# BMW Price Prediction Analysis

Areka Raza, 501071674

CIND820: Big Data Analytics Project

Instructor: Ceni Babaoglu

February 25, 2022

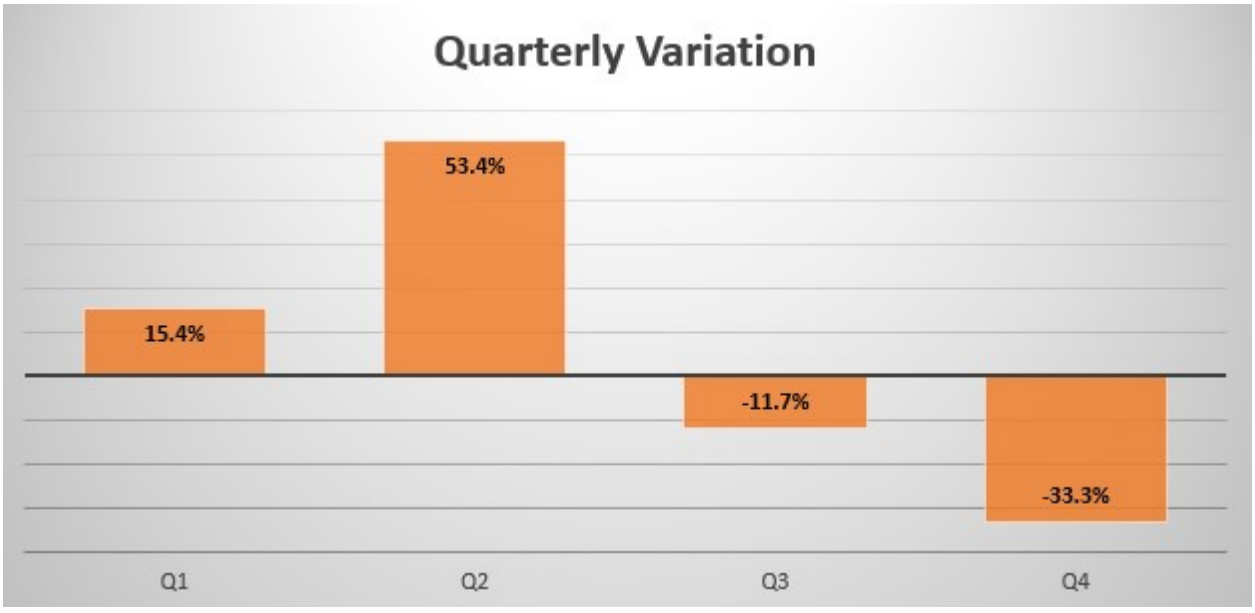
## Introduction

The SARS-CoV-2 pandemic has led the entire world through changes one would not have deemed possible. The world is still recovering from its effects. A notable issue that was created due to SARS-CoV-2, was the microchip shortage. The ramifications of the microchip shortage has led to a decline in electronics and vehicles productions to name a few. Due to a shortage in supply, the demand has increased worldwide, causing an uptick in prices.

At the beginning of 2021, car sales were soaring, and crashed in the second half of the year as there were no available new cars to sell. In the first two quarters, sales had increased by 68.8% in comparison to 2020, however in quarters 3 and 4, sales had decreased by 45%.

**Figure 1**

Quarterly Variation




*Note:* Information retrieved from article published by Focus2Move

Due to the aforementioned decrease in supply, in July 2021, the average used car was being listed at a 12.8% markup, in comparison to July 2020 (Yun, 2021). It was reported by Global News that "pickup truck prices had increased at least 20 per cent" (*Canadian Used Car Sales Rose 5% in 2021 Amid Semiconductor Shortages - National | Globalnews.ca*, 2022). I noticed the increase in aforementioned prices while I was making the decision of whether or not I should purchase a used car. Observing prices of used cars soar intrigued me to learn what influences the pricing of used vehicles, BMW in specific as I work there.

When determining the price of a new vehicle, most individuals know what the dependent variables are, such as brand name, engine power and model type. However, unfamiliarity arises when establishing the price of a pre-owned vehicle. More factors come into light such as total mileage, condition of the vehicle's exterior, condition of a vehicle's interior, how many accidents the vehicle has been involved in and what the model year is, for example. "As a consequence, customers who plan to purchase a pre-owned car often struggle to find an appropriate car within a budget. Even if a customer knows the type of car they want to purchase, it becomes challenging for them to estimate the price of the car" (Rahman et al., 2021). Not everyone has hours to spend researching which factor decreases or increases cost, and by how much. There are simply too many to consider, this presents a dire need to depend on a model that would predict the price and account for all the variables and factors involved.

Machine Learning for price prediction is beneficial for many reasons, notably 4, as stated by Rojewska in 2021:

1. Machine learning can cope with price volatility
2. Machine learning models can analyze multiple data sources at once

3. Machine learning improves the accuracy of price predictions
4. Machine learning can help you improve your profit margin

The aforementioned also make the automotive industry one that benefits majorly from price prediction (Rojewska, 2021).

**Related Work**

Similar studies have been conducted in the past whereby machine learning has been used to predict pricing of used vehicles. Pudaruth conducted a study in 2014 based on used vehicles in Mauritius. The author applied multiple linear regression, k nearest neighbors, naive bayes and decision trees for the predictive analysis. However, this study only consisted of 97 records and posed as a limitation, the sample size was small and only included car makes: Honda, Nissan and Toyota. Pudaruth found that the weakness discovered in decision trees and naive bayes was "their inability to handle output classes with numerical values" (Pudaruth, 2014). Another similar study was conducted by Babu et al., in 2021, they utilized linear regression, random forest algorithm and decision trees. The authors faced similar data issues during their analysis. They found the dataset was not large enough, they did not have many variables and also had missing values (Babu et al., 2021). In addition to this, Kiran S conducted a similar study in 2020 using only a linear regression model. The author predicted an accuracy of 90% with an error of 10% (S., 2020). A research paper by Ashok, K and Samrudhhi, K, focused on K-Nearest neighbor to build a model. The authors obtained an accuracy of 85%, and also validated the model with 5 along with 10 folds (Samruddhi & Kumar, 2020). And lastly the research paper titled "Vehicle Price Prediction System using Machine Learning Techniques" by Kanwal Noor and Sadaqat Jan in 2017 focused only on linear regression to create their model and selected only the highly

influential features, and removed the remainder of the features. They created a model with prediction precision of 98% (Noor & Jan, 2017).

**Methodology**

The purpose of this research is to utilize the aforementioned studies for reference while building a model that has higher accuracy and depicts a clearer picture when predicting the price of pre-owned BMW vehicles. I aim to apply the learnings into my research. The dataset was retrieved from Kaggle which consists of mostly verified vehicle sales in a business to business auction in 2018. Some initial data cleaning has already been conducted, such as removal of vehicles that had engine issues and would skew the results.

The dataset I will be using has 4843 observations, which highlights that records are at an average sample size. If it were too large, it would be difficult to conduct analysis on, and if it were minimal then results would not be depicting an accurate image. Furthermore, I will perform logistic regression to predict price using its attributes and will be evaluating the model performance. Furthermore, I will be using a train/test approach, along with a 10 fold cross validation approach to understand which one yields better results. In addition to this, I will be creating a confusion matrix to evaluate the models and will be using accuracy, recall and precision. Thus, my research will be different from studies done in the past as it will take the limitations of those to create an enhanced model, and will be using a mixture of all techniques utilized.

**Figure 2**

*Descriptive Statistics*

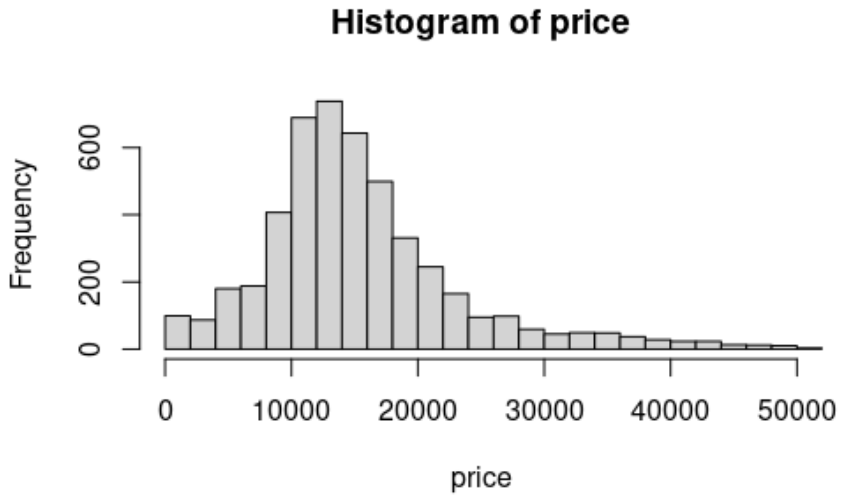
```
> sum(is.na(data))
[1] 0
```

```
> summary(data)
  maker_key          model_key            mileage
 Length:4843        Length:4843        Min.   :    -64
 Class :character   Class :character   1st Qu.: 102914
 Mode  :character   Mode  :character   Median : 141080
                                       Mean   : 140963
                                       3rd Qu.: 175196
                                       Max.   :1000376
  engine_power registration_date        fuel
 Min.   :  0   Min.   :1990-03-01   Length:4843
 1st Qu.:100   1st Qu.:2012-07-01   Class :character
 Median :120   Median :2013-07-01   Mode  :character
 Mean   :129   Mean   :2012-11-22
 3rd Qu.:135   3rd Qu.:2014-04-01
 Max.   :423   Max.   :2017-11-01
 paint_color          car_type         feature_1
 Length:4843        Length:4843        Mode :logical
 Class :character   Class :character   FALSE:2181
 Mode  :character   Mode  :character   TRUE :2662


 feature_2       feature_3       feature_4       feature_5
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:1004      FALSE:3865      FALSE:3881      FALSE:2613
 TRUE :3839      TRUE :978       TRUE :962       TRUE :2230


 feature_6       feature_7       feature_8          price
 Mode :logical   Mode :logical   Mode :logical   Min.   :   100
 FALSE:3674      FALSE:329       FALSE:2223      1st Qu.: 10800
 TRUE :1169      TRUE :4514      TRUE :2620      Median : 14200
                                                 Mean   : 15828
                                                 3rd Qu.: 18600
                                                 Max.   :178500

    sold_at
 Min.   :2018-01-01
 1st Qu.:2018-03-01
 Median :2018-05-01
 Mean   :2018-04-29
 3rd Qu.:2018-07-01
 Max.   :2018-09-01
```

There are 18 variables in the dataset, and none have missing values. Some data seems to be incorrect, for example the minimum mileage being listed as -64 and minimum engine power being listed as 0. Both are not possible and will have to be taken into consideration when conducting the analysis. Furthermore, there are 8 unlisted features in the dataset which will be renamed for better understanding during analysis.

**Figure 3**

*Histogram illustrating the variation in prices and their frequencies*




The independent variable in this research is price. And price is ranging from $100 to $17,8500 in the dataset, with mean being at $15,828. There are outliers in the dataset which will be taken into consideration during the analysis. All of the listed variables will affect the price of the used vehicle, and a correlation analysis will be conducted to understand which dependent variables are highly correlated with price. The hypothesis is that mileage, engine power and registration date will have the strongest correlation with price, followed by model key.

Some further hypothesis are:

- BMW 3 series will be the most popular model in used cars.
- The cost of vehicles with mileage of over an average of 20,000km per year will be low in comparison to the cost of a vehicle from the previous year.

## References

Babu, S. K., Sk., R., N, N., M, N., Kumari, L. K., & B, L. (2021). *VEHICLE RESALE PRICE PREDICTION USING MACHINE LEARNING*. Juni Khyat journal. Retrieved February 14, 2022, from http://junikhyatjournal.in/no_1_Online_21/68.pdf

*Canadian used car sales rose 5% in 2021 amid semiconductor shortages - National | Globalnews.ca*. (2022, February 9). Global News. Retrieved February 12, 2022, from https://globalnews.ca/news/8606989/canada-used-car-sales-rose-2021/

*Focus2move| Canada Auto Sales - Facts & Data 2022*. (2022, January 24). Focus2Move. Retrieved February 14, 2022, from https://www.focus2move.com/canada-auto-sales/

Noor, K., & Jan, S. (2017). *Vehicle Price Prediction System using Machine Learning Techniques*. International Journal of Computer Applications. Retrieved February 14, 2022, from https://www.ijcaonline.org/archives/volume167/number9/noor-2017-ijca-914373.pdf

Pudaruth, S. (2014). *Predicting the Price of Used Cars using Machine Learning Techniques*. Research India Publications. Retrieved February 14, 2022, from http://ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf

Rahman, F., Lanard, A., & Ismat, A. (2021). *Application of Machine Learning Techniques to Predict the Price of Pre-Owned Cars in Bangladesh*. MDPI. Retrieved February 12, 2022, from https://www.mdpi.com/2078-2489/12/12/514/htm

Rojewska, K. (2021, September 15). *Price Prediction: How Machine Learning Can Help You Grow Your Sales*. DLabs.AI. Retrieved February 12, 2022, from https://dlabs.ai/blog/price-prediction-how-machine-learning-can-help-you-grow-your-sales/

S., K. (2020, July). *Prediction of Resale Value of the Car Using Linear Regression Algorithm*. https://ijisrt.com/assets/upload/files/IJISRT20JUL388.pdf

Samruddhi, K., & Kumar, A. (2020, September). *Used Car Price Prediction using K-Nearest Neighbor Based Model*. IJIRASE. Retrieved February 14, 2022, from https://www.ijirase.com/assets/paper/issue_1/volume_4/V4-Issue-3-686-689.pdf

*Used car prices in Canada up 12.8 per cent from last year as microchip shortage continues*. (2021, August 12). CTV News. Retrieved February 14, 2022, from https://www.ctvnews.ca/autos/used-car-prices-in-canada-up-12-8-per-cent-from-last-year-as-microchip-shortage-continues-1.5545158