Homework 6

Your Name: \_Vikas Sanil\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student ID:\_\_A20511552\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**(100 points) Python practice for classification**

**Use the best setting of decision trees, from HW4**

**1). Run different ensemble methods with this best tree model**

**2). Use oversampling to deal with imbalance issues, though the imbalance issue may not be serious in this data. Run decision tree again, and observe the results**

**For dataset Loans\_20K after oversampling the decision tree accuracy decreased but precision and recall score increased pretty significantly.**

**By 10-fold Cross Validation: Original Decision tree accuracy = 0.71745 , precision = 0.358725 , recall = 0.5**

**By 10-fold Cross Validation: Oversampling Decision tree accuracy = 0.7022789448180295 , precision = 0.7515710463933435 , recall = 0.7022794843011192**

**For dataset Loans\_200K after oversampling the decision tree classifier accuracy and precision decreased and the recall score remained the same.**

**By 10-fold Cross Validation: Original Decision tree accuracy = 0.61126 , precison = 0.30563 , recall = 0.5**

**By 10-fold Cross Validation: Oversampling Decision tree accuracy = 0.49788118261097203 , precision = 0.24894059130548601 , recall = 0.5**

**3). Use feature selection and reduction, and re-run decision tree models, and observe whether the results can be improved.**

**For dataset Loans\_20K after performing feature selection and reduction the decision tree classifier result didn't vary compared to results without it.**

**By 10-fold Cross Validation: Original Decision tree accuracy = 0.71745 , precison = 0.358725 , recall = 0.5**

**By 10-fold Cross Validation: Oversampling Decision tree accuracy = 0.7022789448180295 , precison = 0.7515710463933435 , recall = 0.7022794843011192**

**For dataset Loans\_200K after performing feature selection and reduction the decision tree classifier accuracy and precision increased for the dataset without oversampling. The rest of the parameters results didn't change compared to results without feature selection and reduction.**

**By 10-fold Cross Validation: Original Decision tree accuracy = 0.61119 , precison = 0.305595 , recall = 0.5**

**By 10-fold Cross Validation: Oversampling Decision tree accuracy = 0.49788118261097203 , precison = 0.24894059130548601 , recall = 0.5**

**Note, for part 1) and 2), you do not need to change parameters for decision trees, just used the best setting from HW4.**

**For part 3), you need to tune up parameters to decision trees again**

* Use Loans\_20K.csv data by using 10-fold cross validation
* Use Loans\_200K.csv data by using 75% as training, 25% as testing
* Loan term as label

Note:

* You need to change different/multiple parameters to find the best model.
* You can find data sets from “slide & data” on blackboard system

Submission

* The ipynb and saved html files

A comparison of different parameters and metrics (accuracy, F1)