Data Mining with Linear Regression

Introduction to AI – CMP4294

Birmingham City University

Kacper Popis - 23161791

Kacper.popis@mail.bcu.ac.uk

Introduction

Data mining is a process of abstraction in data, where datasets are cleaned (removal of duplicate entries, formatting data correctly and removing invalid data) “in order to identify patterns and extract useful information” **[1]**, such as graphs or charts.

There are multiple methods of data mining, some of which are:

* Regression – allows for a model to develop predictions based on variables provided by a dataset
* Association rule mining – shows relationships between different variables inside a dataset
* Decision trees – allows for predictions and decisions to be made alongside classification of data inside a dataset
* Neural networks – allows computers to “process data in a way that is inspired by the human brain” **[2]** and allowing them to “learn from their mistakes” **[2]**

However, the most suitable method of data mining for the used dataset will be **Linear Regression**, this is because this method does not need sorted datasets to function, and it provides a simple model to the user which can help the user derive an estimate of their car’s value based on different factors, such as mileage (odometer) or body type of the car. Alongside the model being simple to the user, “the modelling speed of Linear Regression is fast” **[3]** making it a viable data mining method for large scale datasets. This method functions on dependent variables – data from the data set, such as last sold price, make, model, condition - and independent variables – the data that is entered by the user to provide them with an estimate value of their car.

As previously mentioned, the dataset that the model will be trained on is a car sales dataset. In the automotive industry there are multiple methods of evaluating a car’s value, be it hiring an expert to physically examine the vehicle and estimate a value of it based on their condition, or by using an application that scans pictures of the vehicle and provides an estimate value. One company that utilizes this method of evaluation is Arnold Clark. However, there are multiple flaws that come with this method, as the value of the car can change based on external factors that aren’t related to the vehicle itself, for example, the picture that is to be uploaded to the online evaluation (which is done by AI) has to be of great quality and clear, otherwise the AI may misinterpret the poor quality as damages done to the car. With the proposed method of car evaluation, the vehicle owner will be able to gather an estimate value of their car based on previous sales of a car of the same type but with some differentiating factors (most common being mileage)

Dataset & Code Breakdown

The main purpose of the dataset is to provide an estimate of a vehicle’s value based on previous recorded sale records of vehicles. Additionally, this data set can also be used to display the quantities of different vehicles sold over years to show any trends and discover most sold brands. The dataset holds multiple types of variables but to provide the user with most accurate readings the model wont focus on all the variables and only on the ones mentioned below.

* Year
* Make
* Model
* Trim Type
* Body Type
* Transmission Type
* Odometer
* Colour
* Selling Price

The dataset contains over 500,000 entries but not all the entries are valid, as such, the dataset can be cleaned (filtered) using different methods. The methods used are shown in Fig 1.1 and Fig 1.2, where any rows with empty values are to be found and dropped (deleted from the table) alongside checking for any duplicate rows, of which there are none in this dataset. During the process of cleaning, there were over 100000 null values found inside of the dataset, requiring the rows with null cells to be dropped. This is done so because the rows of data that are missing values will not be able to provide proper aid towards finding an estimate as the estimate is meant to be generated based on specific data like car model or make, which can’t be missing inside the dataset as it will fluctuate the average estimated value for all vehicles. Duplicate rows also have a similar effect, the presence of duplicate rows will make the model more bias as the accuracy of estimates will be lowered and result in the estimate being different than what it should be.

A screenshot of a computer program

Description automatically generated

Fig 1.1: Checking for and dropping null rows

A screen shot of a computer

Description automatically generated

Fig 1.2: Checking for duplicate rows

Linear Regression works only with numerical values and majority of the values inside the dataset are non-numerical, therefore each unique value inside of the dataset must have a unique value (weight) assigned to them per column. Each column that utilizes non-numerical values has each instance of an entry found (using the **value\_counts()** method) put into a table format based on the number of times it appears. In Fig 2, the process of encoding the non-numerical values into numerical values is displayed, where each column is looked at one-by-one and has each unique value per column replaced by a number and then adding new columns to the current table holding all the columns from the **.csv** file.

A screen shot of a computer program

Description automatically generated

Fig 2: Encoding non-numerical values to numerical values

The model utilizes the columns shown in Fig 3 – some of which are year, modelEnc, makeEnc – to learn and be able to derive a predicted value when the inputs are given. The columns corresponding to the inputs will be under the variable **X** while the column “sellingprice” will be under variable **y**. Additionally, in Fig 4, it is shown that the model is being tested on only 30% of the dataset, whereas it has been trained on the remaining 70% to provide more accurate predictions to the user. Also, the equation for linear regression ( ) is described and how it is applied into this model via the different methods/functions used to train the model along with the Mean Squared Error (MSE) and Mean Absolute Error (MAE). MSE is a method of evaluating the accuracy of a model, and it works by measuring the “average squared difference between the predicted and the actual target values within a dataset” **[4]** . Whereas MAE is the “difference between the measured value and ‘true’ value” **[5]**.

A screenshot of a computer program

Description automatically generated

Fig 3: Selecting columns to train the model

A screenshot of a computer program

Description automatically generated

Fig 4: Training model on a section of the dataset

In Fig 5 is the function that utilises the model that has been fed the dataset to provide an estimate value of a vehicle, described by the user. To evaluate a different vehicle, the user would have to change the values inside of the squared brackets to the corresponding data of the new vehicle. The order for the numbers in the brackets is: year, make, model, trim, body, transmission, condition, mileage, colour. However, an issue arises from applying this prediction type, as the user is able to enter any values inside of the prediction function, and one possible error that could occur is the user typing a body type that isn’t available for a specific make of vehicles, this will provide the user with an untrue estimation as this type of vehicle doesn’t exist.

A black background with white text and numbers

Description automatically generated

Fig 5: Utilising model to get a vehicle evaluation

The car being evaluated in Fig 5 is of the following details:

* Year: 2015
* Make: Nissan (portrayed by 36)
* Model: Altima (portrayed by 63)
* Trim: 2.5S (portrayed by 102)
* Body: Sedan (portrayed by 36)
* Transmission: Automatic (portrayed by 0)
* Condition: 44/49 (portrayed by 44)
* Mileage: 2300 (portrayed by 2300)
* Colour: Black (portrayed by 1)

The estimated pricing is very close to the actual price of the same car with some of the same details as provide in the evaluation. (Actual price of the same care as the evaluated car can be found on the following website <https://www.cars.com/research/nissan-altima-2015/> )

However, when the data of that example is changed, like changing the transmission from automatic to manual, and dropping the condition of the car, the estimated price of the car fluctuates. And it does so according to the data is has learnt from – the price of the same car as tested above dropped when a manual transmission was applied and the condition depreciated which resulted in a $3000 difference between the two cars, as shown in Fig 6.

A black background with numbers and a black background

Description automatically generated with medium confidence

Fig 6: Modified Vehicle evaluation

The dataset used to train the model can be found at the following website: <https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data/data>

Conclusion

In this report, the model described is used to predict prices of vehicles based on data provided by the user, and to also show comparisons as to how the average estimated price of a vehicle changes based on how different factors fluctuate. The model followed a supervised learning algorithm as it took inputs from the user to generate a prediction based on the information provided by the car sales dataset. The main purpose of this model is to aid companies that use various methods to evaluate vehicle prices to get a better average while also enabling businesses to predict future trends of car sales and their prices via past records. However, this model can also benefit regular people as it will help them get a better understanding of their vehicle’s worth which can help them with selling it as it provides more opportunities to strike a better deal.

References

[1] – Twin, A. (2023). *Data Mining: How Companies Use Data to Find Useful Patterns and Trends*. [online] Investopedia. Available at: https://www.investopedia.com/terms/d/datamining.asp [Accessed 9 May 2024].

[2] – Amazon Web Services, Inc. (n.d.). *What is a Neural Network? - Artificial Neural Network Explained - AWS*. [online] Available at: https://aws.amazon.com/what-is/neural-network [Accessed 9 May 2024].

[3] - Satyavishnumolakala (2020). *Linear Regression -Pros & Cons*. [online] Medium. Available at: https://medium.com/@satyavishnumolakala/linear-regression-pros-cons-62085314aef0 [Accessed 9 May 2024].

‌[4] - encord.com. (n.d.). *Mean Square Error (MSE) | Machine Learning Glossary | Encord | Encord*. [online] Available at: https://encord.com/glossary/mean-square-error-mse [Accessed 9 May 2024].

[5] - Stephanie Glen (2016). *Absolute Error & Mean Absolute Error (MAE)*. [online] Statistics How To. Available at: https://www.statisticshowto.com/absolute-error/ [Accessed 9 May 2024].

‌