1. **Exploring - (15 points)**

**Schools in Portland Public School District are affected by churn. In this case, churn is defined as students switching school’s midyear. There are reasons why this occurs. Investigate this issue. Here are links:**

[**http://www.oregonlive.com/education/index.ssf/2018/03/portlands\_housing\_crisis\_is\_an.html**](http://www.oregonlive.com/education/index.ssf/2018/03/portlands_housing_crisis_is_an.html)

[**https://projects.oregonlive.com/eviction/map/**](https://projects.oregonlive.com/eviction/map/)

1. **Describe how you would go about developing a churn prediction model and discuss how you would derive input variables.**

**Solution:** The steps in building a churn prediction model is as follows:

1. **Business discovery/Problem understanding**

The purpose of this step is to understand the education industry, market and the organization geographically.

1. **Perform data discovery**

This step would help me to study and understand the data to analyze decision variables, constraints and different measures involved. This would help me to predict the possibility of churn (Y/N) using classification model.

1. **Understand the purpose of the Data mining project**

To predict the churn rate, there could be another model with a continuous dependent variable using linear regression model.

1. **Obtain the necessary data**

Required data for the analysis might be from different sources and formats and could be a large database. A sample of the data needs to be obtained and made sure that it reflects the original dataset with all the attributes in it. If there are multiple data files with required information, data blending must be performed in order to get complete and accurate data. The dataset is expected to have students’ personal and demographical details.

1. **Data Exploration and cleaning**

Data must be checked for accuracy and consistency. Missing data must be imputed to populate records. The outliers must be removed, and data must be normalized to make well distributed. Correlation between the variables can be checked using scatterplots, box plots etc.

1. **Data dimension correction**

Those attributes or dimensions that may not contribute to the analysis can be removed. New variables can be generated to support the analysis while making sure there is no data duplication which would account for model inaccuracy.

1. **Data Partition**

Divide the data in to training, testing and validation datasets.

1. **Choose the data mining methodology**

Data mining methods can be in two ways. Either prediction of possibility of churn which would use classification or using a linear regression model to predict the number of churns.

1. **Build model and evaluate the results**

For prediction, various algorithms such as logistic regression, neural network, decision tree, knn, Naïve Bayes etc. can be used. For prediction of number of churns, linear regression, Poisson regression, Bayesian regression etc. can be used.

1. **Provide a preliminary model (conceptual) in the form**

**Churn\_Value = M(x1, x2, … Xn)**

**Where M is a model. Here you want to suggest the variables for the model.**

**Solution:** Churn\_Value = M(Student\_Number, Name, Class, Address, Gender, Age, Parents’\_Income, School, Siblings, Vehicles\_Owned, Ownhouse/rented, Parents\_Savings)

1. **Create your own sample data (fictitious) to determine the churn\_Value. Keep it small. You only want to provide an example. Provide a sample of the data.**

**Solution:**

**A screenshot of a cell phone

Description automatically generated**

1. **Understanding measures (5 points)**

**We have used several classification costs besides accuracy. Below is the r-code for deriving a specific measure – F1-measure (also known as F score for the given table.**

**df #data frame with contingency table**

**Prediction One Zero**

**1 One 37 43**

**2 Zero 19 131**

**# Precision: tp/(tp+fp):**

**df[1,1]/sum(df[1,1:2])**

**[1] 0.4625**

**# Recall: tp/(tp + fn):**

**df[1,1]/sum(df[1:2,1])**

**[1] 0.6607143**

**# F-Score: 2 \* precision \* recall /(precision + recall):**

**2 \* 0.4625 \* 0.6607143 / (0.4625 + 0.6607143)**

**[1] 0.5441177**

**A. Review the results for two models below. Calculate the accuracy (% correct) precision, recall, and F-Score and enter the results here.**

**Solution:**

**Model 1**

**PREDICTED.E PREDICTED.A  
TRUE.E         512       488  
TRUE.A          11         899**

Assumption – E belongs to class 0 and A belongs to class 1.

1. **Accuracy** = (Total correctly predicted Es + Total correctly predicted As)/ (Total Observations)

= (495+98797)/ (495+98797+1203+505) x 100%

= **98.31 %.**

1. **Misclassification rate** = (Sum of incorrect Es and As)/ (Total Observations)

= (1203+505) / (495+98797+1203+505) x 100%

= **1.69 %**

1. **Recall** = (Number of As predicted correctly)/ (Actual number of As)

= 98797/ (98797+1203) x 100

= **98.79%.**

1. **Precision** = (Number of As predicted correctly)/ (Predicted Number of As)

= 98797/ (98797+505) x100

= **99.49%.**

1. **F1 Score** = Harmonic mean of precision and recall

= 2PR/(P+R)

= **0.991.**

**Model 2**

**PREDICTED.E PREDICTED.A  
TRUE.E         495         505  
TRUE.A        1203       98797**

Assumption: E belongs to class 0 and A belongs to class 1

1. **Accuracy =** (Total correctly predicted Es + Total correctly predicted As)/ (Total Observations)

= (495+98797)/ (495+98797+1203+505) x 100%

= **98.31%**

1. **Misclassification Rate =** (Sum of incorrect Es and As)/ (Total Observations)

= (1203+505) / (495+98797+1203+505) x 100%

= **1.69 %**

1. **Recall =** (Number of As predicted correctly)/ (Actual number of As)

= 98797/ (98797+1203) x 100

= **98.79%**

1. **Precision =** (Number of As predicted correctly)/ (Predicted Number of As)

= 98797/ (98797+505) x100

= **99.49%.**

1. **F1-Score =** Harmonic mean of precision and recall

= 2PR/(P+R)

= **0.991.**

1. **Discuss the differences in the scores between them. Which model is better? Why? Which measure(s) are more useful and why?**

**Solution:** Model-2 has higher precision and recall compared to Model-1. So, in my opinion, Model-2 is a better model. As precision and recall important in order to know how many of the predicted positive are actually positive and true positive. Precision and Recall are the better model metric compared to Accuracy which is used to select a best model. On the other hand, F-score is also high for Model-2 which shows a balance between Precision and Recall.

1. **Logistic regression (10 points)**

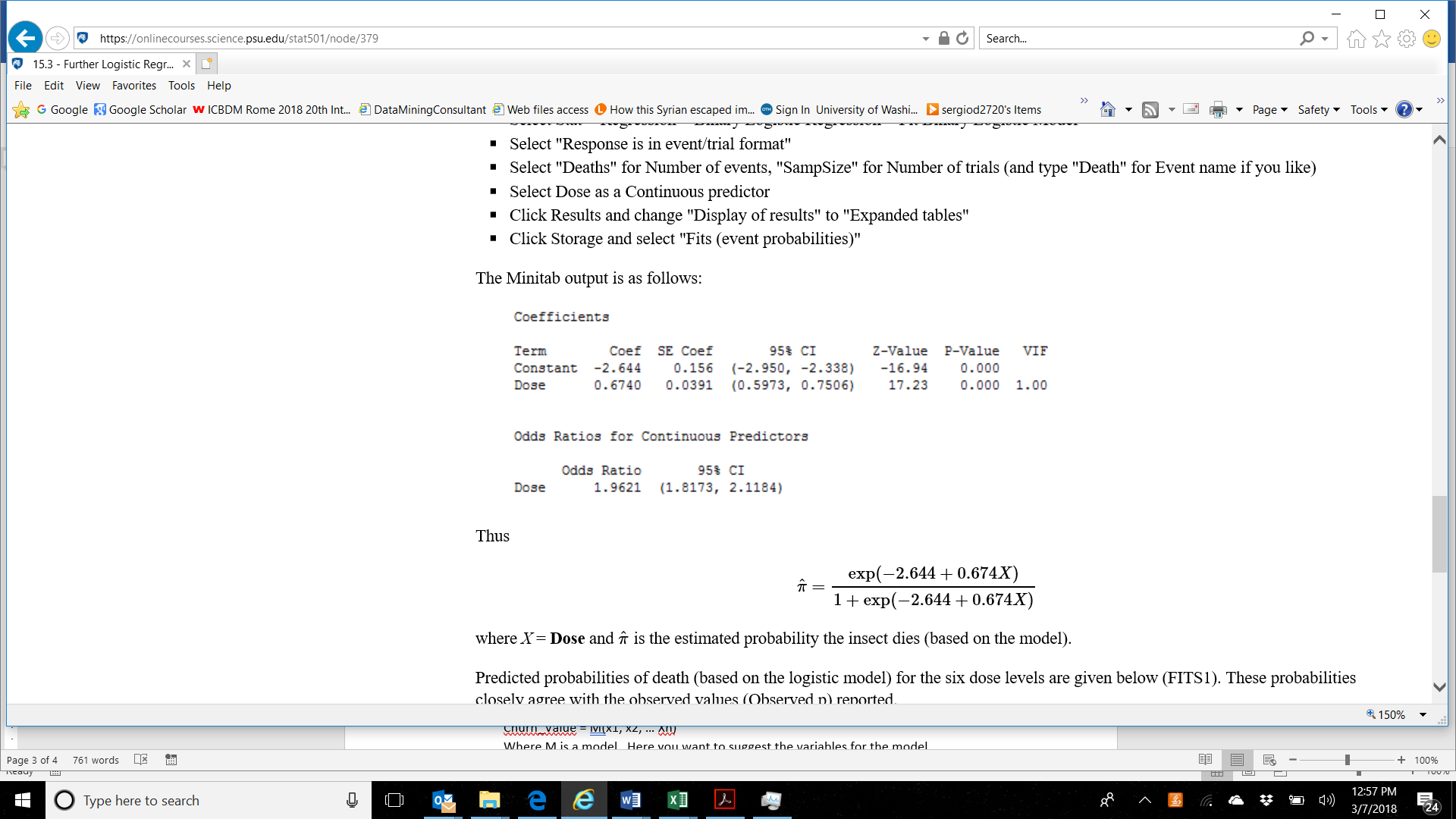
**An experiment is performed to test the effect of a toxic substance on insects. The data is from the textbook, Applied Linear Statistical Models by Kutner, Nachtsheim, Neter, & Li.**

**At each of six dose levels, 250 insects are exposed to the substance and the number of insects that die is counted.**

**Below is a summary table of the data:**

|  |  |  |
| --- | --- | --- |
| **Dose** | **Sample Size** | **Deaths** |
| **1** | **250** | **28** |
| **2** | **250** | **53** |
| **3** | **250** | **93** |
| **4** | **250** | **126** |
| **5** | **250** | **172** |
| **6** | **250** | **197** |

**Below is the output from the program**

****

* 1. **Interpret the output.**

**Solution:** The output represents that the odds of dying increases by 96% as the dose increases by factor 1. Also, that the log (odds of death) increases by 0.67 increases in death with the increase in dose.

* 1. **Use the data set called LRTest.xls associated with this test to conduct logistic regression. Confirm that you get the same results as the output shown. Show your output here. Note: you can use any software.**

**Solution:**

A screenshot of a cell phone

Description automatically generated

* 1. **Calculate the observed probabilities as the number of observed deaths out of 250 for each dose level.**

**Solution:** Observed Probability = Deaths/ Sample size

|  |  |
| --- | --- |
| **Dose** | **Observed Probability** |
| 1 | 0.112 |
| 2 | 0.212 |
| 3 | 0.372 |
| 4 | 0.504 |
| 5 | 0.688 |
| 6 | 0.788 |

* 1. **Calculate the predicted probabilities**

**Solution: Predicted Probability =** π = Dose + exp (constant−co-efficient of dose (Dose))

π = Dose + exp(2.644−0.674(Dose))

|  |  |
| --- | --- |
| **Dose** | **Predicted Probability** |
| 1 | 0.1224 |
| 2 | 0.2148 |
| 3 | 0.3493 |
| 4 | 0.5130 |
| 5 | 0.6739 |
| 6 | 0.8022 |

1. **Association rules – (10 points)**

**Describe an application of association rules method. Do not base your discussion on purchases such as grocery store transactions. You want to apply it in another area – marketing, healthcare, finance, law enforcement, fraud, etc. You want to include in the discussion of the benefit, the specific aspects of the application, and the potential itemset content.**

**Solution:** Association rule mining is a data mining concept to look for trends or patterns in databases like transactional, relational or organization repositories in various business sectors. Government institutions and other big organizations collect data overtime forming large databases. Association rule method is useful in finding correlation between decision variables and patterns in the data. Association rule mining is applied in various analyses like classification, clustering, cross-marketing, basket data etc.

This concept of association rules or affinity analysis is designed to find general association patterns between items in large databases. Suppose an e-commerce tourism company is collecting data about the hotels booked over years, the company can apply affinity analysis to generate patterns associated with the highest number of bookings. This analysis can help the company to strategize their marketing campaigns and bundle products based on the historical data. Similarly, a hospital can use association rules method to predict future symptoms for returning patients based on the analysis “what symptom is followed by what other symptom”. *(Data Mining and Business Analytics Text).*

A potential itemset for association rules method should include the historical data, what-if business questions, major business considerations and decision variables to be able to make predictions.

1. **Neural network (20 points)**
   1. **Conduct 8 experiments using any data set and any neural network implementation (does not have to be R).**

**You will vary the learning rate (2 different ones) as well as the number of hidden layers (2 different ones) as well nodes per layers (2 different ones).**

**Solution:** I chose Haberman’s Survival dataset from Kaggle for my Neural Network experiment. The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. Link to the dataset: <https://www.kaggle.com/gilsousa/habermans-survival-data-set>

The survival status is 1 for the ones who survived breast and 2 for the ones that did not survive the breast cancer. The data was partitioned as 60:40 for Training and validation respectively. Firstly, I considered Regularization factor as 0.1 with 3 input nodes, 3 nodes in hidden layer #1 and 1 node in output layer. Performed 4 different experiments with varying learning-rate (0.3, 0.85, 0.01, 0.95). When compared the accuracy for all the four, learning-rate of 0.3 represented the highest average accuracy.

Next, I considered Regularization factor as 0.1 with 3 input nodes, 2 nodes in hidden layer #1, 1 node in hidden layer #2, and 1 node in output layer. Performed 4 different experiments with varying learning-rate (0.3, 0.85, 0.01, 0.95). When compared the accuracy for all the four, learning-rate of 0.3 represented the highest average accuracy.

**Output:**

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**b. Compare the results and discuss the outcomes.**

**Solution:** When compared both the results, we can observe that the Learning\_rate of 0.3 has the highest average accuracy for both the models. By comparing the overall accuracy for different Learning\_rate we can see that the second set of experiments have a higher average accuracy. By this we can conclude that more nodes led to a better model accuracy and model fit. Smaller learning rate led to more accuracy than bigger learning rate. However, extreme values of Learning\_rate caused the models to become less stable, indicating that the model is still significant and can be optimized to make the model better.

1. **K-means. (15 points)**

**Customer segmentation identifies customers into distinct groups.  You want to examine a set of customers and identify those that behave similarly to tailor marketing.  You want to group them based on purchases of four products.**

Here is the data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Customer#** | **ProductA** | **ProductB** | **ProductC** | **ProductD** |
| **1** | **0** | **0** | **1** | **1** |
| **2** | **0** | **1** | **0** | **1** |
| **3** | **1** | **1** | **0** | **0** |
| **4** | **1** | **1** | **0** | **1** |
| **5** | **1** | **0** | **0** | **0** |
| **6** | **0** | **0** | **1** | **0** |
| **7** | **1** | **0** | **1** | **1** |
| **8** | **1** | **1** | **0** | **0** |
| **9** | **1** | **0** | **0** | **0** |
| **10** | **0** | **0** | **1** | **1** |
| **11** | **0** | **0** | **1** | **1** |
| **12** | **1** | **1** | **0** | **0** |
| **13** | **1** | **0** | **1** | **0** |
| **14** | **0** | **1** | **0** | **0** |
| **15** | **0** | **1** | **0** | **1** |
| **16** | **0** | **0** | **1** | **1** |
| **17** | **1** | **0** | **0** | **0** |
| **18** | **0** | **0** | **0** | **1** |
| **19** | **0** | **1** | **1** | **1** |
| **20** | **0** | **1** | **0** | **1** |

Data Exploration**: 20 customers.  4 products + customer ID.  Binary where 1 = purchased.  0 = no purchase.**

* **9 / 20 (45%) have purchased product A**
* **9 / 20 (45%) purchased product B**
* **8 / 20 (40%) purchased product C**
* **11/ 20 (55%) purchased product D**

**Solution:**

**Output:**

**A screen shot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generatedPlots:**

**A close up of a map

Description automatically generated  
A close up of a map

Description automatically generated**

**A close up of a map

Description automatically generated**

**A close up of a map

Description automatically generated**