ITEC 202: ASSESSMENT 2 CASE STUDY ANALYSIS

BY: ASHWANTH SAI POTASSERY SIVAPRASAD

Table of Contents

[EXECUTIVE SUMMARY 5](#_Toc53849166)

[CASE STUDY: DON’T GET KICKED 6](#_Toc53849167)

[OVERVIEW 6](#_Toc53849168)

[DATASET DESCRIPTION 6](#_Toc53849169)

[DATA PRE-PROCESSING AND FEATURE ENGINEERING 7](#_Toc53849170)

[STAKEHOLDER ENGAGEMENT 7](#_Toc53849171)

[DATA VISUALIZATION 11](#_Toc53849172)

[REGRESSION ANALYSIS 18](#_Toc53849173)

[LOGISTIC REGRESSION MODEL 18](#_Toc53849174)

[WHEEL TYPE ID:ONE 19](#_Toc53849175)

[WHEEL TYPE ID: TWO 19](#_Toc53849176)

[WHEEL TYPE ID: THREE 19](#_Toc53849177)

[CONCLUSION 19](#_Toc53849178)

[DOES THE EQUATION FIT THE DATA? 20](#_Toc53849179)

[WHEEL TYPE Id: ONE 20](#_Toc53849180)

[WHEEL TYPE ID: TWO 20](#_Toc53849181)

[WHEEL TYPE ID: THREE 20](#_Toc53849182)

[OVERALL CONCLUSION 20](#_Toc53849183)

[RECOMMENDATIONS 21](#_Toc53849184)

[REFERENCES 23](#_Toc53849185)

# EXECUTIVE SUMMARY

One of the greatest threats to second-hand auto dealerships, is the risk of purchasing functional, damaged vehicles with issues that prevent it from being sold. These unfortunate purchases are called “kicks”. Kicked purchases are due to tampered odometers, mechanical problems that the dealer is unable to address. These purchases could be very costly to dealers.

The objective of this report is to use data analysis to predict whether the buy was a kick. The data set was described, visualized and analysed. Recommendations were provided to the business.

# CASE STUDY: DON’T GET KICKED

## OVERVIEW

Buying and selling used cars is an emerging trend. Purchasing used vehicles provides advantages such as lower ownership expenses and significantly reduced retail price as the car takes a great depreciation hit. The market for used cars is profitable. However, there is a large risk that buyer faces, purchasing faulty vehicles. Faulty vehicles, commonly referred to as kicks. Seemingly functional used cars sold at auctions pose a risk to dealerships as they are at times difficult to identify. These purchases critically hurt the finances of a car dealership business. Working capital is lost as money is blocked on commodities that do not sell.

Kicked cars are sold with tampered odometers, and hidden mechanical issues that the dealers are unable to address. Car dealerships must take the significant risk into account and make sure every car they purchase at auctions is sellable. Greater predictability will reduce the likelihood of bad, costly purchases. The goal of this project is to use data analysis methods to predict whether a car purchased at an auction is a kick.

## DATASET DESCRIPTION

* The data set contains 72,562 instances and 30 features. Datatype of features include categorical, numeric and date type.
* “RefId “is the primary identifier (key) of every instance. “PurchDate” describes the actual date, month and year of purchase. “VehYear” represents the year in which the vehicle was made. “Vehicle Age” is a related feature, it represents the age of vehicle as per the present date. A categorical attribute named “Auction” indicates where the car was purchased from.
* A vehicle’s brand characteristics are represented by “Make”, “SubModel”,” Trim”, “Model” features. Other features such as “Color”,” Transmission”,” Nationality”, “Size”,” WheelType”,” WheelTypeID” are also included. Off 30 attributes, two variables namely “WheeltypeID” and “Vehicle Year” are dependant. Both of these are removed during pre-processing.
* “VehOdo” presents the odometer reading of the car. “TopThreeAmericanName” points out if the vehicle is from one of the three top American manufacturers. Features which represent the different types of costs associated with each vehicle includes “MMRCurrentAuctionAveragePrice”,” MMRCurrentAuctionCleanPrice”,” MMRCurrentRetailAveragePrice”,” MMRCurrentRetailCleanPrice”,” VehBCost”.
* Auction refers to the expected price of the vehicle on auction. Retail refers to the expected selling price of the vehicle. “Clean” refers to the price of the vehicle while on good condition. Clean prices are higher than vehicles in “Average” Condition.
* “VehBcost” indicates the acquisition cost for the vehicle at time of purchase. “
* VNST” describes zip code where the vehicle was purchased.
* Another corelated feature, “VNZIP1” points out the state where the purchase was made.
* “PRIMEUNIT” indicates the level of demand in respect to a conventional purchase.
* “AUCGUART” describes the level of guarantee on the functionality of the vehicle, provided while auctioning.
* “BYRNO” represents the unique number assigned to the purchaser of the vehicle.” IsOnlineSale” indicates whether the purchase was made online. “WarrantyCost” is associated with the price of warranty.

## DATA PRE-PROCESSING AND FEATURE ENGINEERING

* Out of the 72,562 instances there were 330 duplicates, these rows were removed by identifying duplicate RefId values.
* All blank values and spaces were filled with Null. PRIMEUNIT, AUCGUARD were removed as they consisted of almost all missing values.
* RefId was also dropped as it does not have any relation to prediction.
* VehYear was dropped as VehicleAge is a better identifier.
* VNZIP1 was removed as VNST was a sufficient feature for prediction.  This was the same case for WheelTypeID and WheelType. WheelTypeID was kept.
* Purchase date was sorted according to date for easier comprehension of the data. PurchDate was further split into “PurchDay”, ”PurchMonth”. Finally, data was further cleaned using Tableau.

# 

# STAKEHOLDER ENGAGEMENT

Stakeholders are individuals that are actively involved with the work and project and have something to gain or lose as a result of the project (McHale,2019). The need for communication with the stakeholders is essential, and it can either make or break the project. The process of obtaining information from stakeholders is known as Requirements elicitation (An overview of requirements elicitation, 2010).  Requirement elicitation and analysis work in pairs, is about obtaining information and relaying that information to all the concerned parties so everyone understands what needs to be done (An overview of requirements elicitation,2010).

 The lack of stakeholder engagement is caused by passiveness and lack of activity, having trouble understanding concepts, expectation and direction not clearly explained and communicated, complexity of some of the tasks and presence of some demotivators (Ajani, Diagram

Description automatically generatedn.d).

Figure A: five stages of stakeholder engagement

The figure 1a showcases five factors that can help in obtaining information, data and requirements from stakeholders and those factors are explained below:

* Creating the correct picture about the magnitude, importance and impact of the initiative (Ajani,n.d)
  + It is essential for the organization to identify key stakeholders that should be involved in elicitation. It is necessary that a stakeholder’s register is created which includes:
  + Names
  + Department/organization
  + Contact information
  + Interest level
  + Influence level
  + Project Contribution
* Organization modelling should be utilised to demonstrate the organization before and after the project to the stakeholders (Ajani, n.d).
* Utilising activities to keep stakeholders active and alert by:
  + Brainstorming and keeping tracks by creating a list of possible requirements and concerns (McHale,2019).
  + Utilising collaborative games to help find and uncovers otherwise hidden requirements, assumptions and features (McHale,2019).
  + Creating a mind map of all of the concepts, requirements and stakeholders’ concerns, which will help in creating relationships between them.
  + Utilising user stories to demonstrate the stakeholders on how the solution will work
* Organizing workshops and clarifying the goal of these workshops to the stakeholders and the expected contributions (McHale,2019).
  + Workshops are created to strategically drawing out information from stakeholders. The goal of these workshops is to include collaboration and deep dialog, learning, which can help in gathering information.
  + This can be achieved by explaining the stakeholders before the workshops why communication and participations is necessary in these workshops and this information can be useful in the long run.
  + The agenda of the workshop expected outcome and planned discussions and activities should be agreed upon before and communicated to all stakeholders in advance.
  + Utilizing strong facilitation skills and techniques to ensure the workshop goes as planned and avoiding unnecessary discussions to allow the stakeholders to stay on track. This can be avoided by reenforcing the agenda of the workshop and the progress that has been made by these discussions (McHale,2019).
  + When the workshop has ended, it is important to recap on what activities were completed and the progress that has been made by these workshops.
  + The discussions and the progress made in the workshops should be emailed to the stakeholders to help reinforce the current position.
* Utilizing Diagrams and Visual Aids
  + Utilization of visual aids such as UML diagrams, mind maps and presentations can demonstrate these complex ideas in a simpler way to stakeholders, and it can help them in thinking and making informed decisions.
  + The solution if it includes some IT components that are not easily explained, diagrams and models can help in showcasing how the system works and how the system would look (McHale,2019).
  + If various solutions are available and needs some customization which will help it adapt to the stakeholder’s requirements, they should be presented, and any changes or requests should be recorded. The changes and requests should be used to make a better solution.
* Utilization of time wisely
  + Stakeholder’s time is important, and they should only be involved when their involvement is necessary.
  + The workshops should run for the time decided and avoid any discussions that are not in the scope for the meeting.

Graphical user interface, application

Description automatically generatedThe figure 1 b showcases how the involvement of stakeholders can gather their interest and how their influence can help the organization to meet their needs.

# 

FIGURE 1B: STAKEHOLDER MAP

# DATA VISUALIZATION

Chart, bar chart

Description automatically generatedData visualization is when you represent information and data through the medium of graphs. It can allow for relationships and patterns in the data to be discovered and all this can help in making informed decisions. It also allows for huge amount of data to be represented in a simpler way, aiding in understanding it. The case study “don’t get kicked” presented data that had to be analyzed and the analyzed data can help predict if the car purchased at the auction is a bad buy or not.

FIGURE 2A: HISTOGRAM OF VEHICLE AGE

The figure 2a demonstrates the distribution of vehicle age in the dataset. The mode of the data being 4 years, showcasing that the cars being bought in the auction are 4 years old. This dataset also has the characteristics of the bell curve shape. The distribution of odometer reading, vehicle year and nationality demonstrated the same features as the vehicle age.

Chart, histogram

Description automatically generated

FIGURE 2B: HISTOGRAM OF WARRANTY COSTS

The histogram 2b for warranty costs demonstrates a positive skewness distribution, showcased by the long tail at the end. The values at the end were removed due to them being outliers. All of the MMR features represented similar distribution to warranty costs. Other features such as size and etc. didn’t demonstrate any distribution that can be analyzed or present any unusual features.

Chart, bar chart

Description automatically generatedThe figure 2c demonstrates the relationship between auction locations and average of bad buys. It can be demonstrated by the graph that the average of bad buys at ADESA is significantly higher than at Manheim or other. The organization should avoid or do more research when buying cars from ADESA auction location. The average of bad buys is 0.13 and the average of bad buys at Manheim or other is lower than the average, demonstrating that the locations are suitable and will lower the number of bad buys.

FIGURE 2C: AVERAGE OF BAD BUYS AND AUCTION LOCATIONS

Chart, bar chart

Description automatically generated

FIGURE 2D: AVERAGE OF BAD BUYS AND VEHICLE AGE

The figure 2d showcases the relationship between vehicle age and average of bad buys. It is demonstrated by the graph that as the vehicle age is increased, the average of bad buys also increases. The organization should be wary of only buying vehicles which are not older than 5 years as there is a significant increase in average of bad buys if the car is 6 years or older.

Chart, histogram

Description automatically generated

The figure 2e below demonstrates the effect of independent variable odometer reading on the dependent variable average of bad buys. The filter was used to showcase the relationship and it can be grasped that the average of bad buys increases as the vehicle odometer reading

FIGURE 2E: AVERAGE OF BAD BUYS AND VEHICLE ODOMETER READING’S

increases. The result of the graph demonstrates that the organization should avoid buying cars with high odometer reading. As the odometer reading reaches 90K or above, they should be avoided as the average number of bad buys significantly increases.

Chart

Description automatically generated

FIGURE 2G: AVERAGE OF BAD BUYS AND VEHICLE BCOST

The figure 2g demonstrates the Vehicle BCost and the average number of bad buys. The vehicle BCost stands for acquisition cost paid for the vehicle at the time of purchase. The graph represents that the cheaper cars are more likely to be bad buys. The cheaper cars are significantly above the average line, reinforcing the information that they should be avoided by the organization. The outliers from this dataset were removed, to demonstrate the relationship.

Chart, bar chart

Description automatically generatedThe figure 2f showcases the relationship between purchase day and average of bad buys. By interpreting the graph, it can be understood that cars bought on Sunday are more likely to be bad buys as the average number of bad buys on Sunday is significantly greater than any other day (average being 0.1358 and Sunday average of bad buy being 0.2143) , whereas every other day is below the average line. The day Saturday was removed from the data set as no cars were bought. The auction locations being closed on Saturday could be one of the factors.

The organization should avoid buying cars from auctions on Sunday.

FIGURE 2F: AVERAGE OF BAD BUYS AND PURCHASE DAY

# REGRESSION ANALYSIS

## LOGISTIC REGRESSION MODEL

Chart

Description automatically generatedA regression is a statistical method used to determine a relationship between two variables [Beers, 2020]. The two variables that are used are the response variable (historically known as dependent variable) and explanatory variable (historically known as independent variable) [Helmenstine, 2020]. When the two variables are plotted against each other on a cartesian plane, that has x and y axis’, can display a pattern. This pattern can be explained further through using a least square regression trend line. The regression line will have a linear formula that will look similar to this y = mx + c

FIGURE 3A: LOGISTIC REFRESSION MODEL

The regression plot in figure 3a are to determine if the business is a good buy is based on the variables above Average Condition at time of purchase (y-axis) and the years elapsed since purchase (x-axis). The plot has been broken up further into 3 different plots, this is to separate them into their own wheel type ID’s.

### WHEEL TYPE ID:ONE

The green scatter plot illustrates the two, x and y, variables for the cars that have a wheel type ID of one. The equation that this graph has is,

**The above average cost of the car = 1,099.79 x years elapsed since purchase + 4,888.06**

### WHEEL TYPE ID: TWO

The red graph displays that the years elapsed since manufacturers purchase and the above average cost of the cars, with wheel type ID two, has the equation of,

***The above average cost of the car* = 1,188.10 x *years elapsed since purchase* + 6,036.32**

### WHEEL TYPE ID: THREE

The last scatterplot is the blue one, this graph represents the relationship between years elapsed since purchase and the above average cost of the car for the cars with wheel type three. The equation for this graph is,

***The above average cost of the car* = 974.53 x *years elapsed since purchase* + 5,585.08**

## CONCLUSION

When looking at the equations for each plot it gives a great indication of whether or not the business is a good buy. Starting with the gradient (value m) from each of the plots they all demonstrate a positive value. This means that the more years that elapse that greater the above average price is. Some wheel type IDs have a greater gradient than others, for example type ID three (blue) has a gradient of 974.53 which is less than type ID two (red) which has a gradient of 1,188.10. Moving to look at the y-intercept (value c) all graphs have a positive intercept. Which means that for all the wheel type IDs even if no years have elapsed the business will not have any cars that are worth nothing or come with any debt. Overall based just off the equations the business is worth the investment, however, maybe in the long term only taking on certain wheel type IDs that have a high gradient and a high y-intercept.

# DOES THE EQUATION FIT THE DATA?

In determining how well the equation actually fits the data set there is a value called Coefficient of determination (also known as r-squared). The coefficient of determination measures the distance between each dot on the scatterplot and the regression line [Bloomenthal, 2020]. This will then calculate a value between the numbers 0 and 1. One means that there is a strong relation between the two variables and zero means that there is no relationship between the two variables.

### WHEEL TYPE Id: ONE

The correlation coefficient for the cars with wheel type ID: one is 0.3032. This correlation falls between 0.25 and 0.50, therefore the two variables have a weak relationship to each other.

### WHEEL TYPE ID: TWO

Similarly, to cars with wheel type ID: one, cars with wheel type ID: two have a coefficient of 0.3992. This correlation is still weak but is just 0.05 (Approximately) away from having a moderate relationship. In other words, the relationship is at the stronger end of the weak correlation category.

### WHEEL TYPE ID: THREE

Whereas cars with wheel type ID: three have a correlation of 0.2227. This correlation just falls in the very weak relationship category as it is between the values 0 and 0.25. As this coefficient of determination is very weak it’s almost pointless to use this data as a means of prediction.

## OVERALL CONCLUSION

Ultimately, each of the scatterplots have different coefficients of determination. Wheel type ID: one and two have their scatter plots have a weak correlation and wheel type ID: three has a very weak correlation. These correlations are not only caused by having outliers, like the plots contained in wheel type ID: one and two but can also be caused from having a wide spread of data points on the scatterplot. This means that it's difficult to predict the price of the cars in the future and whether the business will make profit due to having a weak or very weak correlation of all the equations. Although it’s difficult to predict the future, having an equation that has a strong correlation with its plots is a great start when predicting the future. But when making a decision about investing in this business it becomes a difficult decision to make as the positive gradient equations shows promise overtime, but the coefficient of determination explains that there isn’t a strong fit with the actual data.

# RECOMMENDATIONS

* It has been observed that the Average of bad buys from ADESA is much higher as compared to Manheim or other states. Cars bought from ADESA should be scrutinized and further inspected before purchasing
* Data analysis showed that with increase in vehicle age, the probability of it being a bad buy increased. The business should completely avoid buying cars which are older than 6 years, as after this period there is a notable increase in the average of bad buys.
* Analysis of odometer data showed that as the odometer value increased, the probability of the car being a kick increased. Cars with extremely high odometer values, however expensive or cheap should be avoided.
* It is also observed that cars bought through auctions on Sunday had higher chances of being a kick. Auction days on other weekdays should be preferred.
* Cheaper cars bought on auctions reported higher chances of being kicks. Purchasing of cheap cars should be avoided.
* Features such as size, transmission did not affect the probability of bad buys.
* Least square regressions are very sensitive to outliers, therefore removing any outliers from within the business will help to have a coefficient of determination that is strong making it the regression equation a good method of predicting future car sales.
* If the business chooses to take on outliers, the business needs to make sure that there is good reason, such as the car could have high demand on the market.
* For a business taking on outliers it’s a good idea to figure out the cause of the outlier, this can include whether the car has any external damage, etc.

# REFERENCES

Australian government 2013, Engaging Stakeholders, Department of the Prime Minister and Cabinet, viewed on 14 October 2020, < <https://www.pmc.gov.au/sites/default/files/files/pmc/implementation-toolkit-3-engaging-stakeholders.pdf>>

Australian government 2013, five-stage stakeholders engagement framework, digital image, viewed on 14 October 2020, < https://www.pmc.gov.au/sites/default/files/files/pmc/implementation-toolkit-3-engaging-stakeholders.pdf>

Ajani, R n.d, “Engaging Stakeholders in Elicitation and Collaboration”, Iiba.org, viewed on 14 October 2020, <https://www.iiba.org/professional-development/knowledge-centre/articles/engaging-stakeholders-in-elicitation-and-collaboration/>

Beers, B 2020,” What Regression Measures”, Investopedia, viewed on 14 October 2020, <https://www.investopedia.com/terms/r/regression.asp>

Bloomenthal, A 2020,” How the Coefficient Of Determination Works?”, Investopedia, viewed on 14 October 2020, <https://www.investopedia.com/terms/c/coefficient-of-determination.asp>

Helmenstine, A 2020, “Independent and Dependent Variables Examples”, Science Notes and Projects, viewed on 14 October 2020, <https://sciencenotes.org/independent-and-dependent-variables-examples/>

McHale, B 2019, “How to Engage and Communicate with Project Stakeholders?”, Project Central, viewed on 15 October 2020, <https://www.projectcentral.com/blog/project-stakeholder-engagement/>

Modernanalyst.com 2010,” An Overview of Requirements Elicitation”, viewed on 15 October 2020, <https://www.modernanalyst.com/Resources/Articles/tabid/115/ID/1427/An-Overview-of-Requirements-Elicitation.aspx>

Stakeholder Map: Who Needs What 2020, digital photograph, smartsheet, viewed on 15 October 2020, < <https://www.smartsheet.com/what-stakeholder-analysis-and-mapping-and-how-do-you-do-it-effectively>>