Data mining challenge

Worth 30% of final grade in CS405/505 Data Mining Fall 2019.
Prof. Russell Butler, Johnson 114A, Office hours MWF, 9:00-11:00 am.
Due date: December 13th 2019, 11:59pm

Predictive modeling

Your goal is to build a model that predicts a **probability** that a given customer will default on a loan
A customer profile consists of a list of bank transactions preceding the loan request. Each transaction is either a debit (money going out of account, negative values) or credit (money coming into account, positive values).

Customer profile (attributes) (15000 customers total):

Id – id of each customer
dates – dates of each transaction
transaction_amount – numpy array of credits and debits, length varies across different customers
(your predictions will be primarily based on information in this array)
days_before_request – days before loan request for each transaction
loan_amount – amount loaned to customer by bank
loan_date – date of loan
outcome:
isDefault – did the customer pay back (isDefault=0) or not pay back (isDefault=1)?

isDefault is given for the first 10000 customers. Your job is to assign a probability to isDefault for the remaining 5000 customers.
Train your model on the training data (instances 0 - 9999) and make predictions on the test data (instances 10000-14,999). The test data is the same format as training data, except it does not contain the isDefault column.
The data is available at the following link:
https://drive.google.com/file/d/1oPSNCYeCVGJsTX60X-PW088R8S0AMmeT/view?usp=sharing
The data can be loaded from dataset.pkl in python using:

```
import pandas as pd
data = pd.read_pickle('path/to/data/dataset.pkl')
```

Submission:

A) Your submission should be a CSV file with your name firstname_lastname, containing:
5,000 rows corresponding to instances 10000-14999 from the dataset
Each row has two columns:
column 1 – id of customer
column 2 – probability that isDefault==1 (probability the customer does not pay back the loan)

B) Create a small (1 page) presentation that you would hypothetically give to a salesperson of the company when presenting your algorithm, so that the salesperson could understand your algorithm and explain how it works to a potential customer.

In the 1 page, answer the following:
Which algorithm did you use and why?
How certain are you of the results?
Is there a subset of potential customers that are very safe to lend to? A subset that is very dangerous?
What features of the dataset best predict isDefault?
Do these features make sense intuitively? Justify your use of the features
Submit this as a separate pdf, along with your CSV of predictions.

Evaluation:

(75%) – A) The quality of your predictions will be assessed using the ROC area under the curve
(25%) – B) The clarity and correctness of your interpretation of the model