**FDA Homework 3-2**

**Understanding the data**

The following tables will briefly describe each column representation.

[Link to table](https://www.semanticscholar.org/paper/Real-time-prediction-of-online-shoppers%E2%80%99-purchasing-Sakar-Polat/747e098f85ca2d20afd6313b11242c0c427e6fb3/figure/2)

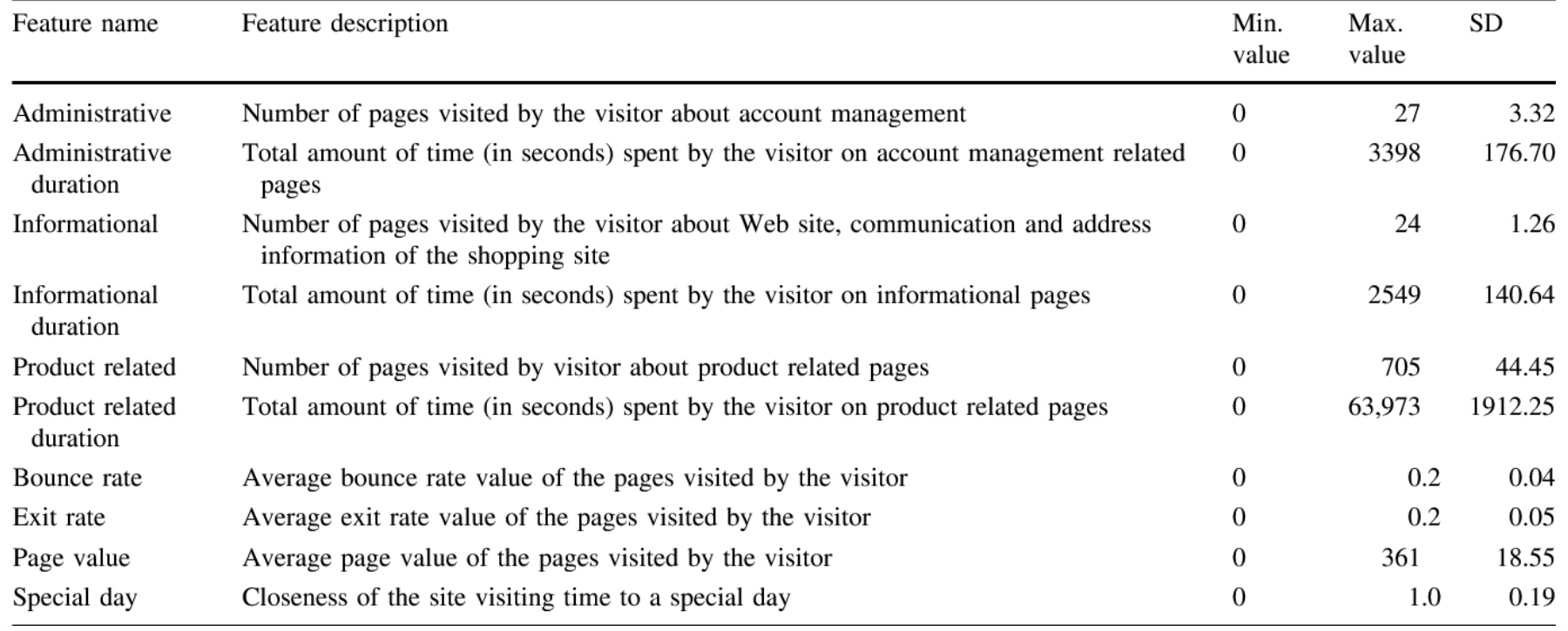


Table 1. feature description

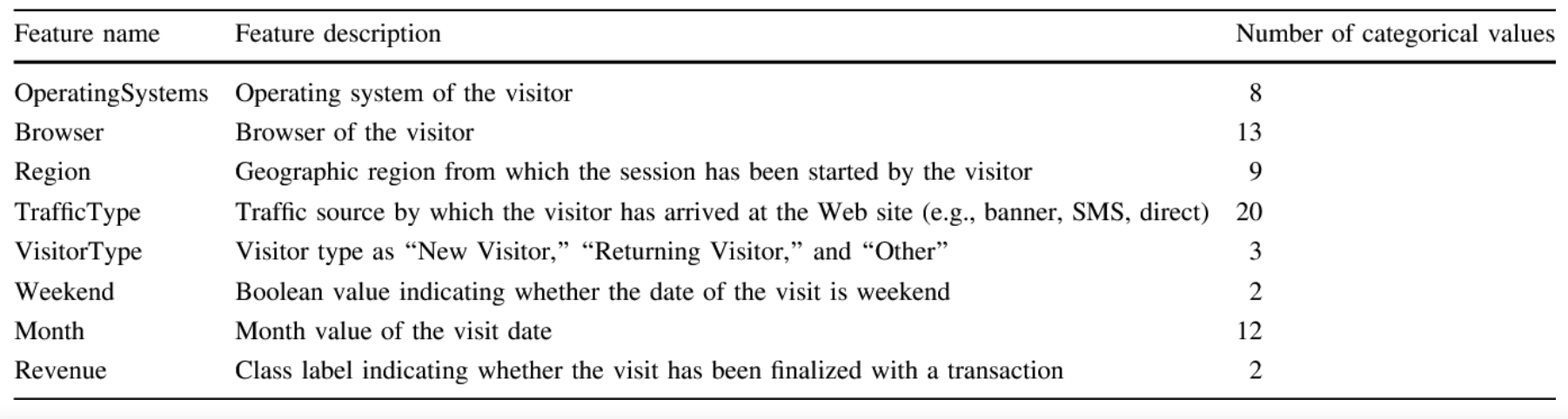


Table 2. feature description

**Data Visualization**

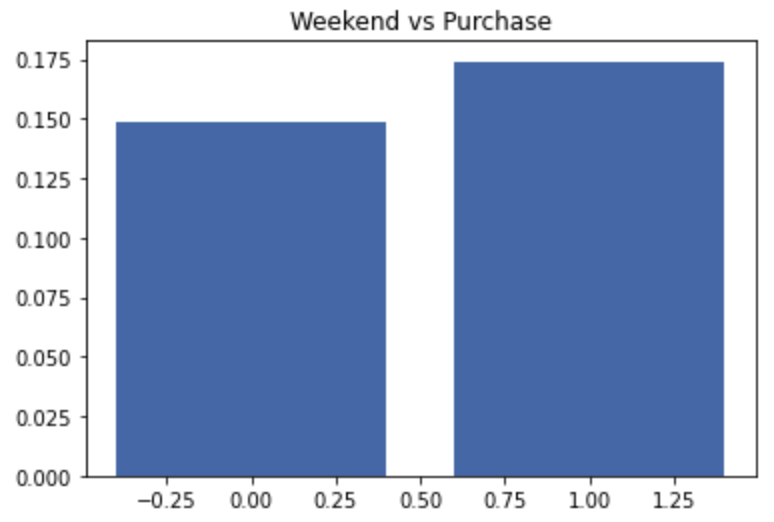
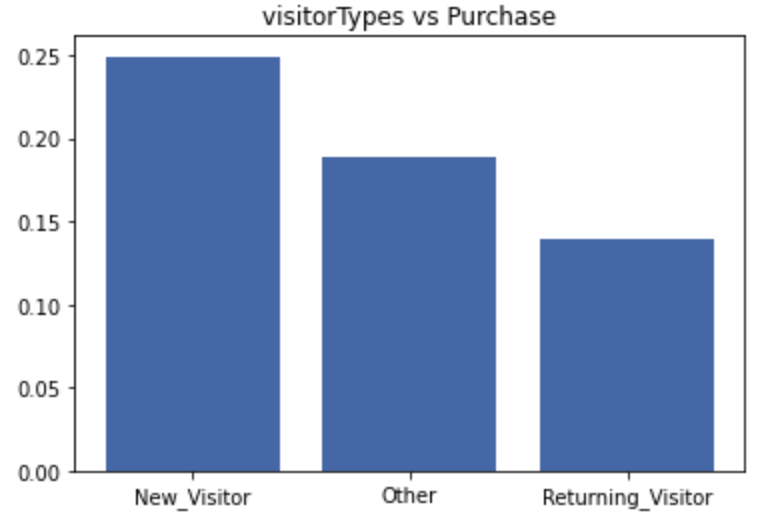
 

Fig1. **Image on the left** indicates the percentage of people that purchased during weekend vs those that didn’t purchased during weekend (we can see there is no too much difference, **image on the right** indicates percentage of each type of visitor to make a transaction.

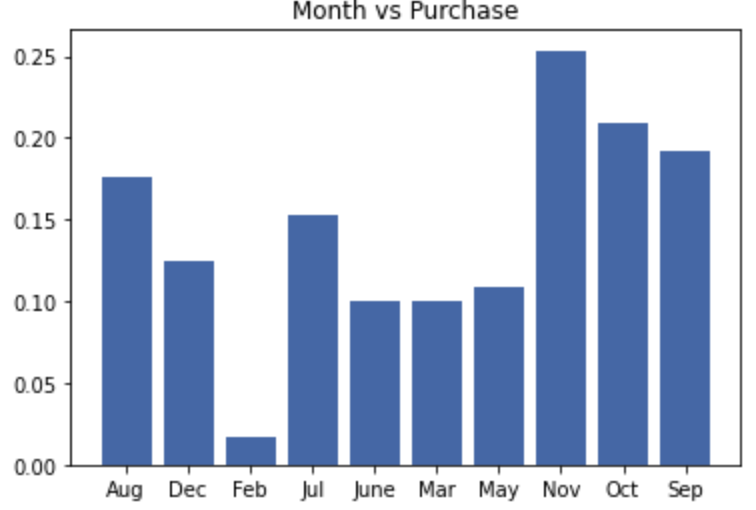
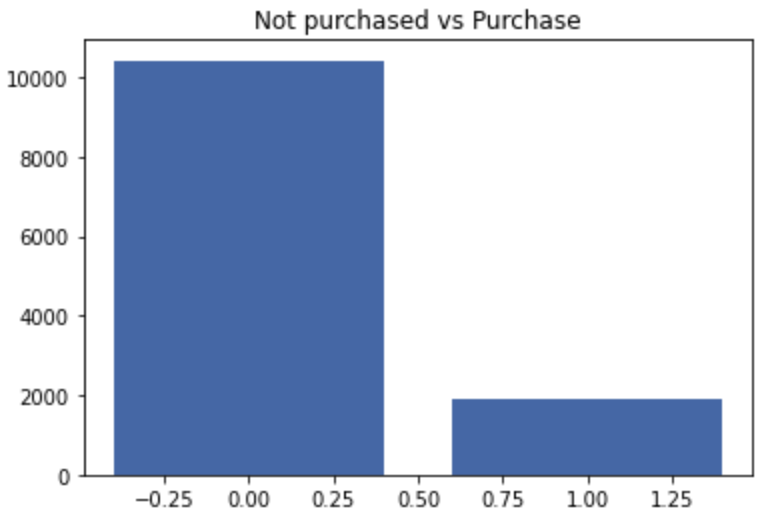
 

Fig 2. Image on the left indicates percentage of purchased made in each month,

Image on the right indicates connection that made transaction vs no transaction

We can notice a very imbalanced dataset.

Preprocessing:

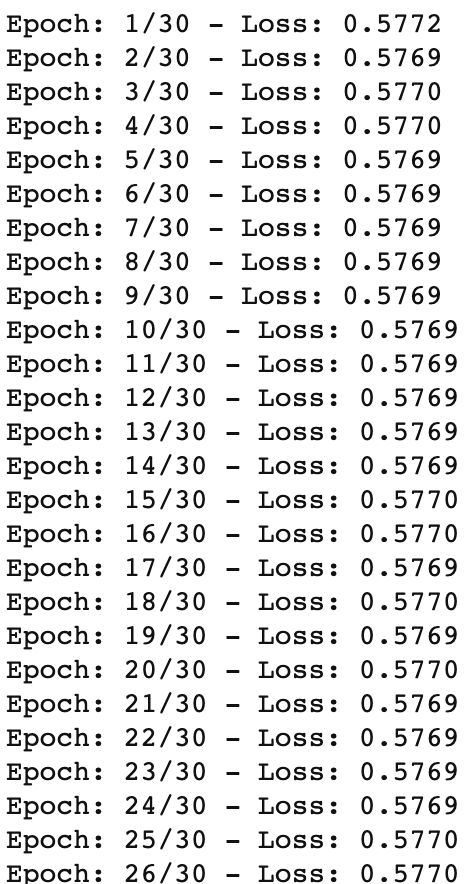
* Categorical encoding(Months, VisitorType, Revenue)
* Drop irrelevant features   
  

Training descriptions:

* Model: Multi-Layer perceptron
* Optimizer: Adam
* Loss function: BCEwithlogits

**Result 1**

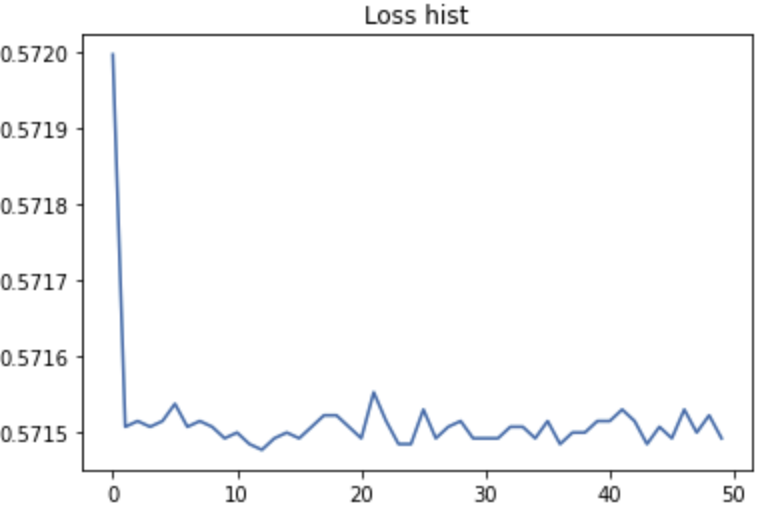
* Validation Accuracy 80%
* First train loop the loss doesn’t seem to update, model is barely trained



**Result 2**

Main changes:

* Get rid of outliers (get rid of > 0. quantile)
* Normalize data (0,1)

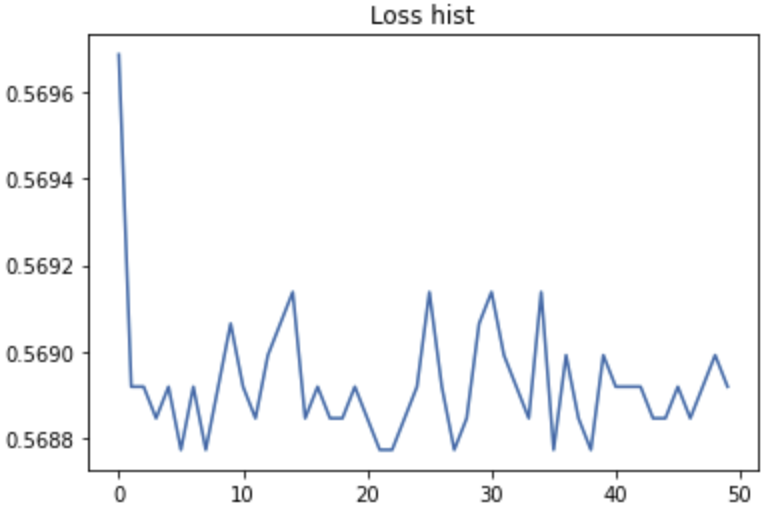


Validation Accuracy: 81%

**Result 3**

Main changes:

* Get rid of outliers (get rid of > 2.5\*75%quantile)
* Add weight initialization

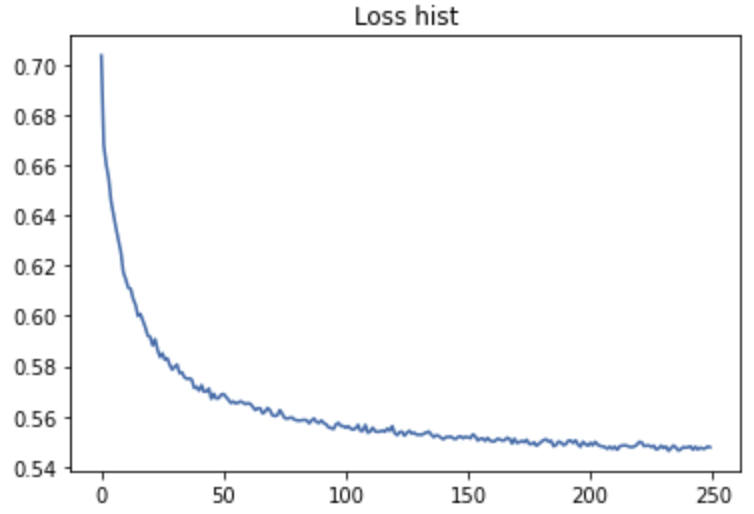


Validation Accuracy: 82%

**Result 4**

Main changes:

* Oversampling to deal with data imbalance
* Learning rate reduce overtime



Validation Accuracy: 85%