**Business Analytics – Finals – Fall 2021**

This take-home exam requires that you adhere to the academic integrity principles specified by the University of Portland. Please indicate that you have read the following and electronically sign the Honor’s pledge.

When you finish reading, please place an X mark in the bracket:

[ ] I will not take assistance from anyone or provide assistance to anyone during the test. Because this is an open-book, open-notes test, I may consult my notes, textbook, Moodle, and other material related to the class.

[ ] I will not provide hints or solutions to any of my classmates during the test. Further, I will not communicate about the test or its contents with my classmates during the test.

[ ] I understand that this test represents the copyrighted material of Dr. Naveen Gudigantala and I will not share this material with anyone (other than the instructor) at any time.

[ ] I agree to abide by the [standards](http://up.smartcatalogiq.com/en/2017-2018/bulletin/University-Academic-Regulations/I-Code-of-Academic-Integrity) stipulated by University of Portland’s academic integrity.

**Honors Pledge:**

As a student of the Dr. Robert B. Pamplin Jr. School of Business Administration I have read and strive to uphold the University’s Code of Academic Integrity and promote ethical behavior. In doing so, I pledge on my honor that I have not given, received, or used any unauthorized materials or assistance on this examination or assignment. I further pledge that I have not engaged in cheating, forgery, or plagiarism.

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student Signature (electronic signature is ok): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Instructions:**

The finals is a take-home, open-book, open-notes exam. The Word document containing your solutions and other solution files (Excel, and/or other files) should be uploaded to Moodle to a link called “Upload Finals files here” by December 5, 11:55 p.m. Please see below for naming conventions. Late upload will result in zero grade. Incorrectly uploaded tests will get appropriately penalized. In other words, if I can’t open the test, or spreadsheet, or Tableau file, you’ll be deducted the whole points associated with it.

The test contains three questions with sub parts. The test is for 100 points and it has seven pages.

Save this test as your last name\_Finals.docx and start working on answering the questions. When you’re done, you must upload the following sets of files:

1. Your last name\_Finals.docx (this test with solutions)
2. Problem # 1 files: Please see the question for what is required to uploaded.
3. Problem # 2 files: Please see the question for what files to be uploaded.
4. Problem # 3 files: Please see the question for what files to be uploaded.
5. I’m setting a maximum limit of 20 files to be uploaded just in case you may feel the need. The maximum limit for each file is 10 MB. If any file exceeds 10 MB, consider zipping it or sending it through OneDrive.
6. I’ll hold office hours on Zoom (https://zoom.us/j/3915496130) on the following days, so if you’ve any clarifying questions, please reach out to me:

Friday (December 3): 4:45-5:45 pm

Saturday (December 4): 10-11 am

Sunday (December 5): 2-3 pm

Other than these times, if you’ve questions, please email them to me by expressing your question clearly and adding any files if necessary. Please expect some lag time.

Question 1 (25)

I want you to build an explanatory model where in the response variable is a *country’s GDP in US $ (Gross domestic product)*. You may choose between *two to five predictor variables of your choice*, visualize the data, build a model, and analyze the results.

Hints:

1. The question is purposefully vague to test your analytical modeling skills.
2. Hint for where to get data: Tableau may have an in-built dataset that may contain useful data; you may also consult the course’s library dataset: <https://libguides.up.edu/bus355_519_543> or use any resource on Web.
3. You may choose data of any of the previous years, but data must be consistent. All of it should belong to one year.
4. Use your best judgment on number of predictors and size of the dataset, etc.

Please answer the following questions. Please include your answers in RED font:

1. Describe your data sources(s) below. Include any assumptions you made about your work. Your excel file’s name should be *solution\_Q1* where in tab1 should be named your *last name\_data1\_Q1* in which you include your final data. Perform your analysis. All your analysis should be shown in the second tab named *last name\_analysis 1\_Q1*. (you may have more tabs in this workbook if you need). If you use Tableau, call the file *last name\_analysis\_Q1.twbx* (packaged workbook file.) (5)

I have chosen the in built sample superstore data and done my analysis on that I assumed that this data is really helpful to explain my insights and it can be seen in that moreover, it was perfect for predictive modeling of my data I assumed that predictive model of my data is really close to my values.

1. What is the difference between an explanatory and predictive model? (2)

In explanatory modeling we explain the outcome of the data, or you can see we use it to test the already existing set of hypothesis and on the other hand predictive modeling as it is also obvious from the name is use to predict the value or new observation.

1. Explain the rationale for including each of your predictors. (3)

I have choose some of these predictors to show how close data is to be predicted and what value is to be predicted in different ways

1. Work on appropriate visualization(s) for this data set (*at least 3* visualizations are needed), paste them below, and describe for each visualization what your insight is. (5)

The following visualization shows sales of the segment, and it also shows the trend that what trend it is following and where it can be gone in future it will be close to this trend line. It is showing time series.

Chart, line chart

Description automatically generated

The following visualization shows the sales in different regions and in addition it tells where there is a greater number of sales on average and where there more sales in total.

Map

Description automatically generated

The following visualization show the actual and predictive sales in different regions and it can be seen that they are close more over the upper and lower bound shows us that these value could lie with in this range with 5% error.

Graphical user interface, application

Description automatically generated

The following visualization clearly shows how much actual and predicted values are synchronized and both of them lie with in the confidence interval so I feel this the authentic model to explain my model also.

1. What model did you choose? Show your model below. (5)

I have chosen the model quantile and it is a table calculation that can provide a calculation based on some expression and predictors.

1. Using appropriate measures, discuss how well your model did? (5)

Well, my model is really close to original value and so that is the reason I believe it is a great value because in future it would help me to predict the future data values. And more over I have also defined the confidence intervals and due to that my data shows where my value could lie more over in the next tab it can be seen the value are so close together which shows its authenticity

Question 2 (40 points)

The city of Seattle sells its used fleet vehicles and equipment through contracted auction services. The dataset in the spreadsheet titled “Q2\_Sold\_Fleet\_Equipment” (to be found in your final exam folder) is a **real data set** downloaded from:

<https://data.seattle.gov/City-Business/Sold-Fleet-Equipment/y6ef-jf2w> (see the link for meta data)

<https://www.seattle.gov/fleet-management/vehicle-auction> (some more information here)

Imagine you’re working as a data analyst for Mr. Calvin Goings, Director of Fleet management, Department of Finance, City of Seattle. Also imagine the department has given this dataset to you and asks you to perform descriptive and predictive analytics on this dataset in a way that your findings should be useful to the fleet management department. I’m deliberately keeping the problem vague to test how you go about performing analytics on this dataset. For this test, here is what you will be evaluated for:

1. Perform descriptive analytics using Excel and/or Tableau and *discuss your questions, findings, and insights*. Answer below in RED font. Please *paste and use two or three key images* here in your answer, but for the rest, refer me to your Excel/Tableau files. This is done to keep the size of this test file manageable. Upload solution files to Moodle. Save the files as Excel\_problem 2 and/or Tableau\_problem 2. The Tableau file should be **packaged workbook** (.twbx). (Note: You must have at least four meaningful visualizations and one dashboard. No limit on the maximum). (10)

Please perform Working in Tableau

So I have done analysis on the data and I have made several graph and visualization from the data so in the very first pie chart I get to know that SPU Department has the highest number of sales overall and OH Department has the least sales price overall. Moreover, from other visuals I interpret that BIDADOO ONLINE AUCTION in 2008 has the highest number of sales price.

Chart, application

Description automatically generated

Chart, histogram

Description automatically generated

1. Perform predictive modeling using Analytical solver and describe your question, description of your target and predictive variables, what models/algorithms did you pick, how did you model the problem including any assumptions made, how did you evaluate the model, and findings? Building good models requires you to think critically at the predictors and try to use as much information and as creatively as possible. Answer below in RED font. Only include two or three key tables/images here and for the rest refer me to your Excel work. Upload solution files to Moodle. Number them as Solver\_problem 2.1, Solver\_Problem 2.2, etc. (for different models/algorithms). I expect you to run three appropriate algorithms and provide your boss with the best predictive model. (20)

(Note: Algorithms are: multiple regression, logistic regression, CART, Neural Nets, etc.; Model: there could be several different models made within each of these algorithms – remember variable selection in multiple regression? You must run three algorithms).

Question: Develop a model to predict the sale price of vehicle using different predictors.

Response (Target) Variable: Sale Price

Predictors:

Model Year,

Make (Chevrolet, Ford, Toyota),

Model (Astro, Camry, Colorado, Prius, Transit Van),

Equipment Type (Sedan Small, Sedan Medium, Truck Light, Van, MiniVan),

**Model 1 - Regression Model:**

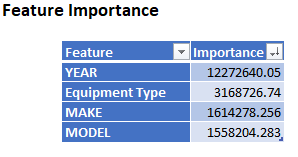
Regression Equation:

Sale Price = 4694.13 + 3958.93\*Year – 360.67\*Make + 609.07\*Model – 300.01\*Equipment Type

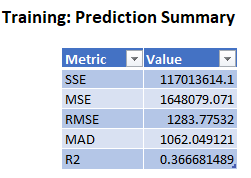
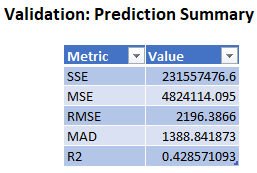
Only Year variable is significant in predicting the price of vehicle as it has p-value less than 0.05.

R-square of model is 0.2971 which means 29.71% variation in sale price is explained by model year, make, model and equipment type. This model is not good for making prediction as R-square is low and the three variables are also not significant.

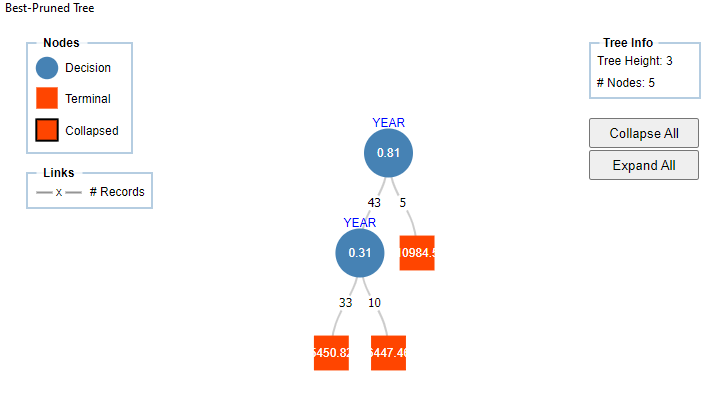
**Model 2 – Regression Tree Model:**



The most important predictor to predict the sale price is Model year and second most important predictor is equipment type.

The error metrics are increased from taining data to validation which shows no indication of overfitting.



**Rule – Best Pruned Tree:**

Left most node:

If (Year<0.8117) and (Year<0.3130), then Sale Price is $5,450.82.

*Modification:* If (Year <0.3130), then Sale Price is $5,450.82.

Second left node:

If (Year<0.8117) and (Year>=0.3130), then Sale Price is $6,447.46

*Modification:* If (0.3130<=Year <8117), then Sale Price is $6,447.46

Right Node:

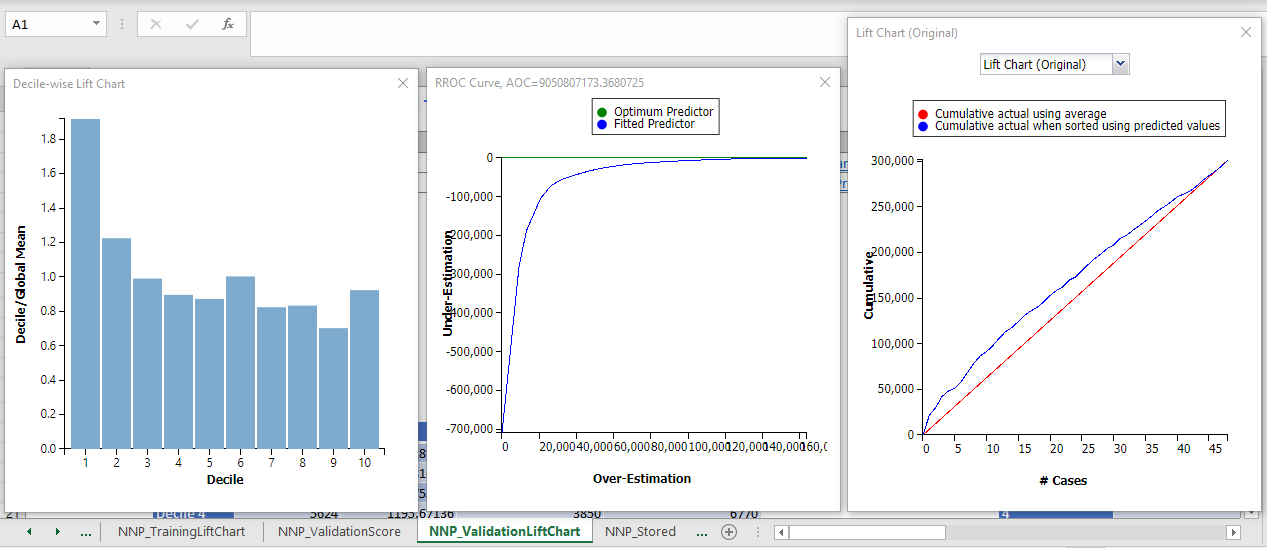
If (Year>=0.8117), then Sale Price is $10,984.5.

The lift of first decile is higher than 1.0 which shows that the model is performing reasonably good. The lift of first decile is 1.9 which shows that the selection of top 10% cases will yield 1.9 times good result as compare to randomly selected cases.

The RROC curve is close to top corner which shows the model is performing well. The smaller the AOC, the better the performance of the model. So, it can be predicted that the model is a good fit to the data.

The lift chart shows that the cumulative actual when sorted using predicted value line is above the cumulative actual using average line which means the model is performing well for making prediction.

**Model 3 – Neural Network:**



The lift of first decile is higher than 1.0 which shows that the model is performing reasonably good. The lift of first decile is 1.9 which shows that the selection of top 10% cases will yield 1.9 times good result as compare to randomly selected cases.

The RROC curve is close to top corner which shows the model is performing well. The smaller the AOC, the better the performance of the model. So, it can be predicted that the model is a good fit to the data.

The lift chart shows that the cumulative actual when sorted using predicted value line is above the cumulative actual using average line which means the model is performing well for making prediction.

1. Please write a memo below, addressed to the fleet management director, describing your findings from both descriptive and predictive analytics in less than 500 words and why such findings are helpful to the organization. Assume your fleet director has no knowledge of advanced analytics. You may include a maximum of two key images as an *appendix* to the memo (and for the rest refer me to uploaded files). *Appendix* is part of the test solution here. Answer below in RED font. (10)

Please write memo

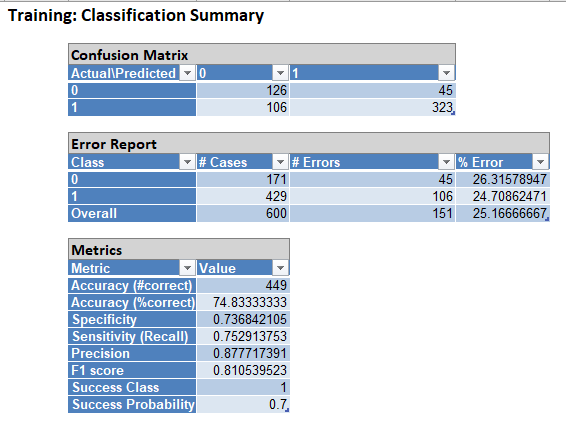
**Question 3 (35 points)**

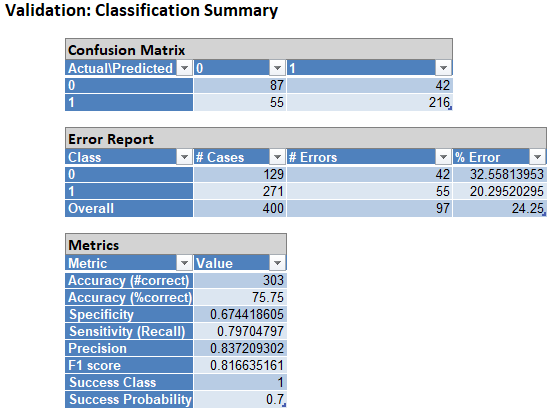
The German credit dataset has 30 variables and 1000 records, each record being a prior applicant for credit. Each applicant was rated as “good credit” (700 cases) or “bad credit” (300 cases). The following figure shows the values of these variables for the first few records.



The variables are described in the last page. A model needs to be developed based on the data presented in the case to classify a new applicant into a good or bad credit risk.

1. Use the data in “Q3\_German credit data” spreadsheet and save the dataset as “Q3\_German\_solution”. No need to develop any dummies in this case as all the predictor variables are either numeric, binary, or ordinal. Split the data into training (60%) and validation (40%), and develop a classification model using analytical solver using the default options. No need to run feature selection but select to show the lift charts and full report. Please use the Logistic regression model. No need to write anything below. I’ll check your Excel output. (7.5)
2. Report the accuracy and error matrices for training and validation datasets below in red. How is the model doing? Is overfitting a problem? Describe below. (7.5)





The overall error of Training data is 25.17% and the overall error of validation data is 24.25. This shows a decreased in error from training data to validation data which means there is no chance of overfitting. Although the error is high for both data, so there is chances of improvement in model.

The accuracy of training data is 74.83% and the accuracy of validation data is 75.75%, so, there is an increase in accuracy of model from training data to validation data. This indicate that the model is performing well.

1. Let’s say the bank gave you some details on average net profit for different predictions you make for each case. See the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Actual | Predicted | Average profit ($) | Explanation |
| Good (credit risk) (1) | Good (1) | 100 | Ideal; you’ll make money on these applicants |
| Good (1) | Bad (0) | 0 | Missed opportunity; no profit to be made |
| Bad (0) | Good (1) | -500 | You want to avoid this mistake;0/;’ (because you gave credit to applicant with poor credit) |
| Bad (0) | Bad (0) | 0 | Ideal |

Use the numbers from each cell of the classification/confusion matrix of your validation dataset and calculate and report the overall net profit. Please show all your work in red below. (8).

|  |  |  |
| --- | --- | --- |
| **Actual \ Predicted** | **Bad (0)** | **Good (1)** |
| **Bad (0)** | 0\*87 = $0 | -500\*42 = -21,000 |
| **Good (1)** | 0\*55 = $0 | 100\*216 = $21,600 |

1. You are asked to help your manager further. All your work for this problem will be done using validation dataset scores tab. Use the validation dataset; consider the “predicted probability of success (or, PostProb: 1)” column (success means 1) and sort the validation dataset on “predicted probability of success” from the highest to the lowest. After doing this, please do the following: (12)
   1. For each case, list the net profit as either 100 or -500 based on actual value of target variable (100 if actual value of the target variable is 1; otherwise -500). <<answer in your spreadsheet solution>> (3)
   2. Add another column for cumulative net profit. Calculate cumulative net profit based on values in part a. <<answer in your spreadsheet solution>>(3)
   3. There are 400 cases in validation dataset. Every chunk of 40 cases now become a decile. Mark the first five deciles (simply highlight each of the five rows representing the end of a decile in yellow). <<answer in your spreadsheet solution by highlighting those rows in red>> (3)
   4. How far into the validation data should you go to get the maximum net profit? Report in terms of the number of deciles. Is it decile 1, or decile 2, or decile 3, or decile 4, or decile 5. Which one and why? <<answer in your spreadsheet solution by highlighting the cell(s) in read>> Report your answer below in red and explain your rationale. (note: it may be possible that the maximum net profit may not coincide with the last row of the decile. It’s ok. Just report the decile in which you find the maximum net profit). (3)

In first decile, we get the maximum net profit of $1000.

Please see next page for variable descriptions. This last question marks the end of the test.

* Here are the files you’ll turn in on Moodle. Incorrectly uploaded/missing files will not get credit.
  + This Word document with your answers. Save file as lastname\_Finals.docx
  + Files related to problem 1
  + Files related to problem 2
  + Files related to problem 3

