Predictive Analytics Midterm:

For all questions 1-4, please provide the answer and 2-3 sentences around the rationale of why that is the answer. For answers 5-8, you will need to provide a separate document of the code.

1. . What would be the rationale for using median over mean when trying to clean up missing values in a dataset?

Whenever the dataset is heavily skewed, then using a median would make more sense as the mean wouldn’t represent the dataset.

1. . At what statistical value would you consider a value to be an outlier. Walk me through a scenario where you would keep the outlier and one here you would delete it (each scenario should be 2-3 sentences)?

First, we find the Q1 and Q3 respectively. First, we find the center of the data, if the even then it’s the N/2 spot, if the odd it’s either the N/2+1 or N/2 -1 where N is the number of the data.

IQR = Q3-Q1 in this case, and then everything > Q3+1.5\*IQR and <Q1-1.5\*IQR.

Then we find whether the outliers are legitimate,

If they are: then if the outliers are only one value, for instance 1,2,3,4,100, 100 is clearly the outlier, and if it’s way out of range yet still correct, then delete it. If there are more than 10% of the outliers, also in the case above its 20%, then we should consider hold it but use median instead of meaning to represent the data.

If the outliers are deemed as incorrect, then delete the outliers.

All in all, we need to know the legitimacy of the outlier and how that might impact the analysis of the dataset.

1. . In order from most important to least important when interpreting a model, how would you order the following: Coefficients, AIC, and P-Value?

AIC>p-value>coefficient

AIC represents the estimation of generalized errors, since it favors smaller residual error and penalizes for including extra predictors or overfitting, it is a one value that can explain many phenomena. The best reason to choose AIC above p-value is that a model selection would definitely favor a low MSE and errors in general, and AIC incorporates that.

It’s a tricky thing between coefficients and p-value, because coefficients are used as the change of mean response. It can be the most direct way to evaluate the predictor’s relationship against the target. However, since p-value would indicate whether the results are statistically significant, it would be more important than the coefficient.

1. . If you are looking at a confusion matrix and are trying to solve the classification problem of approving a loan for someone or not, which would you want to minimize (assuming 1 that they get approved for a loan and 0 that they don’t), False Positives or False Negatives?

FALSE Positives. If you qualify somebody who isn’t qualified for the loan, then the bank would lose money and never get the loan and interests back.

The next 4 questions are based on the dataset attached in the Module. Your goal is to predict *price* which is a binary variable on how high the price is (1 being high range, 0 being low range)?

1. . In 2-3 sentences state what the business problem is and what a successful outcome of this model would be if you are helping a company like Verizon understand why this data could be useful?

With the specification we have, have we priced our phone too expensive (on the high side）？

What spec more importantly in putting the phone in the high price range? What spec should we focus on if we want to price our phone in the high price range?

Can we infer what the most important feature of the phone is in terms of driving more revenue?

6). Which columns have missing or unclean data? How would you handle cleaning those up? Discuss in 2-3 sentences and have the code in a separate document.

Px\_height: column 12. Missing 41 records.

Fc: 1.01% of records, 15 records are outliers

three\_g: 477 data/23.85% are outliers

7). Model the dataset to **predict** *price*. Which **three variables** have the lowest p-value (closest to 0)? If there are more than three (tie) rank the top three by highest coefficient (by absolute value)? For this response, please provide the variables, p values and coefficients (in absolute and non-absolute form).

Table of Variables (Ranked Absolute Coefficient from high to low)

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **P-Value** | **Coefficient (Non-absolute)** | **Coefficient (Absolute)** |
| three\_g | 0.002 | -1.5421 | 1.54 |
| battery\_power | 0 | 0.0007 | 0.0007 |
| px\_height | 0.009 | 0.0004 | 0.0004 |

Referenced:

Logit Regression Results

Table

Description automatically generated

8). What was the overall accuracy of the model? Additionally, if you are a phone manufacturer using this model to understand what differentiates a high value phone vs low value, which square in the confusion matrix would you want to optimize to reduce and why? For this response, provide the following:

1. Confusion Matrix 2. Accuracy Score and 3. 2-3 sentence on which square to optimize and why.

Chart

Description automatically generated

AIC: 1620.52

BIC: 1716.90

Model Score: 0.577

1. I would argue that False Positive results (identify as a higher end phone, yet it’s not, which is also the **top right** grid) should be the most optimized one. We don’t want to misidentify what features made a high-end phone and waste our money on R&D about said feature/spec that are useless in the end.
2. Our business goal is to identify and then find the most relevant feature to ensure our phone will be perceived and sold as high-end while minimizing risk. The FP grid is the risk square and should be optimized.