outcome variable is categorical and has two possible outcomes, typically labeled as 0 and 1.



Example graph of a logistic regression curve fitted to data. The curve shows the estimated probability of passing an exam (binary dependent variable) versus hours studying (scalar independent variable). See § Example for worked details.

Here's how logistic regression works:

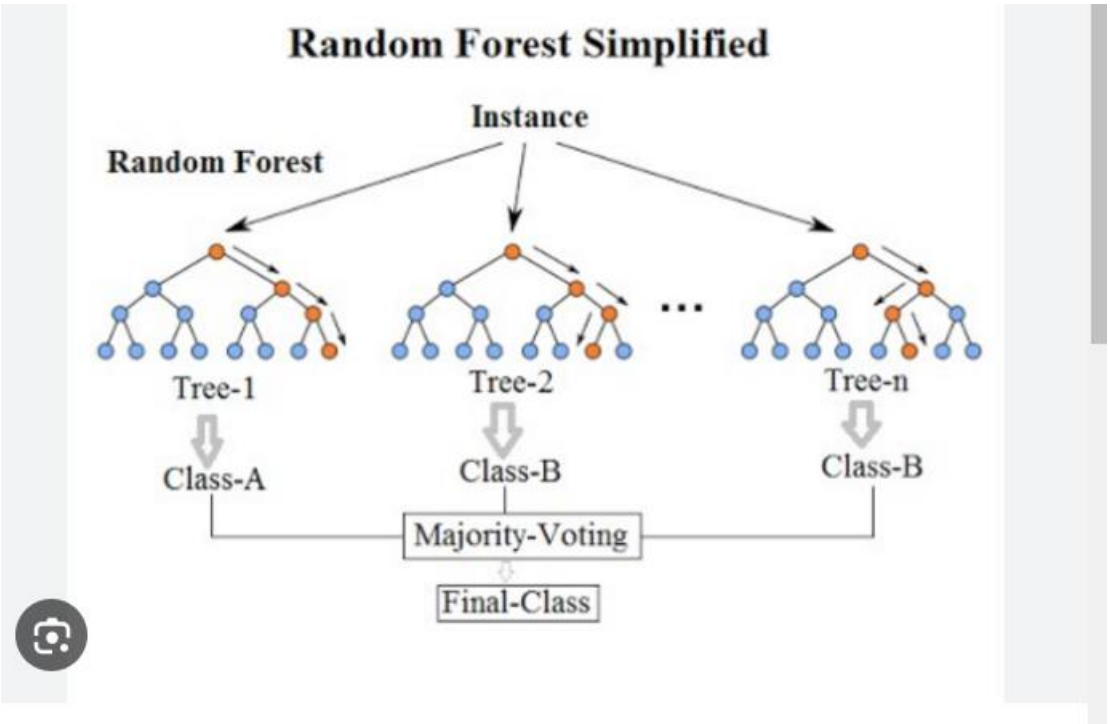
Binary Outcome: Logistic regression is used when the dependent variable (outcome) is binary, meaning it has only two possible values, usually coded as 0 and 1. For example, it could represent whether a patient has a disease (1) or does not have a disease (0).

Logit Function: Logistic regression models the relationship between the predictor variables and the probability of the outcome using the logistic function (also called the sigmoid function or the logit function). The logistic function transforms the linear combination of predictor variables into a value between 0 and 1, representing the probability of the outcome.

Model Interpretation: Once the coefficients are estimated, they can be interpreted to understand the relationship between the predictor variables and the log-odds of the outcome. For example, a positive coefficient indicates that an increase in the predictor variable is associated with an increase in the log-odds (and therefore the probability) of the outcome.

Prediction: After the model is trained, it can be used to predict the probability of the outcome for new observations based on their values of the predictor variables. Typically, a threshold probability (e.g., 0.5) is chosen, and if the predicted probability is above this threshold, the outcome is predicted as 1; otherwise, it is predicted as 0.

Random Forest is a powerful machine learning algorithm that is used for both classification and regression tasks. It belongs to the ensemble learning methods, which involve combining the predictions of multiple individual models to improve overall performance. Random Forest is particularly popular due to its robustness, flexibility, and ability to handle large datasets with high-dimensional feature spaces. Here's how Random Forest works:



Decision Trees: At the core of Random Forest are decision trees. Decision trees are tree-like structures where each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (in classification) or a continuous value (in regression). Decision trees are intuitive and easy to understand, but they tend to overfit the training data.

Random Forest Ensemble: A Random Forest consists of a collection (ensemble) of decision trees. However, unlike traditional decision trees, Random Forest introduces randomness both in the construction of the trees and in the selection of the features used to split nodes. This randomness helps to reduce overfitting and increase the diversity among the individual trees.

Bootstrap Aggregating (Bagging): Random Forest employs a technique called bagging, which stands for Bootstrap Aggregating. Bagging involves training each decision tree on a random subset of the training data, selected with replacement (bootstrap samples). This creates multiple variations of the training dataset, which are used to train different decision trees in the forest.

Random Feature Selection: In addition to using bootstrapped samples, Random Forest also randomly selects a subset of features at each node of the decision tree. This ensures that each tree in the forest is trained on a different set of features, reducing correlation between trees and improving generalization performance.

Voting or Averaging: During prediction, each tree in the Random Forest independently makes a prediction. For classification tasks, the mode (most frequent class) of the predictions from individual trees is taken as the final prediction. For regression tasks, the average of the predictions from individual trees is taken as the final prediction.

Tuning Parameters: Random Forest has several hyperparameters that can be tuned to optimize performance, such as the number of trees in the forest, the maximum depth of the trees, and the number of features to consider at each split. Cross-validation techniques can be used to find the optimal combination of hyperparameters.

Analysis

This should include the modeling analysis with accuracy and score (f1 score) calculated for all the models.

Accuracy: Accuracy represents the number of correctly classified data instance over the total number of data instances.

F1 Score: F1-Score is a metric which considers both precision and recall.

Precision: Positive predictive value

Recall: True positive rate

This should also include feature analysis using various methods to find the best features from the dataset. Best feature in the dataset which shows higher impact to the target variable “Purchase(likely to purchase)” compared to others present in the dataset.

Conclusion

Out of three possible models, I shall gauge which is the best model based on the scores to predict the purchase or the likelihood to purchase tesla car. I shall try to find the best features in the dataset to be able to come up with the prediction model so new and prospective EV car buyers can make that decision confidently.

Assumptions

The datasets being considered may not have all the required features to support the model. I have taken the best possible dataset from the available sources. Also, the data for this model is being limited to United States. The data related to other countries is being excluded for this model. Also some other features which may not be relevant will also be excluded.

Limitations

The dataset considered for this prediction model is to be considered a fictional dataset as it may not represent real-world or factual data. The same goes for the supplemental datasets as well.

Challenges

Key challenge is to ensure if this data is good enough to build the prediction model and if we need more supplemental data to support the model. Might possibly need to explore more supplemental datasets to strengthen the model.

As we are in a real-time change world, this will not represent the overall accurate prediction, due to location constraints and competition on the EV market and government subsidies.

Future Uses/Additional Applications

While this may not exactly represent the real-world data, this model is still similar and can be run against real-world datasets to all other brands or EV manufacturers to get some useful insights.

Recommendations

This model predicts the purchase and relevant useful features that impact the likelihood of purchase with better accuracy with a caveat that the model should be regressed when more or better real-world data is available.

Implementation Plan

As stated in the recommendations, this model can be implemented to predict the likelihood of an EV car purchase for a particular brand along with evaluation of other useful features that may impact the same outcome. While evaluating other features from the datasets, model must be ensured to re-evaluate for no slippage.

Ethical Assessment

There are no possible ethical aspects to this model as the data is public info and doesn't really include any consumer related information.

References

1. Dataset1: https://www.kaggle.com/datasets/omarsobhy14/supercharge-locations
2. Dataset2:https://www.kaggle.com/datasets/richardg9/tesla-car-sales-quaterly
3. https://amplifyxl.com/target-market-for-tesla/
4. Random Forest: https://en.wikipedia.org/wiki/Random_forest
5. Logistic Regression: https://en.wikipedia.org/wiki/Logistic_regression