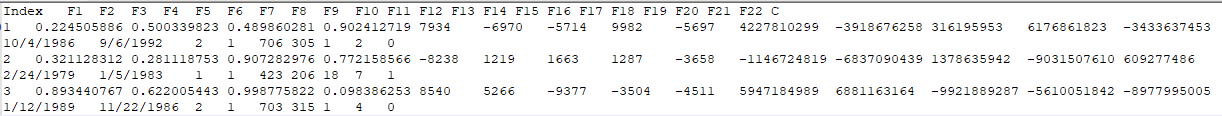
**Affinity Analysis process overview**

**Shivendra Pratap Singh**

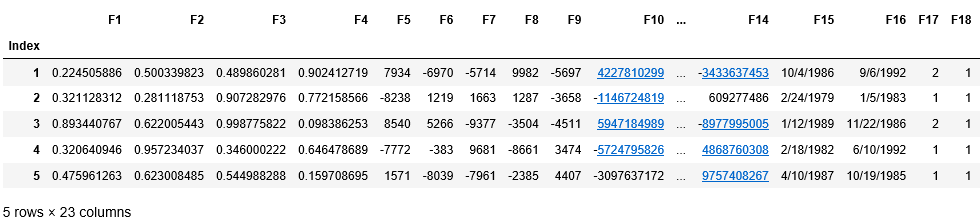
[**Shivendra007pratap@gmail.com**](mailto:Shivendra007pratap@gmail.com) **(8874250931)**

**Problem Statement:** Given an e-commerce dataset where each record is a combination of some user 'u' & some item 'i' & the task is to predict if user 'u' will buy item 'i' or not.

**Data Provided:** data provided was a cleaned data with unknown attributes. Data Contains date columns as well and some of the features were already scaled.

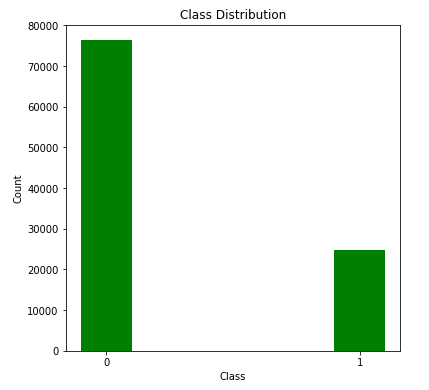
****

**Data Pre-processing:** To Explorer and Analyse data, converted it into pandas DataFrameand after analysing the columns converted them into proper data type ex. Integer, float, datetime.

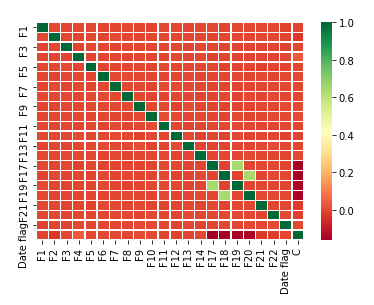
****

**Exploratory Data Analysis**

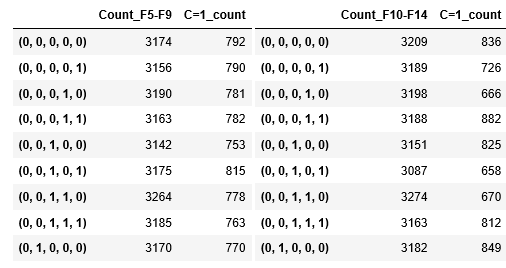
**Imbalanced Class problem:** Firstly decided to find the distribution of class labels and found around 76% labels were ‘0’ and only 24% records were labelled as ‘1’. A Model can achieve 76% accuracy ny just predicting ‘0’ for all records. To handle imbalance class problem we can do Up-sampling of minority classes and Down-sampling of Majority classes using sklearn’s resample module. But I decided to choose confusion matrix as the performance measure instead of accuracy.



**Correlation and Multi-collinearity Check:** Checked the correlation among numerical attributes and found attribute F-19 and F-20 were correlated with F-17 and F-18 respectively with +ve value of .68. In VIF test for multi-collinearity, attributes F19 and F20 were highly multi-correlated, so decided to drop them from dataset.



**Class Distribution Check For numerical columns:** Tried to find a distinguishable pattern in numerical columns to check whether Binnig them would help the model in classification but got to know that distribution was closely same across all the possible Bins and it was disappointing.

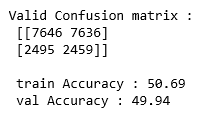
****

Note: [here (0,0,0,0,0) represents that all F5-F9 values are –ve and 3175 shows count of such records. 792 shows count of C=1 across such records]

**Feature Creation:** An additional feature was created using Date columns provided considering one as product expiry date and another as product purchase date. Initially This Feature was categorical (1 if D.o. Expiry > D.o . Purchase otherwise 0).

**Model Iteration 1:**

Split the data into 80-20% ration of train and validation split and Ran down the very first iteration of Logistic Regression. Result wasn’t very satisfactory. Confusion matrix was poor.



Observation:

1. Coefficients for most of high value numerical columns were zero. They were not participating in any prediction.
2. Date flag created wasn’t working in its categorical form.
3. Model bias was too high.

Action Taken:

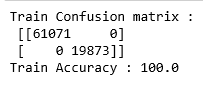
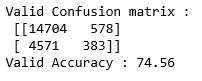
1. Used high Regularisation for the coefficients of logit model.
2. Standardized the high value columns.
3. Instead of Categorical Flag, created feature was directly taken as difference between expiry date and purchase date.
4. Reduced the no. of features using PCA.(Principle component Analysis)

Note: PCA was done on group of five columns (F5-F9 and F10-F14) having same magnitude of numerical values.

**Model Iteration 2:** In iteration 2, I decided to run 2 different classifiers on dataset. i. Logistic regression and ii. Random Forest.

This time Logistic Regression model went rouge and predicted all classes to 0 and achieved the accuracy of 76%.

Random Forest was seems to perform well on train and test dataset as well. It got 100% of accuracy on train set and around 75% on validation set.

Observation:

1. Logistic Regression model was not working with our choices.
2. Random Forest model was overfit.

Action taken:

1. Dropped Logit model.
2. Fine-tuned Hyper-Parameters(no. of estimators) of Random forest model and there was a slight improvement in performance. Yes, it was overfit but considering about our imbalanced dataset it performed well.

**Final Prediction:** Final prediction on test data was made using Random Forest Classifier.