# Abstract [*≈* 1 paragraph] (Luigi)

The abstract is optional, depending on your available space. It should consist of 0.5 paragraph consisting of the motivation for your paper and a high-level explanation of the methodology you used/results obtained.

1. **Introduction [***≈* 0*.*5 **pages] (Luigi)**

Explain the problem and why it is important. Discuss your motivation for pursuing this problem. Give some background if necessary. Clearly state what the input and output is. Be very explicit: “The input to our algorithm is an *{*image, amplitude, patient age, rainfall measurements, grayscale video, etc.*}*. We then use a *{*SVM, neural network, logistic regression, etc.*}* to output a predicted *{*age, stock price, cancer type, music genre, etc.*}*.” This is very important since different teams have different inputs/outputs spanning different application domains. Being explicit about this makes it easier for readers. If you are using your project for multiple classes, add a paragraph explaining which components of the project were used for each class.

The production of cars has been steadily increasing in the past decade and this has given rise to the used car market. The emergence of online second-hand car portals in the United Kingdom and Europe has facilitated the need for both the customer and the seller to be better informed about trends and patterns that determine the value of a used car [1]. A second-hand car price prediction system (for a specific car brand) is thus required to effectively determine the worthiness of the car using a variety of features. The price prediction model and insights/patterns could be then later used as a tool for a second-hand car retailer to give insights to potential customers shopping for a second-hand car.

To tackle the need of a price prediction system, we are going to divide the price prediction problem into two subproblems: a price prediction model for a specific car brand (Mercedes C-Class W205) and a price prediction model for an entire brand (Mercedes). Our main goal is thus to make a price prediction model both a specific model and an entire brand and extract some interesting insights of the entire dataset (all cars and models combined).

For our first price prediction model of the Mercedes C-Class W205, we predict the price (output) based on the mileage (input). We first implement regularized linear regression and will further improve our model by introducing polynomial regression. For our second price prediction model of Mercedes cars, we predict the price (output) based on several features (year, fuel type, transmission type, etc.) by using linear regression with multiple variables.

1. **Related work [***≈* 0*.*5 **pages] (Luigi)**

You should find existing papers, group them into categories based on their approaches, and discuss their strengths and weaknesses, as well as how they are similar to and differ from your work. In your opinion, which approaches were clever/good? What is the state- of-the-art? Do most people perform the task by hand? You should aim to have at least 5 references in the related work. Include previous attempts by others at your problem, previous technical methods, or previous learning algorithms. Google Scholar is very useful for this: <https://scholar.google.com/>(you can click “cite” and it generates MLA, APA, BibTeX, etc.)

Many research used regression techniques, but other techniques such as neural networks (NN), Support Vector Machines (SVM) and decision trees were also used.

The work of Noor K. and Jan S. [2] used multiple linear regression to make a price prediction model for second hand cars in India. The research performed several variable selection techniques in Minitab to extract the most useful features. The result of the prediction model was accurate with an r² of 98%. Similarly, the research of Kuiper S. [3] also used multiple linear regression to make a price prediction model of GM-cars from 2005. The research of Kuiper used several variable selection techniques using Mintab such as Mallows’s CP, Akaike information criterion, stepwise regression and manually exploring multicollinearity on raw data. Kuiper found that the Mallows’s CP was the best for picking the features, but he also concluded that there does not exist a “best” regression model or variable selection technique that guarantees a “best” regression model. In addition, the research of Sameerchand P. [4] also used multiple linear regression for used cars in Mauritius and observed the correlation between features to select the ones used for the model. However, they concluded that their accuracy was relatively low because the dataset, which they collected from daily newspapers, had not enough data to make an accurate model.

Peerun et al. [5] did research on using NN in used car price prediction in Mauritius but concluded that their result was inaccurate. Their research had the same weakness as [4] because their dataset was not sufficiently large. This does not mean that NN performs bad on price prediction. Sun et al. [6] used NN for a car price prediction model and introduced a new optimization method called Like Block-Monte Carlo Method (LB-MCM) to optimize hidden neurons. The optimized NN model yielded higher accuracy compared to other work using NN.

According to the research Listiani [7] for predicting the price of leased cars, SVM proves to yield a higher accuracy than both NN and multiple regression. When a large dataset is available, she found that SVM is considerably more accurate than multiple linear regression in predicting prices. SVM is also superior at handling high-dimensional data and avoids both under-fitting and over-fitting problems.

Lastly, since we only made a model of one specific car and brand and not (yet) the entire dataset of different brands, there also exist research of one specific brand. The research of Erfan S. [8] used different techniques including decision trees, SVM, random forest and deep learning for predicting the price Tesla vehicle. He found that decision tree yielded the best result.

1. **Dataset and Features [***≈* 0*.*5 *−* 1 **pages] (Luigi)**

We used the dataset from Aditya from Kaggle [9]. This dataset contains data from 100 000 used cars from the UK which are divided into separate brands. Each brand is divided into a separate file and contains information about the model, year, price, transmission, mileage, fuel type, road tax, mpg and engine size.

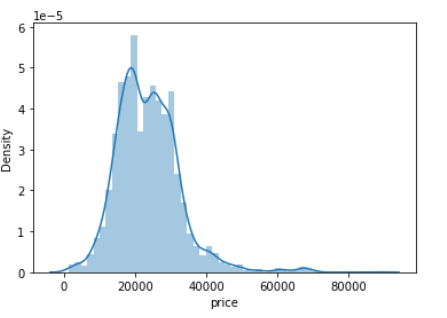
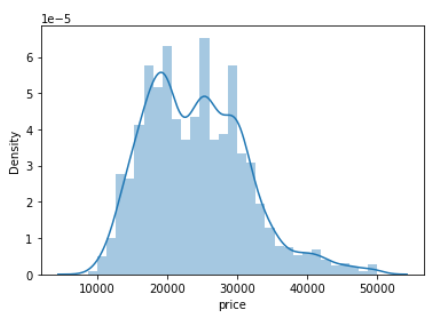
For the price prediction model of the Mercedes C Class W205 we used the separate cclass.csv file and for the price prediction model of all the Mercedes cars we used the merc.csv file. Table 1 shows the records for the C Class dataset. The records of the Mercedes dataset are similar, but with different models. An example of some records is shown in table 1 (of the C Class dataset).

Table 1: records of the C Class dataset



Since we only want to predict the price for the W205 C-Class we filtered the C-Class dataset to only include records ranging from 2014-2020. Secondly for both the C-Class and Mercedes dataset we filtered out the small amount of high prices of both datasets since they do not represent the majority of the prices but can influence our accuracy.

Figure 1: price density C Class dataset before filtering (a) and after filtering (b)



a

b

Figure 1a shows the price density plot for the C Class dataset before filtering, and figure 1b. shows the price density plot for the C Class dataset after filtering (only keeping records ranging from £0-50 000). The Mercedes dataset was filtered too (only keeping records ranging from £0-60 000). Additionally, both the C Class and Mercedes features from both datasets were also normalized.

We further also split our C Class dataset into a training/cross-validation/test-set consisting of 2137, 713, 713 records respectively.

Lastly, only for our Mercedes dataset, we transformed columns that contain more than 2 different non-numerical values (car-models, transmission, fuel type) into separate columns with binary values to help ease our further calculations as shown in table 2.

Table 2: records of the transformed Mercedes dataset



1. **Methods [***≈* 1 *−* 1*.*5 **pages] (Kai)**

Describe your learning algorithms, proposed algorithm(s), or theoretical proof(s). Make sure to include relevant mathematical notation. For example, you can briefly include the SVM optimization objective/formula or say what the softmax function is. It is okay to use formulas from the lecture notes. For each algorithm, give a short description (*≈* 1 para- graph) of how it works. Again, we are looking for your understanding of how these machine learning algorithms work. Although the teaching staff probably know the algorithms, future readers may not (reports will be posted on the class website). Additionally, if you are using a niche or cutting-edge algorithm (e.g. long short-term memory, SURF features, or anything else not covered in the class), you may want to explain your algorithm using 1/2 paragraphs.

# Results/Discussion [*≈* 1 *−* 3 pages] (Kai)

You should also give details about what (hyper)parameters you chose (e.g. why did you use X learning rate for gradient descent, what was your mini-batch size and why) and how you chose them. Did you do cross-validation, if so, how many folds? Before you list your results, make sure to list and explain what your primary metrics are: accuracy, precision, etc. Provide equations for the metrics if necessary. For results, you want to have a mixture of tables and plots. You should include a confusion matrix or AUC/AUPRC curves. Include performance metrics such as precision, recall, and accuracy. Include visualizations of results, heatmaps, examples of where your algorithm failed and a discussion of why certain algorithms failed or succeeded. In addition, explain whether you think you have overfit to your training set and what, if anything, you did to mitigate that. Make sure to discuss the figures/tables in your main text throughout this section. Your plots should include legends, axis labels, and have font sizes that are legible when printed.

# Conclusion/Future Work [*≈* 1 *−* 2 paragraphs] (Luigi)

In this work, a price prediction model was made the W205 C Class and Mercedes cars in general using linear/polynomial regression and linear regression with multiple variables respectively. It makes sense that the linear regression model performed the worst with an r² score of … because there almost is no linear relation between the price and mileage. The polynomial regression model showed a significant improvement, but still relatively low r² score of …. The multiple regression model of the Mercedes brand had the higher r² score of ….

The research showed a clear evolution of techniques used in increasing accuracy. However, as stated in other research, other methodologies can also be used (such as SVM) to obtain a higher r². Also, because manually wrote the functions in Python, the use of external optimized libraries could further improve accuracy, as similar work on Kaggle yielded much higher r² values. Apart from the used techniques, the data could be further cleaned, by for example removing outliers, to obtain a higher accuracy. Polynomial regression, should yield higher r² value than we obtained, but was restrained by unclean data. In addition, our used techniques could easily be expanded to the other brands and even the entire dataset combined to gain more insights.

All sections before this point must fit on six (6) pages. No exceptions. Supplemental material is not allowed. Anything else you want to add to your report (e.g. acknowledgements, author bios, funding sources) is included in the 6 page limit. The report should be in **1 column**. The exception is the section describing the contributions of each team member. **You will be penalized** *−*11 **points per page exceeding this limit**. The max report score is 100.

# Contributions

The contributions section is not included in the 6 page limit. This section should describe what each team member worked on and contributed to the project.

Bibliography

[2] **Vehicle Price Prediction System using Machine Learning Techniques**

[3] **Introduction to Multiple Regression: How Much Is**

**Your Car Worth?**

[4] **Predicting the Price of Used Cars using Machine Learning**

**Techniques**

[5] **Predicting the Price of Second-hand Cars using Artificial Neural Networks**

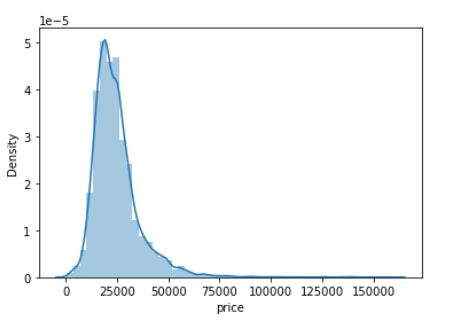
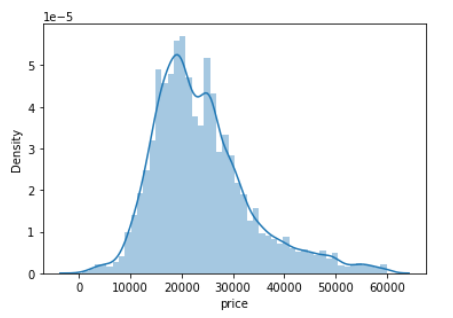
**[6]** [Price evaluation model in second-hand car system based on BP neural network theory | Request PDF (researchgate.net)](https://www.researchgate.net/publication/319412051_Price_evaluation_model_in_second-hand_car_system_based_on_BP_neural_network_theory)

**[7]** Listiani M. 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Master Thesis. Hamburg University of Technology.

[8] Second Hand Price Prediction for Tesla Vehicles

# Bibliografie

|  |  |
| --- | --- |
| [1] | P. Venkatasubbu en G. Mukkesh, „Used Cars Price Prediction using Supervised Learning Techniques,” *International Journal of Engineering and Advanced Technology (IJEAT),* vol. 2019, nr. 3, pp. 216-223, 2019. |
| [2] | N. Kanwal en S. Jan, „Vehicle Price Prediction System using Machine Learning Techniques,” *International Journal of Computer Applications,* vol. 9, nr. 167, pp. 27-31, 2017. |
| [3] | K. Shonda, „Introduction to Multiple Regression: How Much Is Your Car Worth?,” *Journal of Statistics Education,* vol. 3, nr. 16, 2008. |
| [4] | S. Pudaruth, „Predicting the Price of Used Cars using Machine Learning Techniques,” *International Journal of Information & Computation Technology,* vol. 4, nr. 7, pp. 753-764, 2014. |
| [5] | S. Peerun, N. H. Chummun en S. Pudaruth, „Predicting the Price of Second-hand Cars using Artificial Neural,” in *The Second International Conference on Data Mining*, University of Mauritius, Reduit, Mauritius, 2015. |
| [6] | N. Sun, H. Bai, Y. Geng en H. Shi, „Price evaluation model in second-hand car system based on BP neural network theory,” in *18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, United States, 2017. |
| [7] | M. Listiani, „Support Vector Regression Analysis,” *M.S. Thesis, Institute of Software, Technology, and Systems, Hamburg University of Technology,* 2009. [Online] Available: [citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.151.6198&rep=rep1&type=pdf](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.151.6198&rep=rep1&type=pdf) |
| [8] | S. E. Arefin, „Second Hand Price Prediction for Tesla Vehicles,” *arXiv preprint arXiv:2101.03788,* pp. 1-8, 2021. |
| [9] | Aditya, „Kaggle,” 2020. [Online]. Available: https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes. [Opened 19 December 2021]. |



a

b