CIS 9660 – Midterm Exam
10/26/2020 [30 points]
ver-1.

YOU MUST SUBMIT THIS EXAM THROUGH BLACKBOARD BEFORE 8:15 p.m.. FAILURE TO DO SO WILL RESULT IN A FAILING GRADE. DO NOT WAIT UNTIL LAST MINUTE.

By taking this test you are subject to the academic integrity policies depicted in the Syllabus.

Integrity statement:

Every student must respect the right of all to have an equitable opportunity to learn and honestly demonstrate the quality of their learning. Therefore, all students must adhere to a standard of academic conduct, demonstrating respect for themselves, their fellow students, and the educational mission of the University. **Cheating and plagiarism are serious offenses and will NOT be tolerated in my class.** You will not engage in any type of communication with other individuals during the test. **You will report to your professor any attempts of students contacting you in an attempt to gain an advantage over others. You will not send or post the solutions in any internal or external website during or after the exam.** Any case of cheating will be penalized with a score of 0 and a formal report to the Dean of Students. NO EXCEPTIONS! **By continuing and submitting the test you consent that you will act with academic integrity.**

Download dataset file: Download from Blackboard

For both the conceptual questions and the hands-on questions you will create a Jupyter notebook with your answer for each of the questions in a new cell and labeled accordingly. Once you complete all the questions "offline" you will log to Blackboard to submit your answers and the ******ipynb***** file with your answers online. The file name should have the following format:

Firstname_Lastname-midterm .ipynb [first name and last name as it appears on Cuny First/Blackboard]

The Blackboard test will be open from 6:15 p.m. to 8:15 p.m.

Exam steps:

1. Download csv file (dataset) from Blackboard (Midterm exam folder). Will be available starting at 6:15 p.m.
2. Download midterm pdf file with concept questions and questions for hands-on part.
3. Upload ipynb file to the midterm submission folder with answers to all questions. In python, **label** with a comment (#) the question number.

I won't be answering questions about code not working. However, during the time of the test I will be available via the class Zoom. If you have any question related to the test or submission process send it as a private message to directly to me. Note this is an open space where multiple people so only refer to questions related to the problem statement. I won't answer specific questions or confirm whether an answer you have is correct or not. Do not communicate with any other student during the test time and or after as there might be students that have requested extra time through Baruch's Student Disability Services Office (SDS) and continue working on the exam.

Problem Statement

Telco is a provider of telecommunication services. Like many enterprises, Telco is finding mechanisms to prevent customer churn. For a telecommunication service provider like Telco, the cost of acquiring a new customer is a lot greater than that of retaining a current customer. Thus, the importance of developing an effective customer retention program to maintain the existing clientele.

The first task is to build a classification model to predict whether a customer will cancel the service or not. The second task is to analyze the most important factors that lead to customer churn. Last, provide suggestions to prevent churn.

Data available:

This project will use the Telco Customer Churn dataset which was last updated on February 23rd, 2018. In this data set, each row represents a customer and each column represents customer's attributes. The attributes include four big categories:

- Customers who left within the last month – the column is called Churn.
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Attributes:

customerID: the id of the customer
Gender: whether the customer is a male or female
SeniorCitizen: Whether the customer is a senior citizen or not
Partner: Whether the customer has a partner or not
Dependents: Whether the customer has dependents or not
Tenure: Number of months the customer has stayed with the company
Phone Service: Whether the customer has a phone service or not
Multiple lines: Whether the customer has multiple lines or not
InternetService: Customer's internet service provider (DSL, Fiber Optic, No)
OnlineSecurity: Whether the customer has online security or not
OnlineBackup: Whether the customer has online backup or not
DeviceProtection: Whether the customer has device protection or not
TechSupport: Whether the customer has tech contacted tech support or not
Contract: type of contract (i.e., length)
PaymentMethod: payment method
MonthlyCharge: amount billed
TotalCharges: lifetime amount billed
Churn: whether the customer left

End Goal: Create a classifier that is able to predict whether a customer will churn.

Hands-on questions (Submit ipynb file to Blackboard) [1 point each]

[Try to format the notebook so it is nice and clean –I should be able to run the notebook][One cell per question][Note that answers without the respective code that derives the answer will get 0 points]

1) Import the “telco_train.csv” file into the dataframe *df* and print:

a. the attribute names [make sure you import necessary libraries]

b. Number of rows and columns

2) Are there any **null** (i.e., missing) values in the target variable? How many null values were there? [hint: use the isna() method]

3) Drop NULL (na) values that you found in (3) [hint: use the dropna() method and use the inplace = True parameter]. How many instances for each of the levels of the target variable (i.e., churn) are there left?

4) What is the mean and median value for *tenure?*

5) How are customers distributed (i.e., frequency) across gender?

6) What’s the distribution of the *tenure* of customers? Plot a histogram to answer this question.

7) Make sure the data type for Churn are set as Category. [Hint: use the .astype('category') method if you need to change it]. *Do you get the variable churn as category type?*

8) What is the **ratio of** average of Total Charges for Male Senior Citizens **to** that of Female Senior citizens?

9) Create a box plot with the Monthly Charges based on the type of contract the customers have. What can you conclude comparing the month-to-month compared to the two-year contract?

Decision Tree model (default attributes)

10) Run the following line of code in your notebook (copy and paste it). What does it do?

df = pd.get_dummies(data=df, drop_first = True, columns=['gender', 'InternetService', 'PaymentMethod','Contract'])

11) Create a variable predictors with all the predictors (except for customerID and target variable). Create a variable outcome with the target variable

12) Use the train_test_split function from scikit-learn to split the data into train/test using a 70%/30% split, respectively. Set the parameter *random_state = 1. Make sure you import the necessary libraries from scikit-learn.* Instantiate a Decision Tree model **dt** with the following parameters (max_depth = 4; random_state = 1) and fit the training data. *Make sure you import the necessary libraries from scikit-learn.*

13) What is the *precision* and *recall* on the test data?

14) What is the f-measure on the test data?

15) What is the most important variable? [make sure you show how you determined it]

Logistic Regression model (default attributes)

16) Using the data from (10) and (11), instantiate a Logistic Regression model **lr** with the following parameters (random_state = 1, solver = 'liblinear') and fit the training data.

17) What is the *precision* and *recall* on the test data?

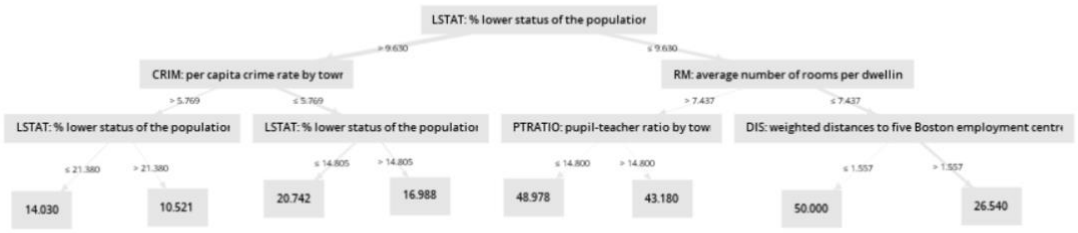
18) What is the f-measure on the test data?

19) Which model (decision tree or logistic regression) performs better?

CONCEPT QUESTIONS [1 point each]

20. **Based on the model below, what would be a possible predicted price for a building with the following characteristics.**

(i) 5 Rooms
(ii) DIS: 2
(iii) LSTAT: 7.3



21. What is the precision of the model with the following confusion matrix: [1 POINT].

		Actual CLASS	
		Class=Yes	Class=No
Predicted CLASS	Class=Yes	50	10
	Class=No	20	120

$$precision = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + FN}$$
$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$
$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

a. 0.85
b. 5/7
c. 5/6
d. 5/8
e. Can't tell based on the data provided

22. In a healthcare context (i.e., diagnosing prostate cancer), which of the following models would be preferable? Why?

Model A

		Actual CLASS	
		Class=Yes	Class=No
Predicted CLASS	Class=Yes	50	20
	Class=No	10	120

Model B

		Actual CLASS	
		Class=Yes	Class=No
Predicted CLASS	Class=Yes	50	10
	Class=No	20	120

23. ARM. Imagine you have the following transactions made by different customers at your POS. Each transaction is represented by the items that were bought on such transaction.

TransactionID	Items
1	Bread, Milk
2	Bread, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Beer
5	Bread, Milk, Diaper, Beer

What is the confidence for the rule {Bread, Milk} => {Beer}?

24. Multiple choice (1 point). Select one. Entropy. We toss 10 times and we get the following results: HHHHHHHHHH. Calculate the entropy based on the outcome: [H: heads][Formula is provided]

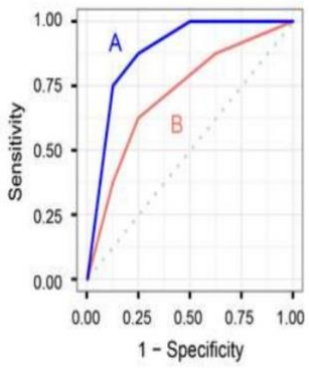
$$entropy\,(A) = -\sum_{k=1}^{m} p_k \log_2(p_k)$$

a. 1/2
b. 1
c. 0
d. 1/4
e. We can't calculate the entropy with the information provided

25. **Which of the following is not true about the 'max_depth' parameter building a decision tree.**

a. The performance of the model in unseen data declines if we keep increasing the max_depth parameter
b. The misclassification error in the training set decreases as we increase the max_depth of the tree
c. The misclassification error in the testing set decreases as we increase the max_depth of the tree
d. The decision tree will yield a larger set of rules

26. **Given the ROC curves for the following two models: Model A and Model B. Which model is preferable? Why?**



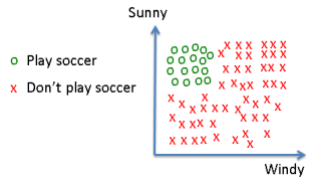
27. You run an Association Rule Mining analysis and get the following output:

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift ↑
15	Concealer, Blush	Foundation	0.115	0.523	0.914	-0.325	-0.003	0.975
20	Blush	Foundation	0.192	0.529	0.875	-0.534	-0.003	0.987
24	Mascara	Foundation	0.192	0.538	0.878	-0.522	0.001	1.003
3	Bronzer	Lip Gloss	0.141	0.505	0.892	-0.417	0.004	1.031
27	Eye shadow	Foundation	0.211	0.554	0.877	-0.551	0.007	1.033

Which of the following rules show a substitution effect?

a. Rule No. 15
b. Rule No. 15 & No. 20
c. Rule Nos. 3, 24, 27
d. All the rules show a substitution effect

28. Given the following training set, you want to build a binary classifier to predict the *Play soccer*, what is the most important variable (i.e., has higher information gain)?



1. Sunny
2. Windy
3. The variables *Sunny* and *Windy* are equally important
4. Can't tell based on the data

29. **Which of the following is not true about the 'max_depth' parameter building a decision tree.**

a. The performance of the model in unseen data declines if we keep increasing the max_depth parameter
b. The misclassification error in the training set decreases as we increase the max_depth of the tree
c. The misclassification error in the testing set decreases as we increase the max_depth of the tree
d. The decision tree will yield a larger set of rules

30. Creating a dataset of emails that are labeled as Spam or Not Spam and using text-mining techniques to create a predictive model that ranks new emails as spam or not spam is an example of:

a. Unsupervised learning
b. Supervised learning - regression
c. Supervised learning – classification
d. Natural Language Processing