Predicting the Utility of an Electric Vehicle for a Specific Mobility Behavior

Seminar paper

handed in to
Prof. Dr. Thorsten Staake
Energy Efficient Systems Group, Faculty WIAI
University of Bamberg

07/28/2016

by
Mathias Renner

Information Systems, M.Sc.

Contents

List of Abbreviations

CRISP-DM	Cross Industry Standard Process Data Mining Model
UNECE	United Nations Economic Commission for Europe
EV	Electric Vehicle
MAPE	Mean absolute percentage error

List of Tables

List of Figures

1 Introduction

This chapter begins by establishing the problem statement and a summary of the approach used herein. Next, this work's approach is compared to that which is practiced today. Finally, I provide an example of how this approach can be applied in practice.

1.1 Abstract

A major barrier for widespread use of electric cars is the high level of uncertainty that potential buyers face when it comes to estimating the car's utility for themselves. Today, the maximum driving range of an electric car is used to evaluate its utility, despite multiple studies demonstrating the limited ability of this metric to evaluate electric car utility. In addition, utility is very much influenced by an individual's driving behavior, which is not considered in the maximum driving range as utility index.

Herein I identify the most significant factors for measuring utility of an electric car, and based on the results I provide a new estimate an electric car's utility. The estimation is provided for specific individual's driving behavior due to the strong influence of such behavior on car utility.

The results indicate that for each driver, the average speed or average distance of single trips are sufficient to estimate a car's utility within 9,5% of the true utility index I established. Manufacturers can use this approach to estimate the utility of an electric car with limited data (i.e. average speed and average distance per trip) from the customer. This data is easily obtained from cars today, making this approach easy to implement.

1.2 Potential Buyers of E-cars Face High Level of Uncertainty

When General Motors started selling the "EV1" car model, customers were primarily concerned about the driving range of the cars. "Some EV1 drivers gave the term 'range anxiety' to their continual concern and fear of becoming stranded with a discharged battery in a limited range vehicle" (Tate, Harpster & Savagian, 2008). Even if this "range anxiety" is irrational and decreases with more experience in using an electric car, this is a major barrier for potential buyers (T. Franke, J. F. Krems, 2013, and Long & Egbue, 2012).

Concerns about the driving range of an electric vehicle (EV) can be reduced to a question of its utility, which for most consumers, is evaluated on the EV's range specified by the manufacturer. However, this range is the result of a standardized test with limited practical significance. As discussed later, the driver's behavior holds a strong influence on the range such that the range indicated by the manufacturer can deviate a lot from the true value.

This work proposes an approach to solve this problem: using machine learning methods, the utility of an EV shall be predicted for an individual driver based on the individual's driving behavior. As a result, a potential buyer shall receive viable information about an EV's utility, tailored to him or her.

The result is quantified by the proportion of goals the driver can reach with a specific EV. For instance, if the result in a given case yields 95%, a specific buyer can be confident that he/she can reach 95% of all targeted goals. In order to estimate the proportion to which goals are achieved, I identify the factors that have the highest impact on achieving the targeted goals using predictive analytics techniques.

The overall approach of this work has been developed by Jürgen Wenig, a researcher at the Energy Efficient Systems Group at University of Bamberg. This work is an implementation and execution of his approach.

1.3 Comparison of this Work's Approach Against the Approach of the European Union

I compare this works's approach with that which is used for all EVs in Europe. This approach is ratified by the United Nations Economic Commission for Europe (UNECE) as the only method with results (= maximum driving range) allowed in promotional sales material[1]. This driving range is semantically seen as equivalent to the utility of EVs.

Our approach to indicate the utility is different in various aspects as depicted in the following table:

Table 1.1: Comparison of approaches that evaluate the utility of EVs

Criterion	**UNECE Approach**	**Approach of this work**
Method	Standardized test based on average driving behavior	Estimation of an EV's utility based on an individual's driving behavior
Result	Max. driving range	Proportion of goals achieved
Significance	Limited due to generalized data	Signifcant for each driver's behavior
Practicability	Limited due to the low significance of driving range, and failing to include the strong influence of driving behavior	Highly practical due to consideration of the driving behavior and other significant influences
Data requirements	Needs *average* driving data as average over many indivduals	Needs *specific* driving data of an indivdual in a specific car
Test frequency	Once for an EV	Once for each driver *and* EV

[1] The regulation can be found at `http://www.unece.org/fileadmin/DAM/trans/main/wp29/wp29regs/updates/R101r3e.pdf`

With the UNECE method as presented in table 1.1, an EV's utility is estimated in a test environment. There, the EV accelerates and breaks according to a predefined sequence that shall represent the average driving behavior of an individual. For more details about this test, please see the drive cycle in Figure A.1 of the appendix.

The method presented in this work estimates the utility based on each individual's driving behavior. Using existing driving data and their according utility, new drivers are matched with a similar behavior to estimate their personal utility.

Regarding the result, the outcome of the UNECE approach is an EV's maximum driving range in kilometers. The outcome of this work is a percentage of goals the driver can achieve with his/her EV, on average. This result can be very significant for a driver since the result is applied on a driver's individual driving behavior. In contrast, UNECE's approach is based on the assumption that individuals have very similar driving behaviors — an assumption which does not hold (discussed later). Therefore, the result of UNECE's approach is rendered impracticable for most drivers.

Furthermore, his work's approach relies heavily on specific data, as depicted in table 1.1. The test performed by UNECE is based on average driving data of many individuals in order to test a car on average driving behavior. In contrast, our approach requires driving data of specific individuals on specific cars. More precisely, the following two data sets are required:

1. A data set with driving information from many drivers of a specific EV. This data set is used to identify influence of driving behaviors to the utility, i.e. the proportion of goals that are achieved. During the analyses in chapter 2, this data set is referred to as 'training set'.
2. A data set from a potential consumer/buyer/driver seeking to obtain a utility score for themselves. For the analyses, a subset of the As I will show in the subsequent chapter, the average speed or the average trip distance is sufficient for providing a good estimation. During the analyses in chapter 2, this data set is referred to as 'test set'.

Finally, regarding the test frequency, the UNECE test has to be performed only once for a new EV, thereby providing an estimation that neglects an indivdual driver's influence on the result. This work's method has to be executed for any new EV and driver in order to provide an estimation depending on a specific driver's behavior.

All in all, the UNECE comes at lower costs, but also but does not provide an accurate score due to the generalizations it makes.

1.4 Case Example

In order to support the understanding of this approach, I show a possible implementation by TESLA, Inc., a manufacturer for EVs in the USA. In case a potential buyer of an EV driver is interested in an EV, he might see this message on TESLA's website:

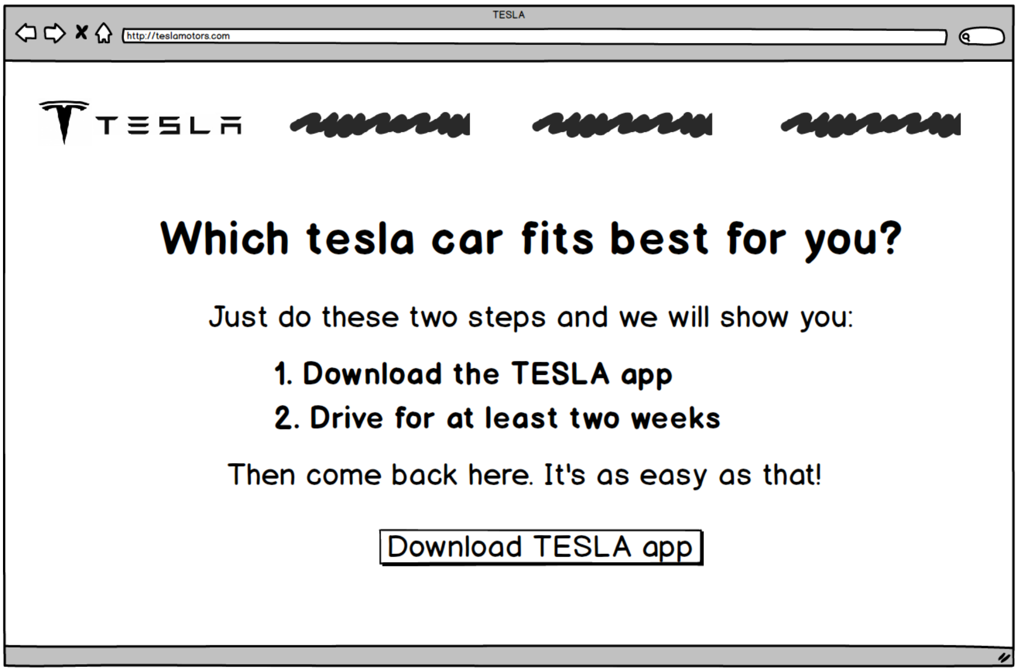


Figure 1.1: Mockup of TESLA's website for potential buyers, requesting data

The app, which the user can download, fulfils the second part of the data requirement stated previously, i.e. via the GPS module of the smartphone, the app gathers data about the driving behavior of the specific user. I assume that the first data requirement has already been fulfilled. After the two week period, if the customer has driven sufficient distances, there is enough driving data available to match the driver's behavior against the available EV models of TESLA. The website then recommends the model with the highest utility, i.e. the highest proportion of goals that the user can reach with this EV:

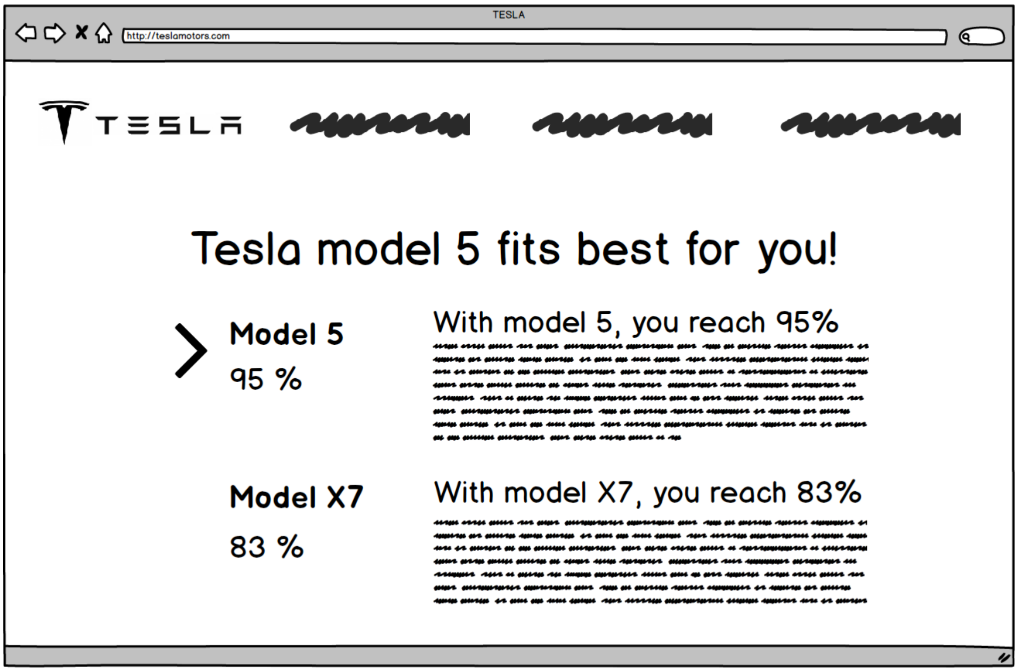


Figure 1.2: Mockup of TESLA's website for potential buyers recommending the EV with best fit

2 Method and Execution of Data Analysis

In this chapter, I describe the data mining tasks and analyses that have been completed to estimate the utility. All analytical tasks have been programmed with the software R[2].

2.1 CRISP-Model as General Approach

When executing the data mining tasks, I follow the reference model Cross Industry Standard Process Data Mining Model (CRISP-DM) (Shearer, 2000). The CRISP-DM model depicts six phases that are successively followed during the data mining process: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. The complete model is illustrated in the appendix, Figure A.2. The following sections follow the model's proceeding. Note that the final phase (Deployment) is skipped because its objective — the presentation of this work's outcome — has already been completed.

2.2 Business Understanding

In this phase, the CRISP model proposes to state the objectives of the project:

I will predict the utility of a specific driver for a specific car, thereby measuring the utility as "percent of achieved goals".

Because the data provided for this work only features a BMW i3 car model, all analyses hold exclusively for this car.

As methods I use use regression models with "percent of achieved goals" as dependent variable and the most signifcant features that can be obtained from the data set as independent variables. The result is a regression model that shall provide a low prediction error compared to the true value of the dependent variable.

2.3 Data Understanding

In this phase, I collect and explore the data and its quality.The Energy Efficient Systems Group at University of Bamberg provided one table containing driving data of 50 different BMW i3 EVs. The table contains 500.000 data points with about 10 features

[2] https://www.r-project.org/

(=attributes) for each point, including important information on GPS location, distance travelled since last GPS location, state of battery charge, type of area (urban area, extra-urban area, highway) and others. A description of all attributes of the data set is depicted in appendix, table A.1. The GPS location of the data points are distributed mainly in northern Italy. I did not find major data quality issues, the minor ones will be handled in the next phase.

2.4 Data Preparation

In this phase, I perform all tasks regarding cleaning and preparing the data, such that I can continue to create the regression models in the next chapter. First, I calculated the speed of each data point and employed an outlier detection on speed values, indicating that several data points exceed the maximum possible speed of the BMW i3. These values were winsorized to the maximum speed.

Next, I calculated the target variable and had to choose between two scenarios. First, in case the EV can be only charged at the driver's most frequented parking spot (e.g., probably his home). Second, in case the EV can also be charged at a secondary location that is his/her second most frequented location (e.g., a regularly visited parking lot of a local super market).

Performing the analysis based on the second scenario is more realistic, because EV charging stations are increasingly found at public places. Therefore this work's analyses concentrates on the second scenario in order to yield results that are representative today and more so in the future.

In the given data set, the difference between the two scenarios is minimal: In the scenario with two charging opportunities, the mean of the target variable is only 3,16% higher. The target variable neither follows a clear skew and nor a standard distribution. In addition to the visual analysis by Figure 2.1, the Shapiro-Wilk normality test has been used that supports this assumption (p-value of 0.022). The following Figure 2.1 illustrates the distribution:

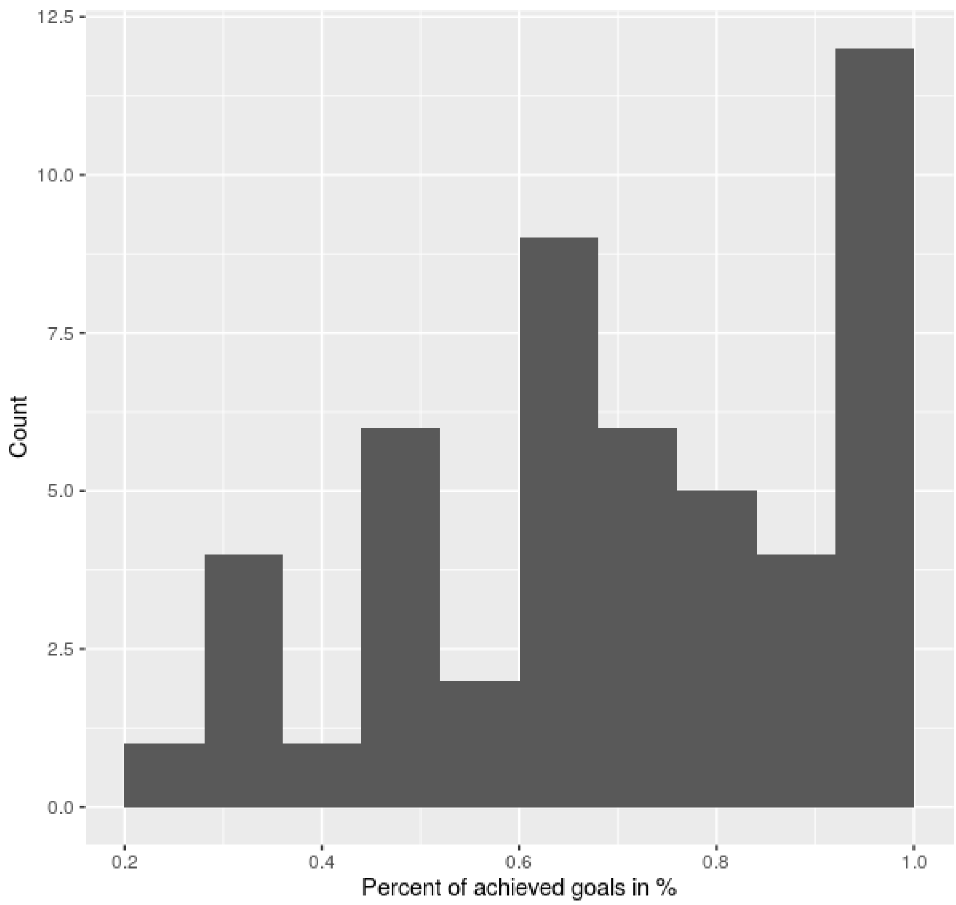


Figure 2.1: Distribution of the target variable ”percent of achieved goals”

Based on this data, I performed feature extraction by identifying and calculating characteristics within the data that might be interesting for post-hoc analysis. In total, 25 features were extracted, e.g. ”speed in average”, ”average distance on highways on weekends”, ”parking time per trip in average” etc.

In addition, many features were calculated for both one-way and round trips. Single trips contain the data points from an EV starting to move to stopping. Round trips contain data points that both start and end at the primary charging station.

2.5 Modeling

In this phase, I create regression models to identify characteristics that most influence the target variable of achieved goals. First, the data was divided into a training set, i.e. 40 EVs, and a test set with the 10 EVs left. The models were built and improved with the training set, while the test set was used to verify the models.

Next, I encountered target variable values greater than one, which should not be possible because the values of the deptendent variable should range from 0% to 100%. Thus, instead of a linear regression, a logistic regression was used that does not allow values below zero or above one. An overview of the mathematical formulas and their impact on the models are depicted in appendix, Figure A.3.

Finding a good model is an elaborate process. To accelerate this greedy search approach, I use the step function in R. This function builds models based on all possible combinations of every feature and tests their quality using the Aikaike Information Criterion (AIC), a quality index. The step function provides a list of the most significant features as an output.

Even though the step function is limited to application on linear regression models, the results are still useful, because the significance of a feature in a linear model is very similar to the one in logistic models. One of the first models I built was a univariate model with only round trip distance as an independent variable.

Then, in an semi-automated way using loops in R, between 200 and 300 models were subsequently built in order to identify the best one. Only the most useful models remained in the source code provided with this paper. The criteria on which the best models have been selected and the results of the predictions will be presented in the next chapter.

Since the next phase also summarizes this work's results, there is a dedicated chapter to present the results.

3 Results and Evaluation

This chapter successively evaluates and summarizes the results of the data analysis performed in chapter 2.

3.1 Selection and Evaluation of the Best Models

In order to select the model that provides the best estimation for a given EVs utility, I state criteria that were found to be important for good models to hold:

1. **A high pseudo-R^2**, which is a well suited quality measure to compare logistic regression models (Hu, Shao & Palta, 2006).
2. **A small error estimation compared to the true value of the target variable**. Thereby, I use the "mean absolute percentage error" (MAPE) adopted from Armstrong (1985). MAPE is calculated by the average difference between the true value of the target variable and the predicted error. All MAPE results were validated with a 5-fold cross validation using 10 cars as a test set for each validation loop. After each loop the results were visualized. One example visualization is presented in appendix, Figure A.3.
3. **Low correlations between the independent variables.** A major threat for models are correlations between independent variables. High correlations designate strong interactions between variables. In that case, each variable does not provide unique information to the model based on the data, instead new information is generated based on the interaction (Erlbaum, L., 2001). This can lead to a very low significance of the model (Graham, 2003). To cope with correlations, I tested all models and improved or rejected the ones with correlations greater than 35%. I created correlation matrices and visually indicated high correlations such as in heat maps, in order to easily identify high correlations.
4. **Small quantities of features**. A smaller number of independent variables in the model is more practicable because less input data is needed, thereby reducing data acquisition and processing times.
5. **The independent variables can be interpreted intuitively.** A feature's influence on the model (i.e. a feature's coefficient) should be easy to interpret for taking the appropriate actions. In the appendix, Figure A.4, I provide an example of a model that has been rejected due to this criterion. Despite it's good estimation results, that one shows counter-intuitive results.
6. **The indeptendent variables are all significant.** All features used in a model should be significant with a p-value smaller than 0.1.

After evaluating all models against these criteria, I found three models with good characteristics, to be presented subsequently for further discussion:

Table 3.1: Comparison of the three best models as a result of this work

Model ID	Independent variables of the model	Pseudo-R^2	Estimation error (MAPE) in %
1	-Average distance on a single trip in km	0.440	9,46
7	-Average speed in km/h, -Overall parking time in hours	0.696	9,62
12	-Average speed in km/h	0.646	9,64

Note that in table 3.1, the 'Model ID' references the model number in the R source code that is delivered with this paper.

The first model (ID = 1) is univariate, which therefore can be plotted nicely:

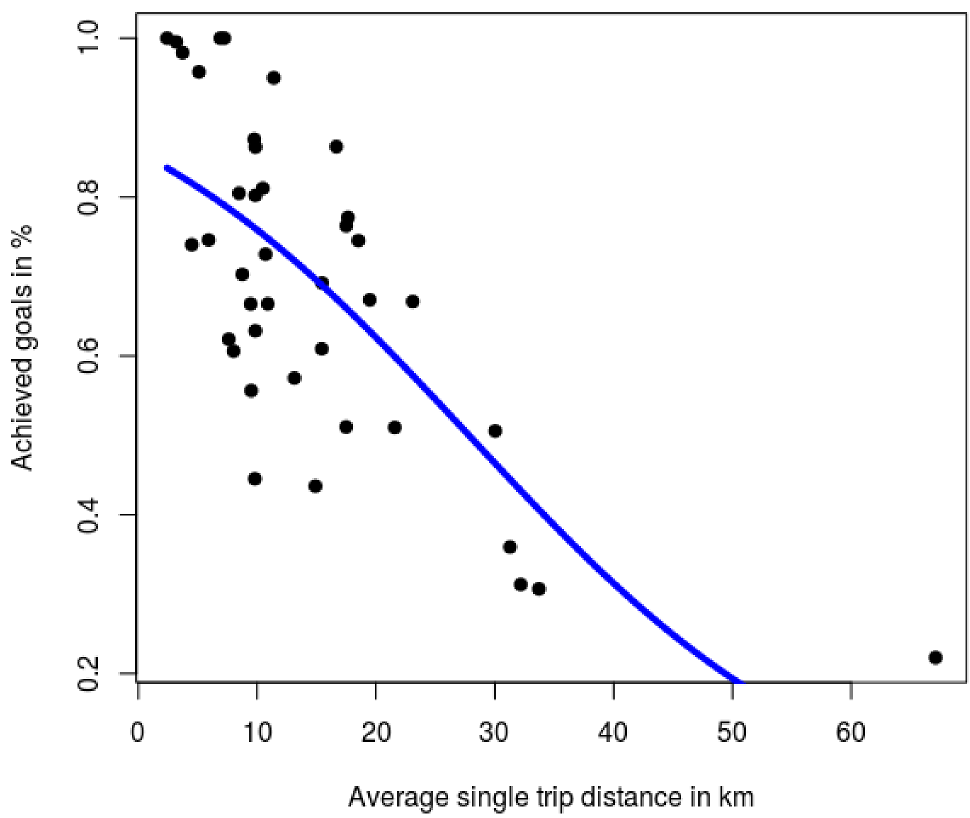


Figure 3.1: Univariate regression model ID 1: Avergae single trip distance vs. achieved goals

As depicted in Figure 3.1, the first model shows that when individuals cover longer distances on average, the percent of achieved goals is reduced. This is an intuitive result for this model, but only has a Pseudo-R^2 of 0.44, which is lower than other models discussed here. Nevertheless, its estimation error is good, with an average deviation from the true value of only 9,46%. This estimation error is similar to the level of the other two models.

The second model with ID 7 as listed in table 3.1 uses the average speed (km/h) and total parking time as independent variables. Compared to the first model, the Pseudo-R^2 (69,6%) is much higher and estimation error is slightly better (9,62%). The independent variables influence the target variable differently. While the percent of achieved goals is inversely related to average speed, the percent of achieved goals is directly related total parking time (hours). The slope coefficient of the speed variable is -0.084, which appears reasonable. While the parking time is significant ($p = 0.0161$), its coefficient of 0.0006 indicates it bears little influence on the target variable. Thus, average speed of a driver has greater influence on the target variable.

The third model with ID 12 as depicted int able 3.1, exclusively uses the average speed, which surprisingly was so significant that the Pseudo-R^2 only slightly decreased (from 69,6% to 64,6%) and the estimation error only slightly increased to 9,64%, from 9,62%. Thus, the model with only average speed as an independent variable is almost as good as the previous one.

Comparing the first and third model (ID 1 vs. ID 12) yielded the interesting result that both models have extremely similar (low) estimation error, but the first model's Pseudo-R^2 was much lower. It is assumed that for a solid comparison, more data is needed to evaluate whether or not the lower Pseudo-R^2 of the first model leads to worse results.

3.2 Results and Discussion

The third model demonstrates how much speed influences the number of goals a driver can achieve. This contradicts the foundational assumption of UNECE test that the distance covered is the most important factor for estimating utility of an EV. As discussed with model ID 1, the average distance is important, but this model has a much lower R^2 while the estimation error stays almost equivalent.

This results' validity and widespread generalizations are limited due to the small sample size of 50 EVs. The final models should be proven with a larger data set consisting of a larger populations and which is better distributed. Driving behavior in other countries with different terrain may deviate.

In addition, the given data set is derived from simulations of EVs on cars powered by combustion engines, whereas a data set with ”real” EVs might provide more realistic data.

While the results of our approach are significant for the given data set, the data requirements for this approach are greater compared to the approach of UNECE, as discussed in section 1.3. Thus, better estimating the utility of an EV with the approach presented in this work comes with cost of higher data needs. Therefore I favored models with less independent variables because of the trade-off between data requirements and model quality.

3.3 Conclusion and Outlook

The goal of this work was to estimate the EV utility for an individual based on his/her driving behavior using predictive analytics. This framework allowes to estimate the utility for a specific EV and driver with an average error of about 9,5%, verified by cross validations.

I address the problem of additional data needs (compared to the UNECE approach) by favoring models with fewer independent variables. The model with ID 12 reveals good results by using only average speed of a driver as input. This is very practicable because the average speed is easy for drivers to retrieve with minimal effort, as most cars show average speed on an internal display. If one wants to use models with higher data requirements, I recommend to retrieve the required driving data this with a GPS-tracking app on a smartphone, as presented in section 1.4.

The approach of this work can be used by companies such as TESLA and any EV manufacturer, as illustrated in section 1.4. It is also useful for comparing portals featuring EVs. Furthermore, the driving data can provide insights of savings from EVs compared to vehicles with combustion engine. This reduces the level of uncertainty for potential buyers additionally.

The widespread validity of results presented here are limited due to small sample size and because the data was based on a simulated, rather than 'real' EVs. Further analysis with a larger population and non-simulated data should be used in future explorations.

4 Bibliography

Armstrong, J. S. (1985): Long-range Forecasting: From Crystal Ball to Computer, p. 348, Wiley, New York.

Erlbaum, L. (2001): Analysis of Variance, Journal of Consumer Psychology, 10(1&2), pp. 5-35

Franke T., Krems J. F. (2013): What drives range preferences in electric vehicle users, Transport Policy, Volume 30, pp. 56-62.

Graham, M. H. (2003): Confronting multicollinearity in ecological multiple regression. Ecology, 84(11), p. 2809-2815.

Hu, B., Shao, J. & Palta, M. (2006): Pseudo-R in logistic regression model. Statistica Sinica 16, pp. 847-860.

Long, S., Egbue, O. (2012): Barriers to widespread adoption of electric vehicles: An analysis of consumer attitudes and perceptions, Energy Policy, Volume 48, pp. 717-729

Shearer, C. (2000): The CRISP-DM model: the new blueprint for data mining. Journal of data warehousing, 5(4), pp. 13-22.

Tate, Harpster & Savagian (2008): The Electrification of the Automobile: From Conventional Hybrid, to Plug-in Hybrids to Extended-Range Electric Vehicles, p. 3

A Appendix

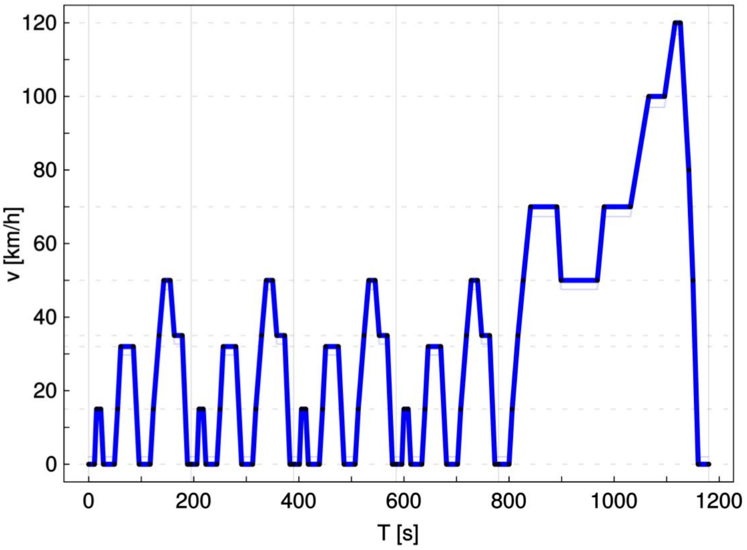


Figure A.1: The standardized driving cycle used by the UNECE

The picture source of Figure A.1 is `https://en.wikipedia.org/wiki/New_European_Driving_Cycle`

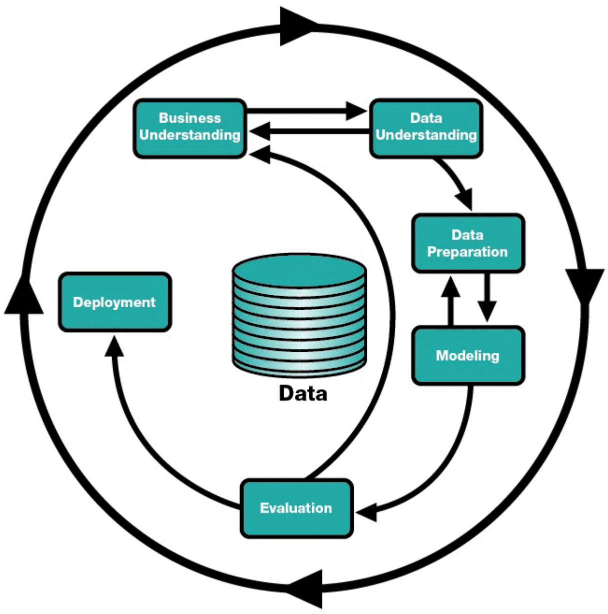
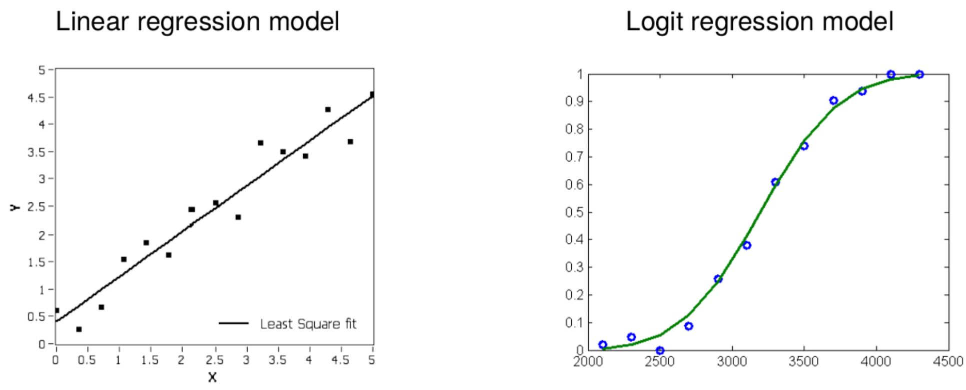


Figure A.2: The CRISP-DM model

Figure A.2 has been derived from `http://crisp-dm.eu/`.



$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Figure A.3: The formulas and their impact of the linear model and the logistic regression model

Images and formulas in Figure A.3 have been obtained from `https://en.wikipedia.org/wiki/Logistic_regression`.

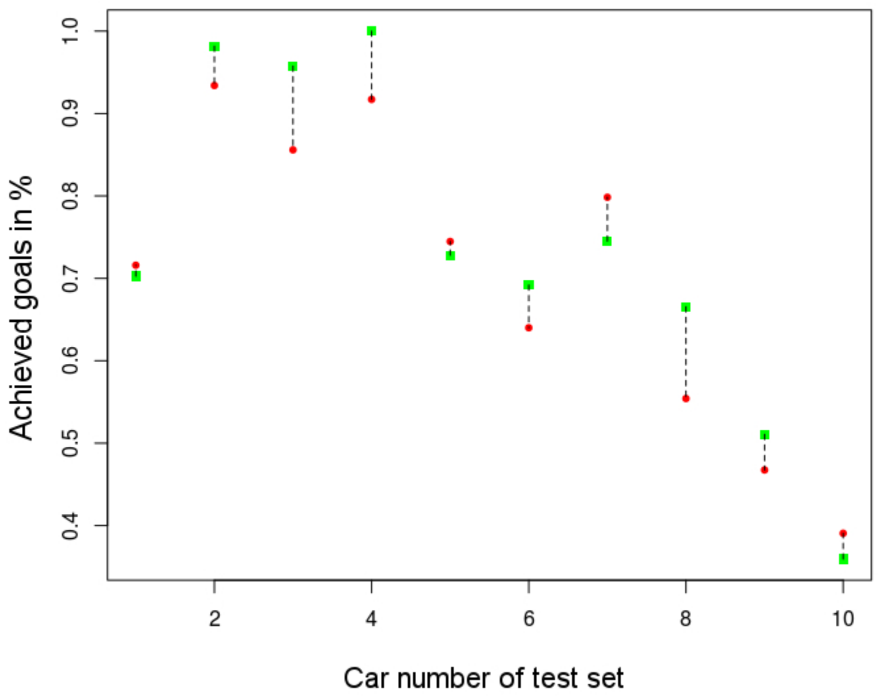


Figure A.4: MAPE deviation plot of one of the a cross validation loops

In Figure A.4, true and predicted values are represented as green points and red points, respectively, while the dashed line indicates the difference between each tupel.

```
glm(formula = REACHED_GOALS_SOCBOTH ~ SPEED_AVG_KMH + … )

                                 Estimate  Std. Error t value Pr(>|t|)
(Intercept)                       5.496e+00  6.778e-01   8.108 2.33e-09 ***
SPEED_AVG_KMH                    -9.935e-02  1.557e-02  -6.381 3.16e-07 ***
DISTANCE_HIGHWAY_KM              -6.441e-05  2.257e-05  -2.854  0.00741 **
ROUND_TRIP_DISTANCE_AVG_KM        8.750e-03  3.466e-03   2.525  0.01656 *
ROUND_TRIP_PARKING_TIME_AVG_H    -5.590e-02  1.740e-02  -3.212  0.00294 **
SINGLE_TRIP_PARKING_TIME_AVG_H    2.871e-01  9.314e-02   3.082  0.00413 **
ROUND_NUMBER_TRIPS_AFTERNOON      2.480e-04  1.070e-04   2.319  0.02676 *
```

Figure A.5: One of the models that has been rejected due to bad characteristics

Figure A.5 has a low MAPE, but high correlations between ROUND_TRIP_PARKING_TIME and SINGLE_TRIP_PARKING_TIME of 93%, and counter-intuitive results such as ROUND_TRIP_DISTANCE_AVG_H has a positive influence on the reached goals.

Table A.1: Attributes of driving data provided by Energy Efficient Systems Group

Feature	**Properties**
ID_TERMINAL	Car/telematics device ID
LONGITUDE	Longitudinal vehicle position in decimal notation (xxx.dddddd)
LATITUDE	Latitudinal vehicle position in decimal notation (xxx.dddddd)
TIMESTAMP	Date (dd/mm/yy) and time (hh:mm) on which the data point was recorded
DELTAPOS	Distance travelled since last recording point in meters
DELTATIME	Time travelled since last recording point in seconds
ID_PANELSESSION	Provides dataset description / dataset purpose: 0 = Dataset recorded on ignition turn-on; 1 = Dataset recorded during vehicle operation; 2 = Dataset recorded on ignition turn-off
ID_LOCATIONTYPE	Road type at recorded location, imputed by Octo Telematics: 0 = Urban; 1 = Highway: 2 = Extra-urban
PARKING	Parking location type: 1 = primary parking location, 2 = secondary parking location, 0 = neither primary nor secondary
SOCPRIMARY	State of charge when charging is possible at a primary parking location (battery capacity = 18.8 kWh, charging power = 3.68 kW)
SOCBOTH	State of charge when charging is possible both at a primary and at a secondary parking location (battery capacity = 18.8 kWh, charging power = 3.68 kW)