**COMP5310 Principles of Data Science**

Student name: Yichen Chen

Student ID: 510138903

Unikey: yche3494

**Setup**

In recent years, the price of PC graphics cards(GPU) has fluctuated, and at the same time, the blockchain has emerged. This work aims to solve two problems:

1. **Find the relationship between Ethereum value and PC graphic card – GPU’s value.**

For this question, we establish linear regression and set null hypothesis – GPU value are not related to ETH value, i.e., R – square equal to 0. Alternative hypothesis – GPU value are related to ETH value i.e., there is a significant evidence that R – square is not equal to 0.

After that, we will find what factors (like processor and memory capacity ) of GPU influence this relationship.

1. **Establish two different models, provide decision support for consumers to buy GPU, compare and select better models.**

Establish logistic regression and decision tree and compare.

One of my data selected is “Ethereum Historical Data”, it was downloaded from the Kaggle website located at the URL <https://www.kaggle.com/kingburrito666/ethereum-historical-data>

All Ethereum data from the start to August 2018. The zip is 551 kb in size. I choose one csv fold named EtherPriceHistory(USD).csv which is 32kb in size.

Another date set I choose ‘’Steam Hardware Survey July 2020’’, which is about player's graphics card using data statistics in the game platform steam, <https://www.kaggle.com/kunwardeepak/steam-hardware-survey-july-2020?select=gpu_directx.csv>

Finally I use data from ‘’Ethereum Effect impact on PC parts prices’’ in Kaggle, which is 277MB in size. I choose files ‘FACT\_GPU\_PRICE.csv’ ， ‘DIM\_TIME.csv’ ， ‘DIM\_MERCHANT.csv’ and’DIM\_GPU\_PROD.csv’ , which are totally 36.4mb. <https://www.kaggle.com/raczeq/ethereum-effect-pc-parts>

文本

描述已自动生成FIGURE 1 is my processed data,

Value: the price of ETH, USD

PROCESSOR: attribute of GPU

Memory Capacity: attribute of GPU

AVG: average price of GPU on a certain day

We will quantify reliability using significant testing and coefficient for each factor of linear regression. We will measure effectiveness for linear regression using r-square, and using f1-score, recall and precision for logistic regression and decision tree.

Figure 1

**Approach**

For question 1, since PROCESSOR is not continuous data, we just remove it for convenient. And regard AVG as target data for model.

For question 2, we convert continuous data into categorical data. We divide the ETH value into five levels, and GPU average price into two levels like ‘0’ and ‘1’, 0 means the price level is low, and 1 means high. We do this to simply provide consumers with buying advice like ‘buy’ or ‘not buy’. Finally, we give every processor codenames.

We met some troubles in question 1.

We first establish a multiple linear regression model in order to verify our hypothesis test. For convenience, we removed the processer because it is not a continuous variable. Although the memory capacity is not a continuous variable, its value has a linear effect on the function of the GPU, so we keep it. The trouble mentioned above is that after establishing multiple linear regression, although the p-value values ​​and R-squared are ideal, the coefficient of ETH value is very small, which makes us suspect that it is a useless variable for the regression. To simply verify our suspicion, we just re-established a simple linear regression using only ETH value and GPU value. We got an R-squared value close to 0. This is bad news. In this case, null hypothesis cannot be rejected. We guess this is because the GPU's memory capacity and processor model have a great influence on the price of the GPU. In order to eliminate other interferences as much as possible to judge whether GPU price and ETH price have a linear relationship, I re-established linear regression and chose a fixed processor and memory capacity to explore what we are really interested in. The result can be seen in the next section, it is ideal.

So far, we have been able to determine that GPU value and ETH value have a certain degree of linear correlation, but this is only for a specific processor and memory capacity. After that, we established several linear regressions of different GPU processor and memory capacity to roughly analyze whether and to what extent the processor and memory capacity affect this linear correlation.

Our reason for thinking about Question 2 is based on the fact that consumers do not need to predict the price of a graphics card specifically, but only need to judge whether the price is high or low. Here, due to our poor data preprocessing ability, we can only roughly divide the historical price of the graphics card into two levels, high and low. In fact, this method is not desirable in our retrospective review. People will not justify the high price GPU is not worth to buy because they may have excellent performance (processor and memory capacity) that can support this price. We believe that a better approach is to establish different historical price levels for different types of processors and memory capacities, but at present, this is not allowed in terms of time and technology.

We establish a logistic regression model and a decision tree model and compare them. For the decision tree, choose the decision tree max depth in 2..6, criterion in 'entropy', 'gini' and splitter in 'best', 'random', showing respectively for logistic regression and decision tree whether each test instance is predicted correctly or incorrectly . Determine whether the classifiers significantly different at p<=0.05 according to McNemar's test, using paired t-test(use f-score measure).

**Results**

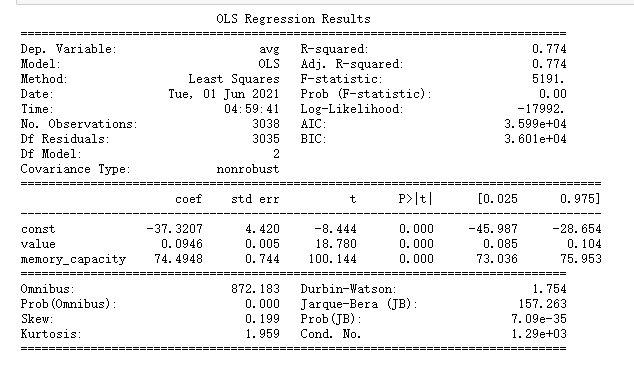
For question 1, according to linear regression evaluation, R - squared is 0.774, p – value is very close to 0, both are ideal. Coefficient of memory capacity is 74.5, but for ETH value, it is 0.1, which shows that ETH value might be useless for this regression. Since memory capacity is a small value compared with ETH value, we do not think it is a negative result. In question 1, we mostly aim to find the relation between ETH value and GPU value, so we did extra simple linear regression for ETH value and GPU value, remove the influence of memory capacity on model, get the R – squared equal to 0.02. It seems that in this condition ETH and GPU value do not have a linear correlation.

Figure 2

手机屏幕截图

描述已自动生成We want to find the reason. After that we select a fixed GPU GTX 1050 Ti. The result shows that R – squared is 0.56 and p – value is 0. It can be concluded that the evaluation R-square shows that regression maintained a medium level linear correlation, we reject null hypothesis that R squared equal to 0. After that, we establish more regression to find what factors (like processor and memory capacity) of GPU influence this relationship. The results (can see them in appendix) shows that R – squared between GPU value and ETH value is about 0.55 when memory capacity greater than 3G, and it is about 0.3 when memory capacity lower than 3G. We conclude that GPU which has more advanced memory capacity may have stronger correlation to ETH value. More information at the end of code file.

Figure 3

For question 2, as you can see in figure 7, 8, 9 shown in appendix, which are the results of 文本

描述已自动生成logistic regression, the prediction is quite good, 9 times get right results for 10 times in total.

Figure 4

Precision, recall and F1-score are all acceptable. You can see result of decision tree in figure 5 and 6, for decision trees, the result is also ideal.

Let us see the details from Figure 10, 11 in appendix

We can see from this picture that if processor smaller than 2.5, which means GTX 1060, and if eth price Less than four-fifths of the highest price in history, the result recommends us to buy.

If higher than that, highly recommended not to buy.

But when processer become more advanced than GTX 1060, eth price is not important anymore.



Figure 5 best parameters

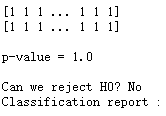
 To compare with logistic regression, we set null hypothesis that the accuracy of decision tree is equal to logistic regression.

Figure 6 p - value

We can compare with figure 4 and 6.

The accuracy is different but not reliably so.

So, we retain the hypothesis that we can choose either model.

No significant evidence that which one is better.

**Conclusion**

From this assignment, I increased my understanding of data science on a practical level. As far as my work is concerned, the preprocessing of data is the most impressive and helpful to my ability of data analysis. I consciously divided the topics into two different types of problems to exercise my ability to deal with different models. By transforming continuous data into categorical data, I successfully dealt with the situation where the data type does not conform to the model. I will be calmer when dealing with the same problem in the future. At the same time, I also realize that I have insufficient data processing ability. Although I have other more ideal methods, I have no means or ability to achieve it. Therefore, I do not think this is a suitable solution for my problem. In question 1, I ignored the possibility of processor as a variable. Although I knew it was very important, just because it is a categorical data, and I am not good at dealing with dummy variables, I gave up. In question 2, I did not classify different types of GPUs, but treat them equally. This will cause all low-performance graphics cards recommended for purchase under this model, and this may just because of their low price, not the impact of the ETH price increase is small at that time.

**References**

1. Graphics processing unit, *Wikipedia*, viewed 10 April 2021,

<https://en.wikipedia.org/wiki/Graphics_processing_unit#Integrated_graphics_processing_unit>

2. Ethereum, *Wikipedia*, viewed 10 April 2021,

<https://en.wikipedia.org/wiki/Ethereum>

3. Kaggle 2021, *Ethereum Historical Data*, Kaggle, 12 April 2021

<https://www.kaggle.com/kingburrito666/ethereum-historical-data>

4. Kaggle 2021,*Steam Hardware Survey July 2020*, 12 April 2021

<https://www.kaggle.com/kunwardeepak/steam-hardware-survey-july-2020?select=gpu_directx.csv>

5.Kaggle 2021, *Ethereum Effect impact on PC parts prices*, 12 April 2021

<https://www.kaggle.com/raczeq/ethereum-effect-pc-parts>

**Appendix**

正方形

描述已自动生成

Figure

表格

描述已自动生成

Figure 7

表格

描述已自动生成

Figure 9

**文本

描述已自动生成**

Figure 10 decision tree

**手机屏幕截图

描述已自动生成**

Figure 11 evaluation of decision tree

图表, 散点图

描述已自动生成

Figure 12 Confidence interval

图表, 散点图

描述已自动生成

Figure 13 prediction of linear regression