Homework 1

*Vinoothna Peruri*

*June 04 ,* *2020*

Class: ALY6020 - Predictive Analytics

\*\*Professor:\*\* Marco Montes de Oca

OBJECTIVE:

Tackle the “Yoga problem” using logistic regression.

Investigate the effects of different thresholding values on the accuracy of the model at different class imbalance ratios.

BACKGROUND:

* Logistic regression returns a probability.
* In order to map a logistic regression value to a binary category, you must define a **classification threshold** (also called the **decision threshold**).
* It is tempting to assume that the classification threshold should always be 0.5, but thresholds are problem-dependent, and are therefore values that you must tune.
* Due to different class imbalance measures like accuracy, precision, recall breaks down.
* A **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. AUC provides an aggregate measure of performance across all possible classification thresholds
* AUC is desirable for the following two reasons:

AUC is scale-invariant. It measures how well predictions are ranked, rather than their absolute values.

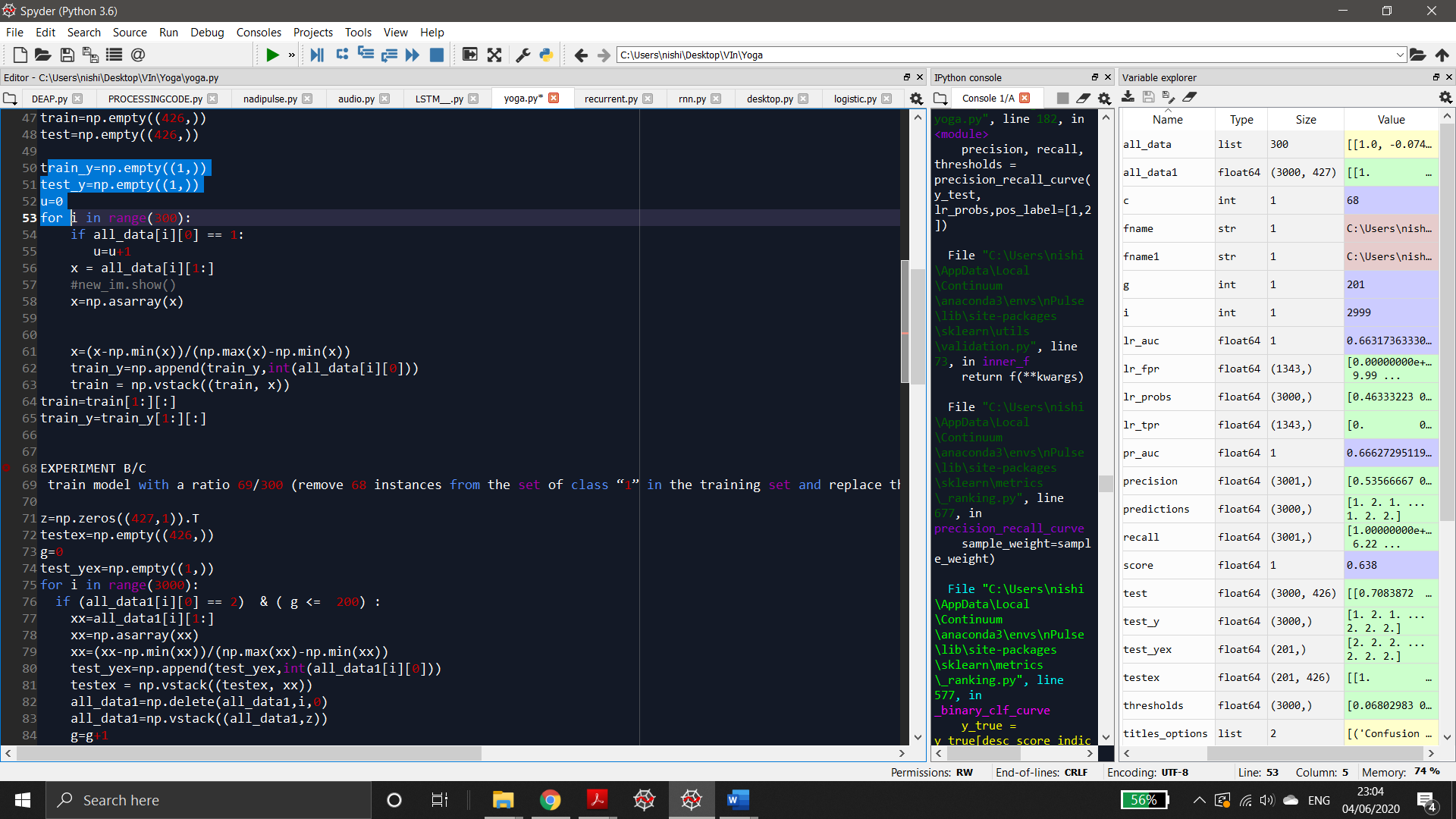
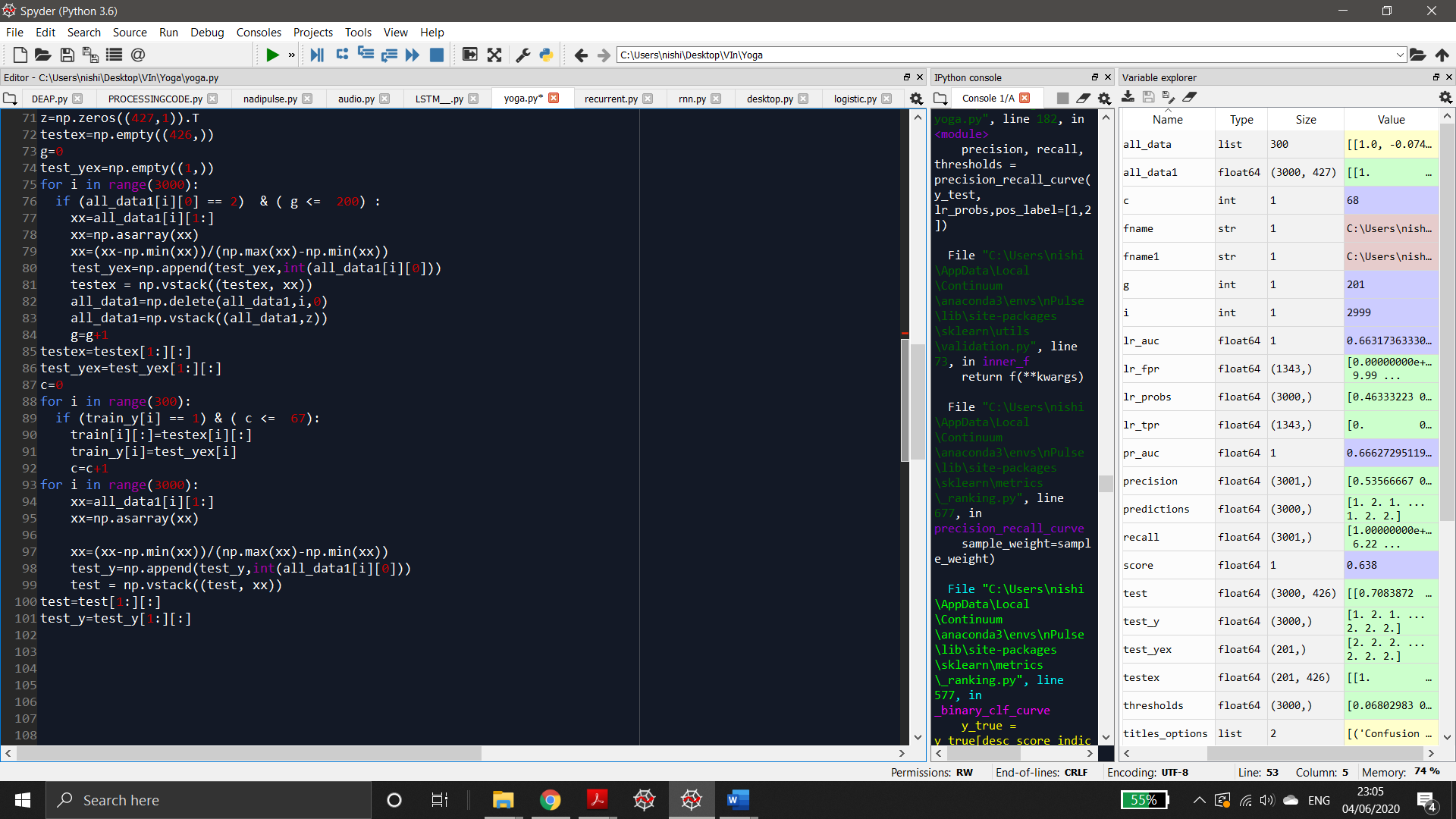
AUC is classification-threshold-invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

We classify images with three different class imbalances and obtain the ROC, Precision vs recall curves and AUC – ROC for each of them.

PROCEDURE:

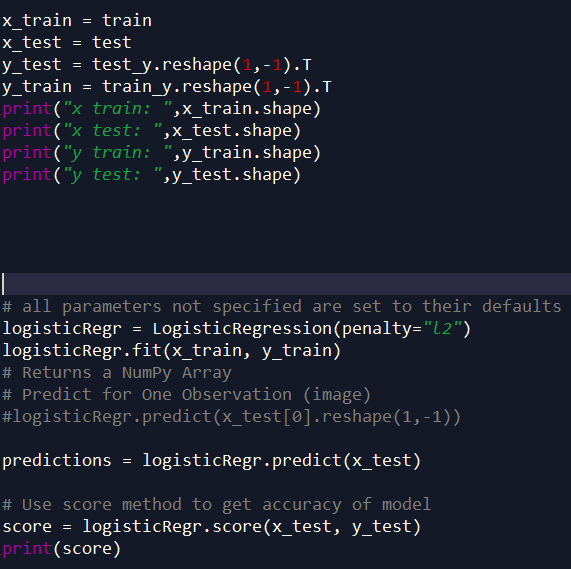
1. Preparing the test train data

The three experiments have different training datasets with an increasing class imbalance. For this data of class ‘2’ from the test data was swapped with n number of datapoints of class ‘1’ form the training data to increase class imbalance.

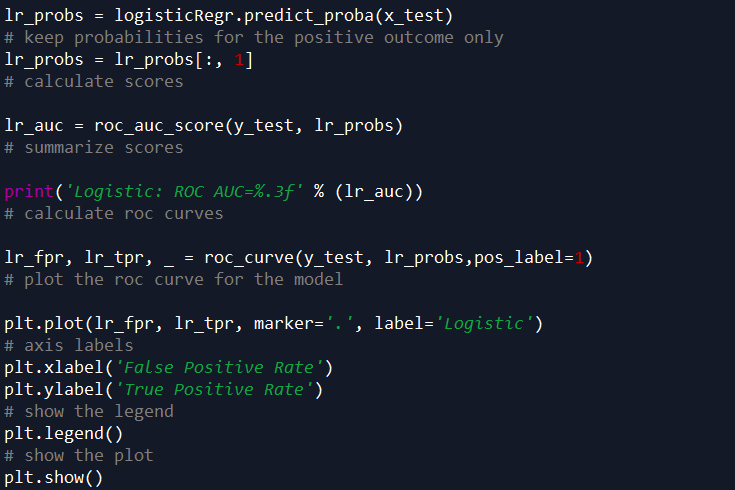
1. Training/ Prediction

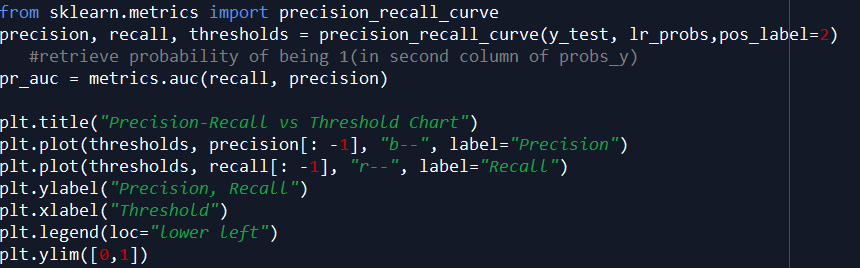
We used classifiers Library Sklearn Logistic regression classifiers to train the dataset and below is the code.



1. Plotting

Below is the code used to plot the charts.





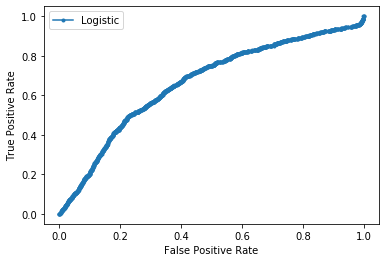
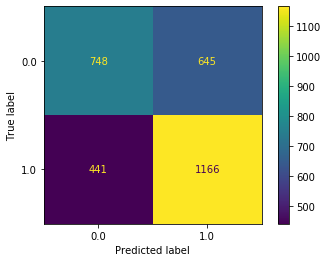
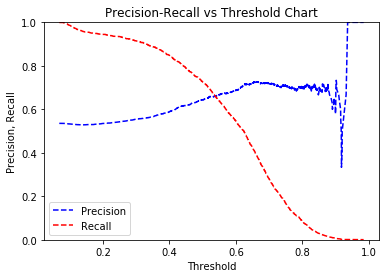
OBSERVATIONS & RESULTS

The original class ratio in the training set is 137/300 (number of class “1” instances over total number of instances). Try three experiments with logistic regression:

* Experiment A: train model with the original set

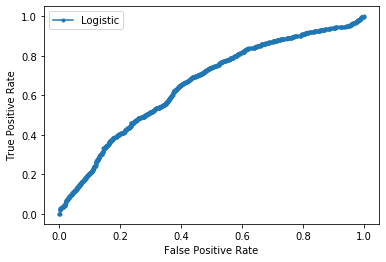
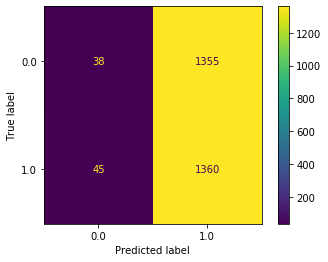
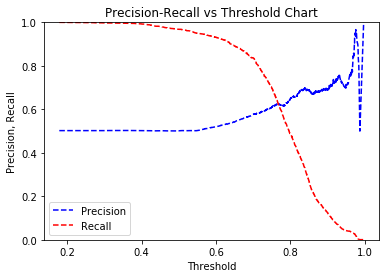
Accuracy =0.6353333333333333

ROC AUC=0.663

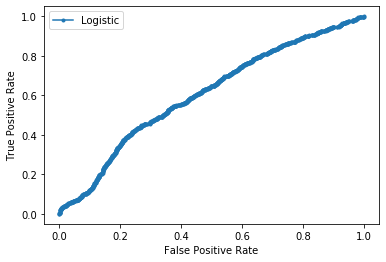
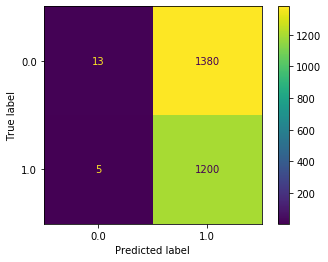
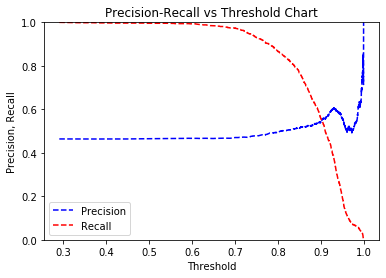
* Experiment B: train model with a ratio 69/300

Accuracy = 0.4996426018584703

Logistic: ROC AUC=0.657   

* Experiment C: train model with a ratio 34/300 Accuracy : 0.46689761354888376

Logistic: ROC AUC=0.606

ANALYSIS

* The accuracy decreases with the increase in class imbalance.
* We also see fluctuations in the precision vs recall curve with respect to different thresholds in each of the three experiments. Making it difficult to rate the model’s performance or tune the hyperparameters.
* This brings us to the use of ROC-AUC to define the aggregate measure of the model. Since AUC is **classification-threshold-invariant**. It measures the quality of the model's predictions irrespective of what classification threshold is chosen. We observe a close to constant ROC-AUC of around 0.6 for all the three cases. In our case this would mean that the probability of the model classifying the gender correctly is 0.6, confidently.

REFERENCES

[1]