**Final Exam**

**MGT 153**

**Professor Zal Phiroz**

Suggested Time: 3 Hours

Please show your work to receive potential partial credit.

Please sign here to confirm the following statement:

“I hereby certify that I have adhered to the university policies regarding ethical behavior in preparing for and completing this exam.”  
  
Please note that any collaboration on this exam is not allowable, and any accusations of cheating will be investigated.

Total Points Possible: 65 Points

**Name (please print) : MAN-FANG LIANG**

**Signature : MAN-FANG LIANG**

**Section 1 (Short Answer and MCQ Questions)**

**Possible Points : 11**

1. (2 pts) An online retailer is interested in predicting *how much* additional revenue they would receive if they offered a 10%-email coupon to customers. Would a classification model which we have covered in class work in solving this problem? Briefly explain why/why not.

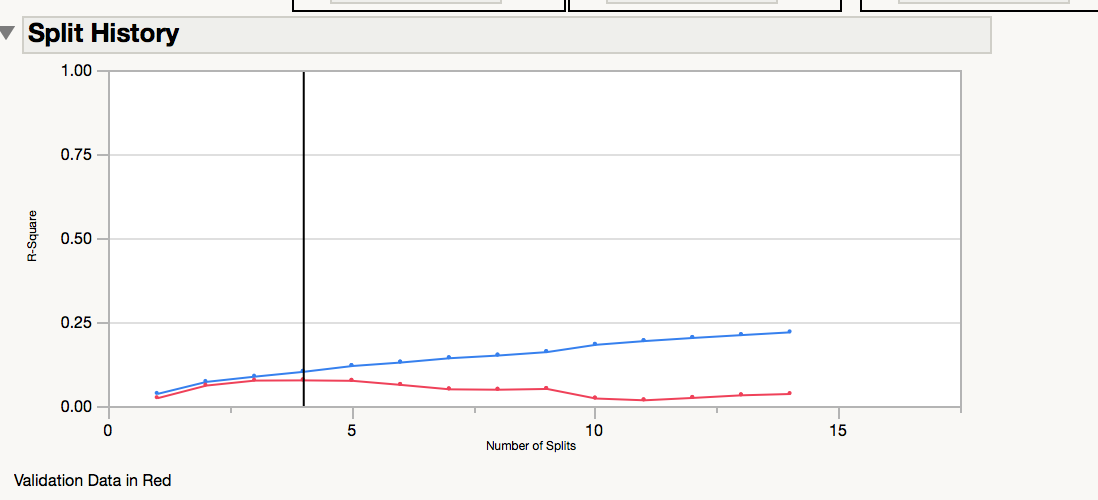
No, a classification model could not solve this problem. Classification is a procedure for predicting the outcome of a binary situation using supporting data. Predicting how much is a numeric outcome instead of a binary outcome, thus classification model would not work in this situation.

1. (2 pts) A private high school is interested in predicting which students are likely to drop-out (i.e. stop coming to class) in the next 2 months, so they can proactively reach out with counseling services and support. They have historical data on student demographics, academic performance, and which students dropped out in the past. How might they do this?

Firstly, they might use a portion of the historical data as a training set to fit the decision tree or linear classification model, where student demographics, academic performance are predictor variables, and whether a student dropped out in the past is a response variable. Secondly, they might use the other portion of the data as a testing set to validate the accuracy of the model. Last, they could predict which student might drop out in the next 2 months with the current data on student demographics and academic performance and know what the most important variables are and how their impacts might be.

**The following information pertains to Questions 3 and 4.**

You and your partner start working for a bank who lends money for mortgages. Your friend uses a decision tree to try and predict if a mortgage is likely to default based on the number of credit cards of the applicant and various other demographic data. He has a large dataset (more than 5000 customers) with many variables, including whether each customer defaulted. He tried to follow the steps we have discussed in class.



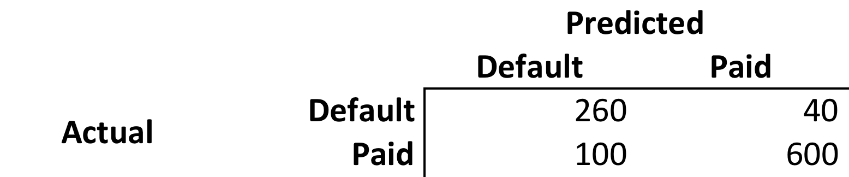
Line 2

Line 1

1. (3 pts) Your friend sends you the Split History from JMP, but for some reason does not use a color printer (See above). Each line shows the Rsquare output. Which line is the training data? Why?

Line 1 is the training data since the r-square increases steadily, which means the more we split, the more we explain with this data. In contrast, the r-square of line 2, the testing data, starts to go down at a certain point. It is the result of overfitting, which means we become too specific in explaining our training data and hurt our prediction in our testing data.

1. (4 pts) Your friend puts aside 1000 data points to use as a testing set. She computed a confusion matrix from this testing data using a threshold of 15%, and it is below. Inspired by your friend, you fit a stepwise logistic regression but instead only put aside 500 data points to use as a testing set. You computed a confusion matrix from this testing data using a threshold of 20%, and it is also below.

Her model:

Your Model:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicted** | |
|  |  | **Default** | **Paid** |
| **Actual** | **Default** | **135** | **45** |
| **Paid** | **10** | **310** |

(**True/False**) Based on the above, are the following statements True or False? **Briefly explain for each**.

1. Her model is more accurate (likely because she used a larger testing set).

False, the accuracy of my model (0.89) is higher than the accuracy of her model (0.86). The size of the testing set is not a point since neither did her model nor my model use a small testing set (both are more than 500, which is large enough).

The accuracy of her model:

(260+600)/(260+40+100+600)=0.86

The accuracy of my model:

(135+310)/(135+45+10+310)=0.89

1. From a business point of view, your model will yield higher amounts of profit because it makes (proportionally) fewer errors.

False, the most accurate is not necessarily the same as most profitable. The bank gain or loss different amounts of money for different combinations of actual results and predicted results. In other words, the importance of each cell in the confusion matrix is different.

We can convert the confusion tables into a percentage view.

Her model:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predict | |
|  |  | default | paid |
| Actual | default | 26% | 4% |
| paid | 10% | 60% |

My model:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predict | |
|  |  | default | paid |
| Actual | default | 27% | 9% |
| paid | 2% | 62% |

The tables show that her model has a less proportion (4%) of incorrectly predicting a mortgage is paid than my model (9%). If the situation of incorrectly predicting a mortgage is paid brings a relatively higher loss for the bank compared to the other combinations of the actual and predicted result, her model will yield a higher amount of profit. From a business point of view, it doesn’t ensure that my model will yield higher amounts of profit, although its accuracy is higher.

**Section 2 (Long MCQ Questions)**

**Question 1**Please use data in **Flights.jmp**

Possible Points: 16

Predicting flight delays can be useful to a variety of organizations: airport authorities, airlines, aviation authorities. At times, joint task forces have been formed to address the problem. Such an organization, if it were to provide ongoing real-time assistance with flight delays, would benefit from some advance notice about flights likely to be delayed.

Airlines will also be very happy to utilize this new information, as it can help them retain customers. A customer who was alerted of a flight delay with enough warning can change their flight, rather than simply canceling it. Flight delays do wind up costing airlines money, but this advance notice can allow them to salvage more customer relationships. If the airlines can be made aware of delays earlier in the process, not only can they avoid these last-minute pileups, but they can keep their customers happier.

The goal is to predict accurately whether a new flight, not in this dataset, will be delayed. The outcome variable is a variable called Flight Status, and it’s status is denoted as delayed or on time. Please use the data noted in **Flights.jmp**

|  |  |
| --- | --- |
| **Variable/Column name** | **Description** |
| FL\_DATE | Flight date |
| FL\_NUM | Flight number |
| WEEKDAY | BusinessDay or Weekend |
| WEATHER | 1 if there was a weather-related delay. 0 if not. |
| CRS\_DEP\_TIME | Scheduled departure time |
| SCHEDULED\_DEPARTURE | Scheduled departure time period: morning, noon, after2pm and evening |
| ACTUAL\_DEP\_TIME | Actual departure time |
| ORIGIN | Origin |
| DEST | Destination |
| CARRIER | Carrier |
| FLIGHT\_STATUS | Flight status: 1 = on time. 0 = delayed |

**The first 1500 rows have been designated as Training Data. The remaining 701 rows are Testing Data.**

Build a **Decision Tree Model** for Question 1-4 to predict **FLIGHT\_STATUS** using the following 5 variables:

WEEKDAY

SCHEDULED\_DEPARTURE

ORIGIN

DEST

CARRIER

1. (2 pts) From the Decision Tree, which destination has the highest probability of being on time?

D. LGA

1. JFK
2. EWR
3. IAD
4. LGA
5. (2 pts) From the Decision Tree, describe the characteristics of the flights which are most likely to be delayed?

A. Carrier (UA, US, DL), DEST (JFK)

1. Carrier (UA, US, DL), DEST (JFK)
2. Carrier (OH, RU, CO, MQ, DH), Scheduled\_Departure (After2pm. Evening), DEST (JFK)
3. Carrier (OH, RU, CO, MQ, DH), DEST (JFK)
4. We cannot tell from this data

1. (2 pts) JMP stopped splitting the decision tree at a certain point. What would happen to the rsquare of the training data if we split the tree one additional time?

A. The rsquare would go up

1. The rsquare would go up
2. The rsquare would go down
3. The rsquare is not a factor with Classification
4. The rsquare won’t change
5. (2 pts) In looking at the model, how many people within the Decision Tree fit the following characteristics: Carrier (UA, US, DL), DEST (JFK)?

B. 20

1. 43
2. 20
3. 33
4. 12

Build a **Linear Classification** **Model** **Using Stepwise Regression** for Question 5-9 using a **threshold of 0.78** to predict **FLIGHT\_STATUS** using the following 5 variables:

WEEKDAY

SCHEDULED\_DEPARTURE

ORIGIN

DEST

CARRIER

1. (2 pts) What is the P-Value of the model?

A. <0.0001

1. <0.0001
2. 0.001
3. 0.1738
4. .6767
5. (2 pts) What is the accuracy of the model?

B. 65%

1. 45%
2. 65%
3. 43%
4. 87%
5. (2 pts) If we wanted to increase the accuracy of the model, what could we do?

B. Decrease the threshold

1. Increase the threshold
2. Decrease the threshold
3. Use only training data
4. Calculate the confusion matrix using less of the testing data
5. (2 pts) Assume the following:

* Every time a flight is correctly predicted to be delayed, there is a gain of $100.
* Every time a flight is predicted to be delayed, but is on time there is a loss of $200.
* Every time a flight is predicted to be on time, but is delayed, there is a loss of $1000.
* Every time a flight is correctly predicted to be on time, there is no gain.

Based on the assumptions above, how much of a gain/loss would this model give us (assume only testing data is used to make this calculation).

D. -$85,100

1. $2,420
2. $12,560
3. -$2,050
4. -$85,100

**Question 2**Please use data in **RadySoap.jmp**

Possible Points: 18

Rady Soaps is an American company which specializes in producing various types of soaps (laundry detergent, dish soap etc.) for sale primarily in Indian retailers.

Recently, the company hired a new president who wants to evaluate the possibility of adding a hand soap category with several brands of hand soap. She notices that purchasing habits in the Indian market work differently in comparison to other markets, and she wants to understand the consumer behavioral patterns of the market, and how to best design each brand to meet the needs of the market.

The data (in **RadySoap.jmp)** includes a subset of 600 people from various regions in India. Included is data on prior transactions, purchase behavior and demographic data, as follows:

|  |  |  |
| --- | --- | --- |
| **Variable Type** | **Variable Name** | **Description** |
| Member ID | ID | Unique identifier for each household |
| Demographics | CLASS | Socioeconomic class (1 = high, 5 = low) |
| DIET | Diet (1 = vegetarian, 2 = vegan, 3 = non-vegetarian, 0 = not specified) |
| SEX | Gender of homemaker (1 = male, 2 = female, 0 = not specified) |
| AGE (1 = 19-25, 2 = 26-30, 3 = 31-34, 4 = 35+) | Age of homemaker |
| EDU | Education of homemaker (1 = minimum, 9 = maximum) |
| ADULT | Number of adults in the household |
| CHILDREN | Number of children in the household |
| TV | Access to TV (1 = available, 2 = unavailable, 0 = unspecified) |
| Purchase summary over the period | BRANDS | Number of brands purchased over the year |
| TRANS | Number of transactions made (number of visits to the store) in a year. |
| PRICE | Average price of purchase per transaction (in USD) |

In using the data above, please answer the following questions:

**Create a Hierarchical Clustering model (using CLASS, DIET, SEX, AGE, EDU, ADULT, CHILDREN, TV, BRANDS, TRANS, PRICE) to answer questions 1 – 4:**

1. (2 pts) One of the Product Managers suggests that based on his overview of the data, having 3 clusters, 4 clusters or 5 clusters of data would present the same value. Do you agree?

a. No. It appears there is more value going from 3 to 4 clusters, as opposed to from 4 to 5.

* 1. No. It appears there is more value going from 3 to 4 clusters, as opposed to from 4 to 5.
  2. No. It appears there is more value going from 4 to 5 clusters, as opposed to from 3 to 4.
  3. It is not possible to tell from the output we have.
  4. Yes, all cluster models provide the same fit and value.

**Use 4 clusters for Questions 2 – 4.**

1. (2 pts) If we assume that people who spend more per transaction are inclined to spend more on hand soap, which cluster would we want to target for the most expensive (highest priced) brand of hand soap ?

c. We would want to target Cluster 3

* 1. We would want to target Cluster 1
  2. We would want to target Cluster 2
  3. We would want to target Cluster 3
  4. We would want to target Cluster 4

1. (2 pts) A Retail Manager suggests that the majority of buyers are from higher socio-economic classes. Based on the data, do you agree?

d. Since Cluster 3 has the most users and the CLASS variable is lowest in this cluster, the managers assumption is correct.

* 1. It is not possible for us to draw this conclusion from the data
  2. Since Cluster 2 offers insight into the CLASS variable, and PRICE variable, the managers assumption is incorrect.
  3. Since Cluster 3 has a low TRANS value, the managers assumption is incorrect.
  4. Since Cluster 3 has the most users and the CLASS variable is lowest in this cluster, the managers assumption is correct.

1. (3 pts) In marketing the various soap brands, a decision is made to only market larger quantities of soap to buyers who have children. The assumption is that the buyers with more children shop more frequently than buyers without children. Do you agree?

d. Since Cluster 3 is comprised largely of buyers who have the most transactions, this assumption is incorrect.

* 1. Since Cluster 4 has buyers with the highest number of children, and also has a low transaction value, this assumption is correct.
  2. Since Cluster 1 has the most buyers who are children, this assumption is correct.
  3. Since Cluster 3 is comprised largely of buyers who spend the most, this assumption is correct.
  4. Since Cluster 3 is comprised largely of buyers who have the most transactions, this assumption is incorrect.

**Create a K-Means Clustering model (using CLASS, DIET, SEX, AGE, EDU, ADULT, CHILDREN, TV, BRANDS, TRANS, PRICE) to answer questions 5 – 8:**

**Use 4 clusters for Questions 5 – 7.**

1. (3 pts) What is the centroid value for Cluster 3?

| **Cluster** | **CLASS** | **DIET** | **SEX** | **AGE** | **EDU** | **ADULT** | **CHILDREN** | **TV** | **BRANDS** | **TRANS** | **PRICE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3 | 3.41150442 | 2.67699115 | 1.95575221 | 3.21681416 | 3.46017699 | 5.10176991 | 2.96902655 | 1.07079646 | 3.07964602 | 28.7035398 | 9.47172566 |

1. (3 pts) Please describe the **buying habits** of a typical consumer in Cluster 1:

c. Purchases relatively few brands over the course of the year, and makes few visits to the store each year. Spends an average amount over the year.

* 1. Purchases relatively high numbers of brands over the course of the year, and makes quite a few visits to the store each year. Spends an above average amount over the year.
  2. Purchases an average number of brands over the course of the year, and makes quite a few visits to the store each year. Spends an above average amount over the year.
  3. Purchases relatively few brands over the course of the year, and makes few visits to the store each year. Spends an average amount over the year.
  4. Purchases a relatively high number of brands over the course of the year, and makes quite a few visits to the store each year. Spends a below average amount over the year.

1. (3 pts) In putting a marketing campaign together for one of the soap brands, an advertising promotion which focuses on targeting customers with a low socio-economic class, a vegetarian diet, who are between the ages of 19-20. Does this make sense?

d. No, this does not make sense, as the clusters do now show these factors to be grouped together in any of the clusters.

* 1. Yes, this makes sense as these factors are well fitted within the model.
  2. No, this does not make sense, as the average number of 19-20 year olds in the data is not accurately represented.
  3. Yes, this makes sense if we only want to target 149 customers.
  4. No, this does not make sense, as the clusters do now show these factors to be grouped together in any of the clusters.

**Section 3 (Long Answer)**

**Question 3**Please use data in **salaries.xlsx** and **satisfaction.txt**

Possible Points: 20

The city of San Diego compensates its employees with a combination of base payment, overtime payment, bonus payment, and benefits. Each year the city evaluates a subset of employees, recording their compensation information, results of a performance evaluation, and results of a satisfaction survey. The city would like to use these data to understand if and how certain compensation structures influence employee performance and employee satisfaction.

The file “**salaries.xlsx**” contains salary and performance evaluation data on each employee.

* ID – Identification number employee
* Year – Year the employee was evaluated
* Employee name – Name of the employee
* JobTitle – Employee’s Position/Job Title
* BasePay – ($) Amount the employee earned in the year they were evaluated
* OvertimePay – ($) Amount the employee earned in overtime in the year they were evaluated
* BonusPay – ($) Amount the employee earned in bonus payments in the year they were evaluated
* Benefits – (binary) Whether the employee received health benefits in the year they were evaluated
* Performance – (1-5 stars) Performance score of the employee for the year they were evaluated

The file “**satisfaction.txt**” contains data on each employee’s satisfaction.

* ID – Identification number employee
* Satisfaction Level – (1-5 stars) Job satisfaction as reported by each employee in the survey for the year they were evaluated

1. (1pt) How many different unique job titles are there in the dataset?

There are 479 unique job titles in the dataset.

1. (2 pts) How much money did the city pay in bonuses to employees who only had a 1 or 2 performance rating? Provide **the total sum of bonus pay.**

The city paid $36709851.54 in bonuses to employees who only had a 1 or 2 performance rating.

1. (3 pts) For each performance rating (1-5 stars), compute **the average of employee satisfaction** for employees with that performance rating. Is there a relationship here? Most managers argue that the more satisfied someone is with their job, the higher they perform. Does this data support that conclusion? Explain.

|  |  |
| --- | --- |
| **Performance rating** | **Average of Satisfaction (1-5)** |
| 1 | 3.045320197 |
| 2 | 2.930232558 |
| 3 | 2.9713999 |
| 4 | 2.977137177 |
| 5 | 3.014917695 |

No, it seems there is no relationship between performance rating and the average of satisfaction. The employees with the lowest performance rating have the highest average of satisfaction, thus the data does not support the conclusion.

1. (4 pts) Starting in 2012, a number of initiatives were introduced to increase job performance amongst employees. As a result, a manager believes that overall, employees are more satisfied with their jobs in 2014, in comparison to 2011. Is there evidence to support this claim? Use A/B testing to justify your response quantitatively (at a 5% significance level)?

No, there is no evidence to support the claim since the average of satisfaction in 2014 is not statistically significantly higher than in 2011. (p-value of 0.39 > 5% significance level)

1. (3 pts) Repeat the question above, but do so separately for employees who received benefits vs. those who did not. Are benefits a confounding variable (at a 5% significance level)?

No, benefits is not a confounding variable since no matter those employees receive or do not receive benefits, the averages of satisfaction in 2014 are not statistically significantly higher than in 2011. (p-value of 0.32 when benefits = 0 and 0.58 when benefits = 0, both are a lot higher than 5% significance level)

1. (2 pts) A consulting team is hired to develop a metric for evaluating ‘potential for a promotion’ amongst employees. They suggest evaluating every employee annually, and calculating the following ratio: ((Job Performance \* 0.6) + (Job Satisfaction \* 0.2)).  
     
   Employees with growth of 1% on average over a 5 year period are fast-tracked for promotion. Those who grow at a rate of 0.99% or lower are re-evaluated and potentially assigned to new roles. If more than half the employees do not grow at least a 1% rate, the city is planning on investing in new training programs.  
     
   Based on your understanding of the fundamentals of a good metric, critique the above approach.

In my opinion, the above metric is a good metric. Tracking metrics can set up incentives and guide decision-making. The above approach not only evaluates the potential for promotion but also alerts management to potential problems early which the city can plan on investing in a new training program. The above approach meets the three important qualities of a good metric which are Simplicity, Comparative, and Influence Behavior. Firstly, the metric is simple since it can be immediately understood how it is calculated from the growth rate of ratio (Job Performance \* 0.6) + (Job Satisfaction \* 0.2), which can be considered as a weighted average of job performance and job satisfaction. And it is a straightforward interpretation for the business. Secondly, this metric is comparative. Since they suggest evaluating every employee annually, they can benchmark this number against the previous number. Lastly, the metric can influence behavior. Depending on the number of metrics evaluated, the city plan to have different actions on employees to affect this number, such as re-evaluating and potentially assigning new roles for the employees growing less than 0.99%. The only thing I would suggest is that they can calculate the growth rate of the ratio on average over a shorter period such as 2 years, which can reduce the risk of metric being not comparative due to employees not staying long enough or leaving soon. It would be more beneficial from the business point of view for tracking the employees’ situation.