Business analytics minor

- [June 2022 BA Minor Exam – Solution Guide](#)

June 2022 BA Minor Exam – Solution Guide

We've selected some actual student responses from the exam to give you a better idea of what is expected. Hopefully this helps you understand where you might have missed points and how you can better prepare for the next exam. This is not meant to be a perfect answer to each question, but more representative of acceptable and unacceptable answers provided by students.

Part I (10 points):

Refer to the Credit Card Fraud report. The dataset this analysis is based on is a list of credit card transactions with some details about each one and an indication whether it was fraudulent (fraud = 1) or non-fraudulent (fraud = 0). A fraudulent transaction means someone other than the credit card owner stole the credit card details and made an illegal purchase.

Question 1 (3 points): What can you say about the relationships between these variables and fraud: repeat retailer, used chip, used pin number, online order? What did you base your answer on? Is there another place in the report that confirms these findings? Explain your answer.

Good Answer:

Fraud and repeat retailer: In the boxplot there is no difference between a repeated retailer and a new one the fraud will have the same chance of appearing. The chances are very high (+-90%)

Fraud and used chips: In the boxplot there is a small difference. Not used chips (the highest bar) are more frequent a victim of fraud. This by 10%. The difference here by repeat retailer is that the information gain is a very small number (0.002) So it is more likely that this information is lest trustworthy,

Fraud and Used pin number: In the boxplot there are no occurences of fraud when used with a pin number. When used without a pin number there is a 10% chance that it is a fraud transaction. The information gain is even lower than used chips so less trustworthy.

Fraud and online order: In the boxplot the online order are 95% fraud while no online orders are +- 38% no fraud orders. These information gains are higher than the pin number and used chips but

lower than repeat retailer. So middle trusted of the 4.

(Note from the lecturers: lower information gain doesn't mean less trustworthy, but it does give us some insight into how well these variables separate the fraud from the non-fraud. Better to mention the Rank widget explicitly as the "other place in the report".)

Insufficient Answer:

Out of these, repeat retailer looks by far the least interesting. The boxplot shows the proportion of fraud appearing totally independent of that variable, and its info gain noted in the Rank widget is so small it's written in scientific notation.

(Note from the lecturers: this isn't a wrong answer, just incomplete.)

Question 2 (3 points): Calculate the recall of fraud = 1 for Naïve Bayes. How does this compare to the Test and Score widget shown in the report? How can you explain this? What would you do differently to evaluate the models if you were making your own analysis? Why?

Good answer:

Formula Recall = number of correctly predicted fraudulent instances / total number of fraudulent instances within the test set data

5120 / 34961 = 0.1464

According to the Test & Score widget output, the Recall for Naïve Bayes is actually 0.9144025. However, this is the Recall based on the average over all classes (as indicated by the Target Class value of the Test & Score), meaning this value represents ratio of correctly predicted instances to the total number of instances over all class values.

If we want to represent the Recall of fraudulent instances within our Test & Score widget, we should set the Target Class value of the Test & Score to fraud = 1, in order to only display the calculated evaluation metrics for that specific class value. The CA will remain the same, however, as this depicts the ratio of all correctly predicted instances over all instances within the test set.

Insufficient Answer:

5120/34961 = 0.146 -- want in confusion matrix zien we dat we er 5120 juist hadden voorspeld van de actual 34961 fraudieuze transacties.

Question 3 (1 point): Agree or disagree: for credit card fraud, it doesn't matter how long it takes to make a prediction as long as it's as accurate as possible. Explain your answer.

Good Answer:

Incorrect, due to the nature of the prediction (a monetary transaction) the system should be able to make a prediction that is as accurate as possible within a limited time frame. This way a client using their credit card won't have to wait long for the prediction to confirm or deny their purchase. It is however in the credit cards company's best interest to try and be as accurate as possible especially when it comes to false negatives since these could result in money loss for their clients and thus would be extremely negative for the business.

(Note from the lecturers: two important points are addressed here. One: how long are you willing to wait for the little machine to confirm everyone's purchase when you're queuing at the shop? Two: accuracy only tells one overall view, different kinds of errors matter. We would much rather cancel some legitimate transactions than allow one large fraudulent transaction go through undetected. So focusing on accuracy isn't ideal here.)

Insufficient Answer:

Disagree, if you take too long to make a prediction, it might already be too late to detect fraud.

Question 4 (1 point): Agree or disagree: the fraudulent transactions are more difficult to correctly identify than the non-fraudulent transactions. Explain your answer.

Good Answer:

If we look at the 4 different techniques than we see that the number of predicted non frauds and these are actual frauds differ from 2013 to 29841. If we compare this with the predicted frauds that are no frauds the number differs from 0 to 4398.

If we compare the numbers:

- KNN dist weight6: 2013 -2776 = +- 700
- Logistic regression: 13798 - 2575 = +- 11000
- Naive Bayes: 29841 - 4398 = +- 25000
- Tree: 9189 -0 = 9182

So if we combine them then we see that the prediction that it is no fraud but the actual data is a fraud is much higher so the prediction for non fradulent is more wrong.

(Lecturer's Note: we probably should look at these are relative rather than absolute numbers since there are so many more non-fraud instances.)

Insufficient Answer:

Agree, we have a lot of examples of normal transactions whereas there is less data to compare to when we look at fraudulent transactions.

(Lecturer's Note: indeed, the class imbalance is part of the problem here – and a known problem in fraud detection research. But some calculations based on the data we have would strengthen the argument.)

Question 5 (1 point): Agree or disagree: the following transaction is definitely not fraud – ratio to median purchase = 5, online order, used pin number. Explain your answer.

Good Answer:

Disagree, when the ratio to median purchase = 5 and we have an online order where the pin number is used, 95.8% of the transactions are not fraudulent, but this means that there remains a 4,2% chance that the transaction is actually fraudulent. We can check this in the tree_viewer.JPG.

Insufficient Answer:

Agree.

If we take a look at the third box plot we can see on that the bottom bar is completely blue, indicating no fraud.

(Lecturer's Note: it's risky to try to use box plots to make predictions.)

Question 6 (1 point): Agree or disagree: if you know the ratio to median purchase price, you can already identify a lot of the fraudulent purchases. Explain your answer.

Good Answer:

disagree. if the value is lower than 4 then you could say with a 97,5% certainty that the purchase is non fraudulent. so saying the data is fraudulent here is a low chance. If the value would be higher, then we only have a 62% chance of the purchase being fraudulent. this is too low to identify anything.

Insufficient Answer:

Agree.

If we look at the rank widget information gain is highest for ratio to median purchase price.

It is also the first split made in the tree. If the distance = 4.0001, 97.5% of transactions are not fraudulent.

(Lecturer's Note: we can identify many non-fraudulent purchases based on this variable. However the other branch in the tree is still highly mixed, about 60%, so we are not able to identify many fraudulent purchases based on this alone.)

Part II (10 points):

There are two datasets for Part II: groceries.csv and Adult (found in Orange)

Question 1 (2 points): Use the groceries.csv file to answer this question. How many rows are there? How many columns? What do the rows and columns represent (you don't need to name each column individually, but describe generally the information they contain)? What kind of analysis can you perform with this dataset?

Good Answer:

Rows: Every one of the 14964 rows represents a transaction for a specific customer

Columns: There are 169 total columns.

- The first column represents the customer_number of the customer making the transaction
- The second column represents the date on which the transaction was made
- The following 167 all represent the amount of a specific item that was included in the transaction, here a ? indicates 0 and a number indicates an amount included in the transaction

We can do different types of analysis on this data, the most useful would be to look at possible associations to determine things like which products are frequently bought together or which products get purchased more on which days of the week.

(Lecturer's Note: several students correctly noticed that there are only ? or 1, so it's a binary feature (bought the item or not).

Insufficient Answer:

How many rows are there - 14962 rows

How many columns - 338

Question 2 (2 points): Continue using the groceries.csv file to answer this question. Set up your analysis. Which widgets did you use? Which settings did you select and/or change? What are some of your findings (list at least 3)? What are some limitations or drawbacks you found with the dataset?

Good Answer:

- Widgets I used:
  - Select columns widget: to discard the meaningless columns. And the columns that are not cooperating at the associatioin rules or frequent itemsets.
  - Association Rules widget. Settings:
    - Minimal support: 0.01%
    - Minimal conf.idence: 95%
    - Max rules: 10.000
  - Frequent Itemsets widget. Settings:
    - Minimal support: 1%
    - Minimal items: 2 (because 1 would be overlapping with Association Rules widget)
  - The settings are set very narrowing for the sake of this exam. In a real-world example, it would be better to include more Rules and Frequent sets.
  - Findings:
    - Frequent Itemsets: when someone buys whole milk or rolls/buns, they will always buy 'other vegetables' as well and vice versa.
    - Frequent itemsets: when a client buys whole milk, there is a really good chance they will buy rolls/buns, soda or yoghurt as well (or more of them), this also works vice versa.
    - Association rules: We can say with a 100% certainty, we can say that if a customer buys prosecco and whole milk/waffles, they will buy sausage, other vegetables and waffles/whole milk as well.
  - Limitations: I found there are too many items to choose from, so that there are too many possible combinations to extract really useful information.

Insufficient Answer:

I use impute to set all the ? as 0, I then discretize to 0.5 so that 0 is the same as = 0.5 and 1 is 0.5

I used frequent itemsets and association rules to see what tend to go together

When you dont buy kitchen utensils you also don't buy baby cosmetics and vice verse which means that when they are bought they are bought together.

(Lecturer's Note: a lot of students seemed to have trouble with the ? even though this is the same format as the dataset we used for the exercises on Association Rules.)

Question 3 (2 points): Now use the Adult dataset that you'll find included in the Orange datamining software.   Check the distributions widget. What do you notice about the target class? Why is it important to know this upfront? Use some other widgets we used during Exploratory Data Analysis to identify some interesting relationships between variables and the target.

Good Answer:

Er zijn veel meer waardes waar y = <50k dan waardes waar y = >50k. Dit is belangrijk in de verdere analyse want een vergelijking tussen beiden wanneer er zoveel verschil is is veel minder correct dan wanneer er even veel van elk is. Ik gebruik RANK en Scatter plot om interessante relaties op te sporen. Zo zien we in RANK dat relatationship, marital status, education en occupation belangrijke featuers zijn, zei hebben een hoge info. gain. In het scatter plot zien we dat dat bij capital-gain en martial-status interessante informatie terug te vinden is. Dit is ook bij bijvoorbeeld education + relationship die een mooie relatie weergeeft tenopzichte van de target.

(Lecturer's Note: we should take note of the class imbalance (many more low income instances than high income) so we can interpret things correctly. Same as the fraud classes from part I.)

Insufficient Answer:

target = y = money above or under the 50k, i noticed that a lot of people in the private sector, they earn more than 50k

Question 4 (2 points): Continue with the Adult dataset.   Set up an analysis that makes sense given the data you have to work with. Which features will you include and how did you decide? Which models have you included and which settings did you try?

Good Answer:

Ik heb eerst Select columns gedaan op data dan mijn data gesampeld in 60-40 procent dan gewoon Logistical Regresion , KNN en Naive Bytes gemaakt.

Als ik laso gebruikt bij Logistical regresion is die better dan ridge.

Als ik in knn mijn neighborgs aanpas dan is die ook verbeterd.

Normaal gezien moest jij ook in uw python script via die tabel zien welke column de groste p-waarde hebt en zo uit u tabel gooien en zien als iets aan verranderd maar mijn python script werkt niet ik heb sommige errors.

(Lecturer's Note: obviously better if your python script works, but using it to eliminate features which are not statistically significant is correct. If you're having trouble with your python script, make sure you have the Timeseries Add-On installed and that you know how to deal with missing values. These are the most common issues.)

Insufficient Answer:

Knn, native bayes, and logistic regresion

Question 5 (2 points): Continue with the Adult dataset. How can you evaluate your models? Which settings did you use for the evaluation? Which ones seem to perform the best? What did you base your decision on?

Good Answer:

De models die ik heb gebruikt zijn KNN met weight = 10, KNN met weight = 20 tussen (deze 2 models wat niet heeveel verschil te zien), Naive Bayes, constant, logistic regression. Naive Bayes heeft bij mij de hoogste CA , Precision, Recall, AUC en kan dus als beste model beschouwt worden.

Wanneer we de target variable veranderen naar <50k dan krijgen we de beste waardes voor CA , Precision, Recall, AUC. Average geeft de tweede beste waardes en target variable et >50k geeft de minst hoge resultaten. Dit alles als we testen op de test data.

We kunnen zien dat er 2 coeficenten niet significant zijn als we het script uitvoeren.

(Lecturer's Note: good to check how they perform on different target classes and look at several models and multiple evaluation metrics.)

Insufficient Answer:

The Test & Score widget determines the accuracy of the model, which is the quality of our model. I determined that the widget kNN is better since it has a higher accuracy as shown by the Test & Score widget.

(Lecturer's Note: accuracy is just not informative enough, especially when we're dealing with a class imbalance.)

File Submission: Create an html report from your Orange workflow.

- Click the Report icon at the bottom of each widget that you want to include in the report
  - You can click the Report icon more than once if you make changes within the widget and want to show it multiple times
- Save your Orange workflow as an ows
- Compress both of these files into a zip file and upload that zip file here.

Important Note: without these files, you may not receive full credit on your answers to the previous questions.