DSBA/MBAD 6211 Assignment 2: Advanced Predictive Models

Instructions: This is an individual assignment. The submitted solution and answers should be your own. The data file for this homework is **BankChurn.csv**, which is to be downloaded from Canvas. You are asked to use Python to build Random Forest and SVM models and answer the given questions. Create a new Word document and save it as Predictive_xxxx (where xxxx is your ninernet login name). Write your full name on the first page of the Word document. Where required, write your answers or paste screenshots in this Word document. You need to submit both the Word document and Python Code file. **Your Python code should run correctly for your assignment to be graded. Code that generates error will result in loss of points (up to a maximum of 20%)**

Variables and models naming requirements:

- Include your ***name initials*** to the data frame names as well as model names in your Python coding. This is required for your work to be graded.
- For instance, my initials are **CS**, and in my coding, I would name the data frames as ***dfCS, dfCS.train***, and ***dfCS.test.*** I would also name the models as ***RFCS, SVMCS***, etc.

Problem description and questions: We will analyze a dataset **BankChurn.csv** which had data of a bank's customers and their churn. The descriptions of the columns are given below. Run both Random Forest and Support Vector Machines models on the data to predict whether a customer will churn or not (i.e., the target variable is "churn") and answer the questions that follow on the next page.

- customer_id: Unique identifier for the customer.
- credit_score: Customer's credit score.
- country: The country of the customer (e.g., France, Spain, Germany).
- gender: Customer's gender.
- age: Customer's age.
- tenure: Number of years the customer has been with the bank.
- balance: Customer's account balance.
- products_number: Number of bank products the customer is using.
- credit_card: Indicates whether the customer has a credit card with the bank (1 = Yes, 0 = No).
- active_member: Indicates whether the customer is an active member (1 = Yes, 0 = No).
- estimated_salary: Customer's estimated salary.
- churn: Indicates whether the customer has left the bank (1 = Yes, 0 = No).

Questions & Problem Tasks

1. Are there any variables which cannot be used in your model? Explain why? (2 pts)

2. What variables need to be dummy coded before you run your logistic regression model? Explain what new dummy coded columns you created. (4 pts)

3. Do you have to consider missing values in your dataset? How did you handle the presence of missing values, if any? (4 pts)

4. First run the decision tree classifier using all independent variables in the dataset, after dropping the ones identified in question 1. You can use any of the criteria of gini index or information gain. What is the accuracy of this classifier? Show the confusion matrix with proper labels. (5 pts)

5. **Running Random Forest Classifier: (15 pts)**
   a. Run a Random Forest Classifier. Choose the appropriate set of hyperparameters. What is the accuracy of this classifier? Show the confusion matrix with proper labels. Does this classifier perform better than the decision tree in question 4?
   b. Re-run the Random Forest Classifier with a different set of parameters. What is the accuracy of this classifier? Show the confusion matrix with proper labels. Does this classifier perform better than the first two you ran?
   c. Fine tune the Random Forest model.
      i. What are the optimal parameter settings for the Random Forest Classifier?
      ii. List the variables in the decreasing order of their importance. Do you agree with the rankings, based on your knowledge of why customers quit their banks?
      iii. Show the accuracy and confusion matrix for this classifier.

6. **Running SVM Classifier: (15 pts)**
   a. Run an SVM Classifier, first with a linear kernel and then with a radial kernel. Give your own values for the other hyperparameters but keep them the same across the two runs of SVM. Which kernel gives a better performance for this dataset? Explain.
   b. Fine tune the SVM model.
      i. What are the optimal parameter settings for the SVM classifier?
      ii. Show the accuracy and confusion matrix for selected classifier by the fine tuning.

7. Overall, which model (random forest or SVM) perform better for this dataset? Explain. (5 pts)

Submit your Word document and the python code through Canvas.