VERSION 1.0

NOVEMBER 16, 2017

BOI CASE STUDY V.3 ANALYSIS REPORT

PRESENTED BY: MINA YOUSSEF

PURPOSE

This report presents the finding for Case Study V3 presented by candidate Mina Youssef

TOOLCHAIN USED

Python, Jupyter (interactive environment), NumPy and Pandas

scikit-learn (for Machine Learning)

R/Rattle (for data exploration)

Assuming all independent packages are properly installed, please navigate to:

On Windows

Run cmd

Navigate to: $> ~\boi-case-study\solution\

Run $> “Jupyter notebook”

MANAUL INSPECTION

Upon receiving the dataset, I carry out a manual inspection and two files are swapped between for training and testing purpose: (TEST - Transactions out of Current Account.csv / Model Build - Transactions out of Current Account.csv)

Also, minor manual editing in the CSV headers for some files so it can be loaded by Python Pandas stream reader.

WORKFLOW

The following Summaries the workflow carried out:

STEP 1: DATA PREPARATION

STEP 2: DATA INTEGRITY VALIDATION/OUTLIER ANAYSIS

STEP 3: DATA EXPLORATION

STEP 4: BUSINESS INTELLIGENCE QUESTIONARE

STEP 5: TRANSACTIONAL NARRATIVE FEATURE EXTRATION

STEP 6: PREDICTIVE MODEL(S) INDUCTION

STEP 7: PREDICTIVE MODEL TESTING

STEP 8: MODEL(S) OPTIMIZATION

STEP 9: POST-DATA EXPLORATION

STEP 10: CONCLUSION AND RECOMMENDATION

STEP 1: DATA PREPARATION

The objective of this step is to successfully load and aggregate all the dataset in single table AbstractBaseTable (ABT) and make sure no duplicate client information (i.e. horizontal data verification)

Jupyter Notebook:

0_data_preparation (for Model Build)

0_data_preparation_test (for Test Sample)

1. We start by loading each of the *.csv file into separate dataframe
2. For each dataframe, we check for any duplicate customer entries
3. A total of 5 customers duplicates are found across all *.csv
   `ClientIDs is be excluded from data set: {1220, 46, 21, 3674, 3675}`

4. Since we do not know the ETL process that has been used to populate the data, the safest action is to remove those duplicated records especially they constitute 0.05% of the whole dataset.
5. Finally, we merge all the loaded dataframes into single dataframe, and persist the clean out data under directory ~/clean/
6. We repeat the whole process for the test sample.

Finally, the prepared dataset is saved under

".../specs/clean/Model Build - AbastractBaseTable.csv"

STEP 2: DATA INTEGRITY VALIDATION/OUTLIER ANAYSIS

The objective of this step is to investigate the data integrity per-attribute (i.e. vertical data verification) . Also, report and act on outliers found within each.

Jupyter Notebook:

1_data_integrity_validation

1_data_integrity_validation_test

One main technique to use for data integrity validation is to plot the frequency count of the attribute data series and observe its distribution

Some irregularity found:

- Some clients age are 100 and 200
- Gender Attribute contains is not uniformed as 1/0 as in the specs
- County Attribute contains Dublin areas, cities, towns even other countries

All the invalid/irregular data has been adjusted, for detail analysis please refer to Jupyter Notebook

Finally the validated dataset is saved under

".../specs/clean/Model Build - AbastractBaseTable - Validated.csv"

STEP 3: DATA EXPLORATION

STEP 4: BUSINESS INTELLIGENCE QUESTIONARE

In this step we are answering the questionnaire of Section I. Business Intelligence. A segmentation based on gender is carried out, while the total number is obtained by the confusion matrix, which can be used in marketing campaign targeting.

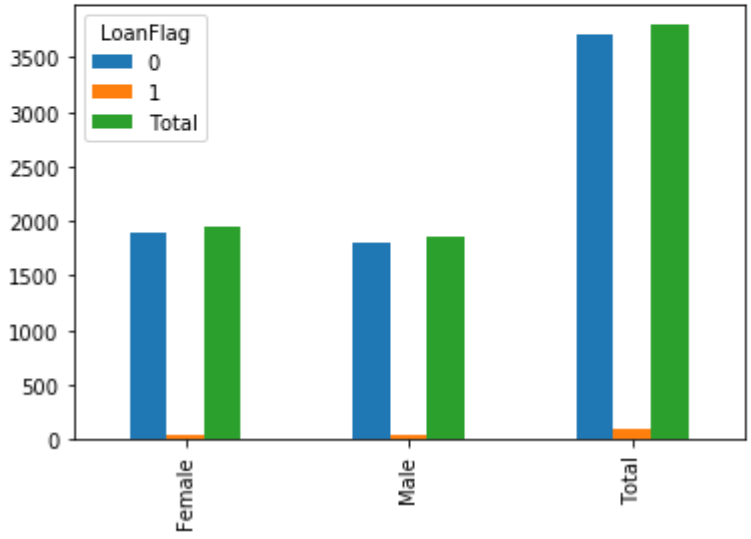
Jupyter Notebook:

3_business_intelligence

How Many Customers above 50 years old have taken up a loan?

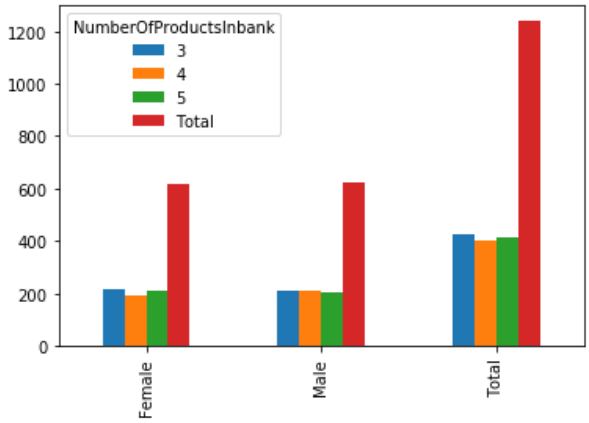
LoanFlag	0	1	Total
Female	1894	48	1942
Male	1810	48	1858
Total	3704	96	3800

```
Total number of customer above 50 yr took a loan is 96
```



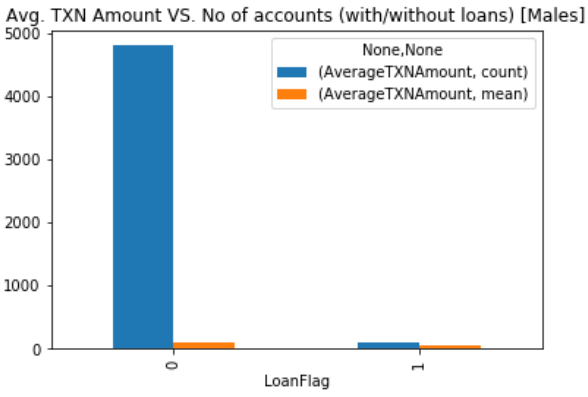
How Many Females aged 30 to 40 have more than 2 products?

NumberOfProductsInbank	3	4	5	Total
Female	215	193	209	617
Male	209	210	203	622
Total	424	403	412	1239



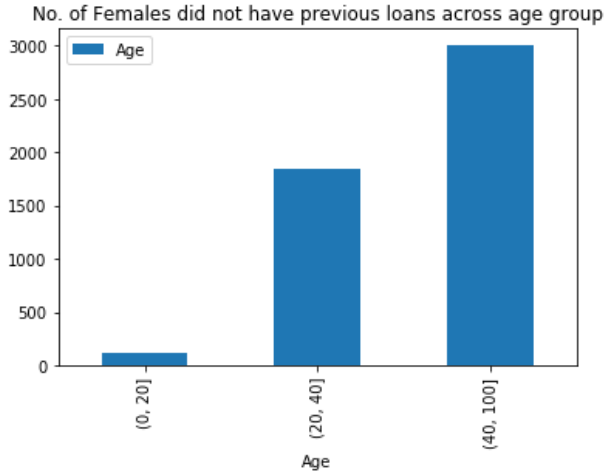
What is the average number of Current Account(CA) Transactions for males who had a previous Loans

	AverageTXNAmount	
	count	**mean**
LoanFlag		
0	4797	92.606891
1	100	62.631880



How many females did not have a previous loans and who are aged (Less than 20, 21 to 40, 40+)

	Age
Age	
(0, 20]	115
(20, 40]	1842
(40, 100]	3009



STEP 5: TRANSACTIONAL NARRATIVE FEATURE EXTRATION

Jupyter Notebook:

4_TXN_features_extraction

The objective of this step is to extract payment pattern and activities for each of the clients. We accomplish that using 3 set of features:

1. Frequency of a given payment (Rank)
2. Activity associated with a given transaction (i.e. if paying in Casino in Las Vagas then we flag a Gambler, 'PayPay' => Internet shopper and so on)
3. Country of transaction, to flag if transaction is domestic or international

First, For the transaction Rank, we first hash the narrative and compute frequency table of each of the hash code and the number of occurrence of the narrative for example

TXN Narrative "LUXOR HOTEL/CASINO LAS VEGAS NV"

hash code " 438e7d777b4277f173f5e4649bc3fb29"

and occurred 17 times throughout the dataset

Therefore for every occurrence of this narrative for any client he/she will have a rank attribute with value 17 and so oone

	ClientID	LastTransactionNarrative	Rank
0	1	nan	NaN
1	2	THE BRIDGE LAUNDRY WICKLOW TOWN	2.0
2	3	LUXOR HOTEL/CASINO LAS VEGAS NV	9.0
3	4	HARVEY NORMAN CARRICKMINES	1.0
4	5	PAYPAL *PETEWOODWAR 35314369001	1.0
5	6	METROPARK HTL KLN HK10900HONG KONG	1.0
6	7	AIR FRANCE E AIRFRANCE.FR	2.0

Second, we associate top activities for each client. We collect first top performed payment activities and associate a behavior for each activity

```
dic_features = {
    # Feature => Keywords
    'gamber': ['CASINO', 'VEGAS', 'LAS', 'HOTEL/CASINO'],
    'luxurious': ['INN', 'SUITES', 'PLAZA', 'HILTON', 'ROYAL', 'HYATT', 'MARRIOTT', 'FAIRMONT', 'RESORT-WDW', 'RESORT/CASINLAS'
    'golfer': ['GOLF'],
    'traveler': ['AIR', 'AIRWAY', 'Limitersd-travel.ie', 'Airport', 'EASYJET.COM', 'RYANAIR'],
    'gamer': ['PLAYSTATIONNETWORK'],
    'shopper': ['STORES', 'AMAZON.CO.UK', 'Amazon'],
    'cinephilia': ['NETFLIX.COM', 'MOVIE', 'MOVIES-AT.IE', 'MOVIEPLEX', 'MOVIES', 'ITUNES.COM/BILL'],
    'car_renter' : ['RENT', 'RENT-A-CAR']
}
```

And then construct a flags vector of each of the client record based on the txn narrative he/she got.

For example, the following are subset of client who golf

	ClientID	gamber	luxurious	golfer	traveler	gamer	shopper	cinephilia	car_renter
492	503.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
2161	2175.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
3407	3421.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
3652	3666.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
3660	3676.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
4219	4236.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
4820	4837.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
5025	5042.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
5123	5140.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0

NOTE: IN PRODUCTION SYSTEM, EACH OF THE CLIENT WILL HAVE A PAYMENT PROFILE AND IN DEPTH (I.E. TIME SERIES PAYMENT DATA WOULD BE USED IN ML INDUCTION)

Finally, Last set would be Domestic or international pay and this can easily be concluded by the last token of the narrative by due to time constraint.

All the above steps has been applied to Test Sample as well.

STEP 6: PREDICTIVE MODEL(S) INDUCTION

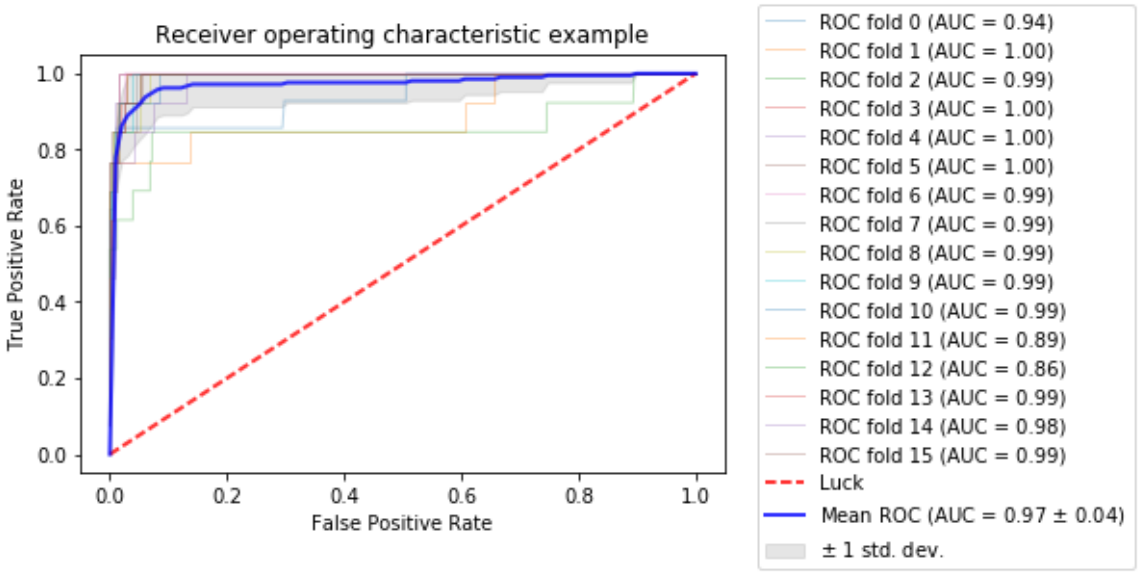
Jupyter Notebook:

5_predictive_modeling

In this step we apply Ensemble Voting ML classifier (using Decision Tree/KNN/RandomForest and AdaBoost) onto our training ABT

As we do not have testing dataset, we employ cross-validation technique so that we use the entire set for learning and before testing.

A final ROC curve with AUC 0.97% is reach with 16 kfold

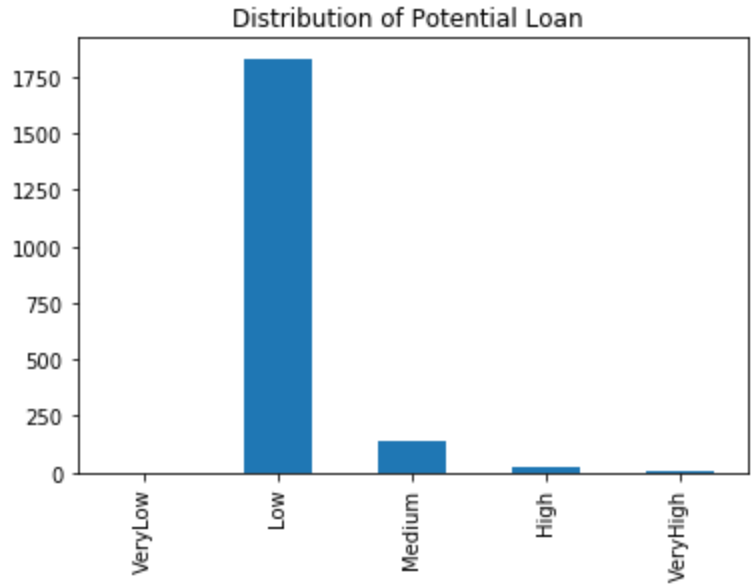


Then test data is applied, with log scale for each of the final category group

```
threathold_groups = {
    'VeryLow' : (0, 0.05),
    'Low' : (0.05, 0.15),
    'Medium' : (0.15, 0.30),
    'High' : (0.30, 0.55),
    'VeryHigh' : (0.50, 1.0),
}
```

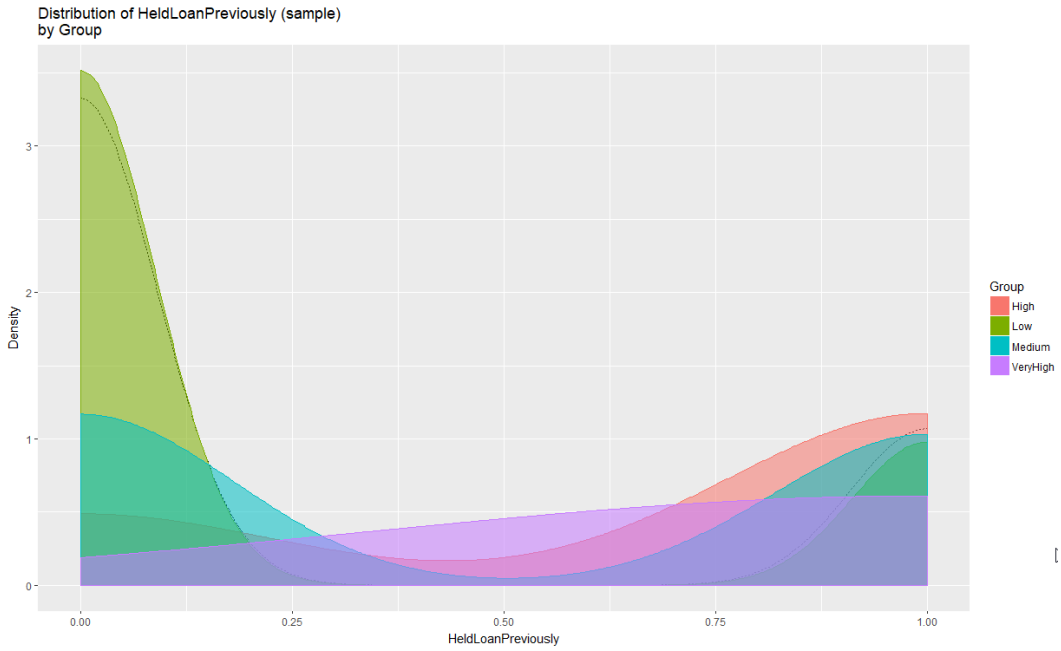
And finally the result is saved at **"../specs/result/RESULT - Model Result.csv"**

```
VeryLow        NaN
Low         1832.0
Medium       139.0
High          21.0
VeryHigh       8.0
Name: Group, dtype: float64
```

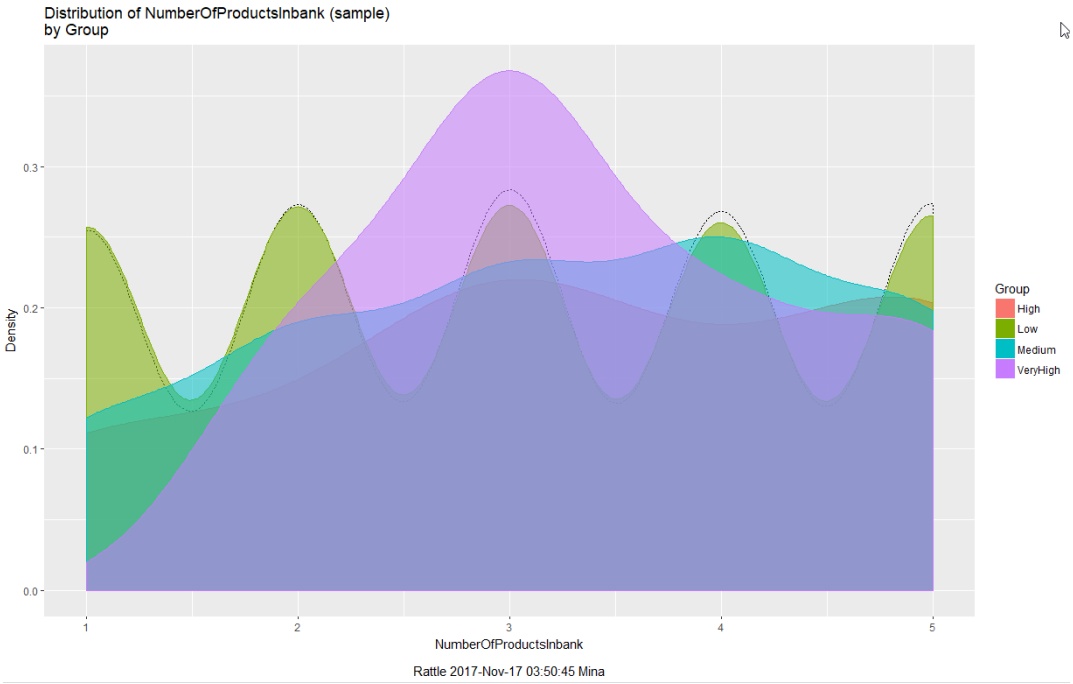


STEP 9: POST-DATA EXPLORATION

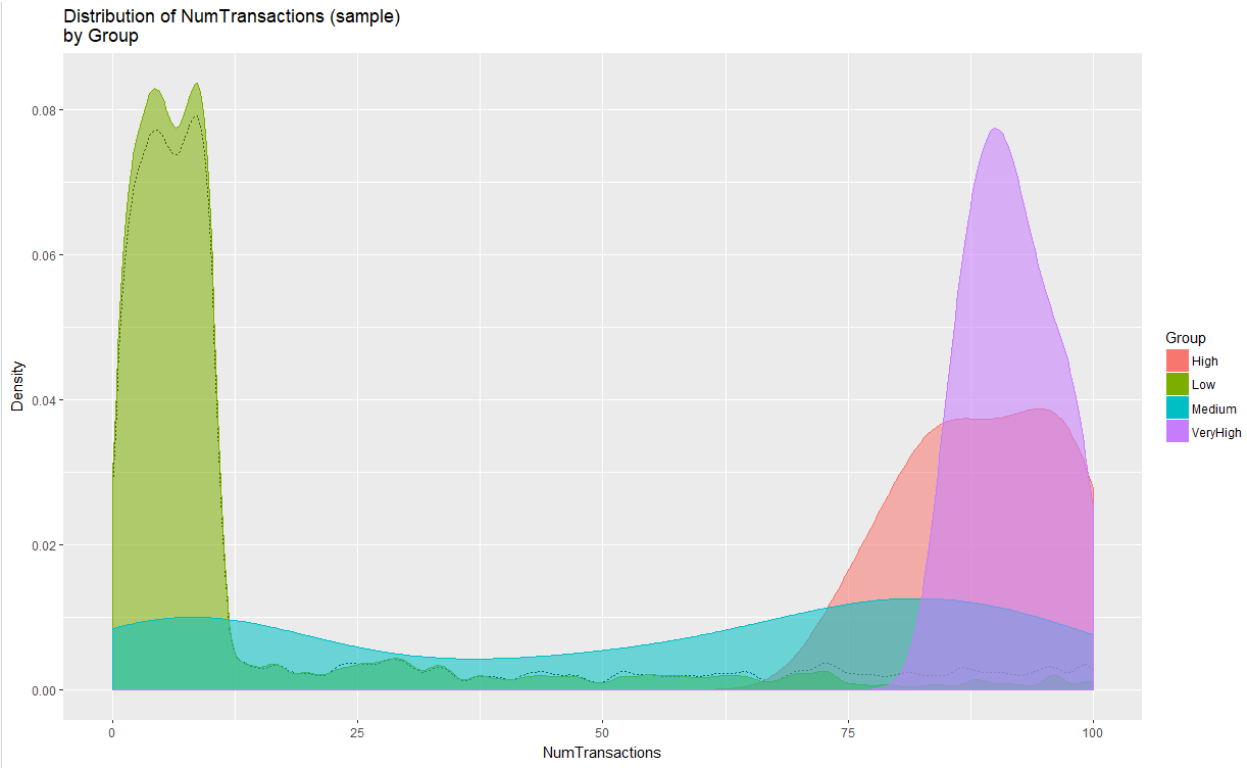
A post data exploration analysis after getting the score, as we can see held previous loan increase the potential of getting new loan.



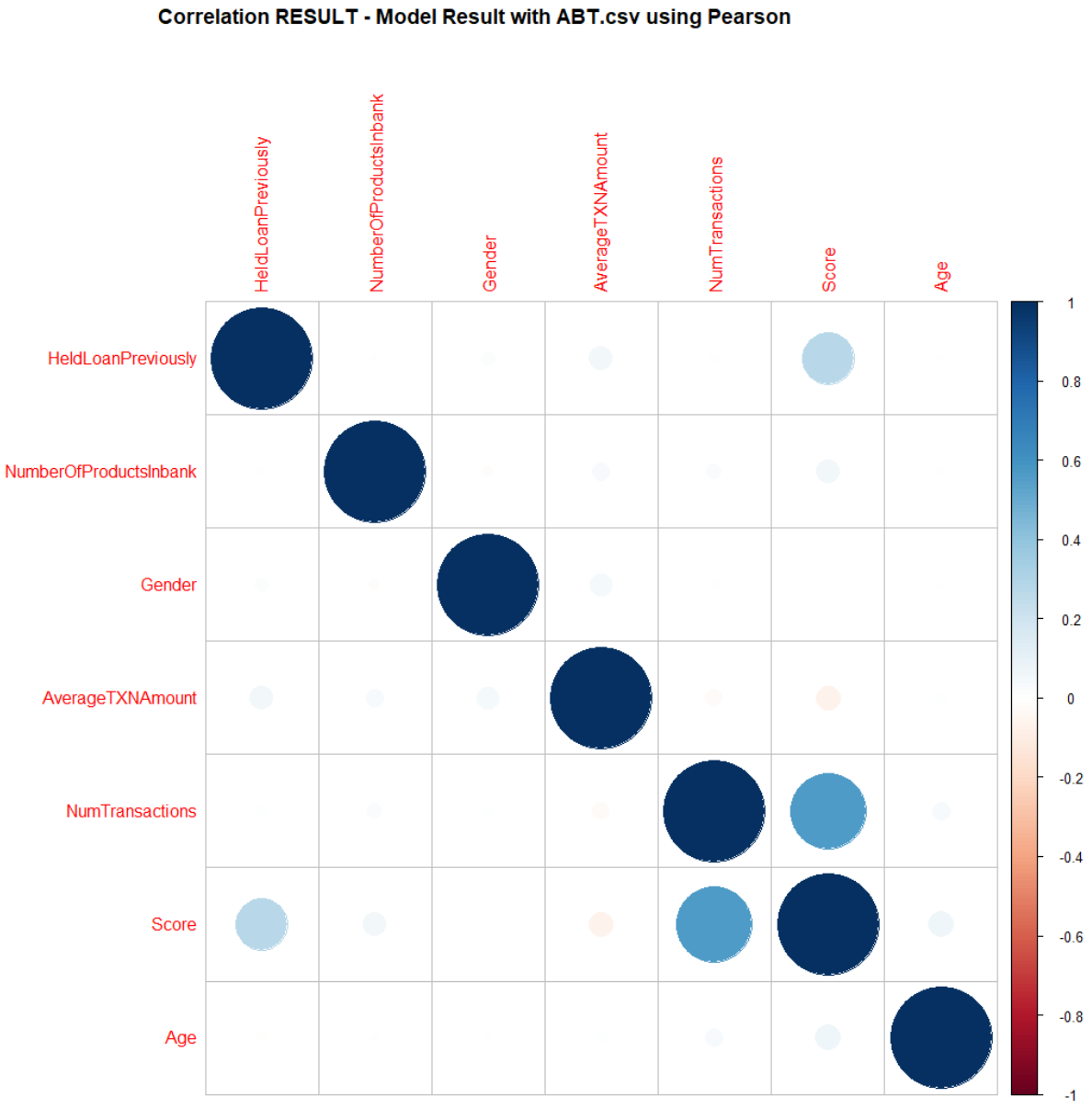
Client holds around 3 bank products are very highly likely to get new loans



High number of transactions have high potential of taking loans



A correlation analysis is performed between the final score and the rest of the features, we can see **HeldLoanPreviously**, **AverageTXNAmount** and **NumberOfProducts** are the main factors influencing the score.



STEP 10: CONCLUSION AND RECOMMENDATION

There is a cut-off age group in the model build data at age 70 while it is up to 80 in the test sample, ETL needed to be investigated and a paired t-test needed to be carried out between number of clients in both trained and test dataset so demographical data is balanced before applying ML.

Two experiments done with/without transaction narrative features and no significant result found out, we either need more transaction payment data and/or expend the activities base.

A Daft.ie rent report data set could be incorporated as a comparative study to see how rent change influence load/mortgage taker, but as the given data set is a time insensitive this won't be feasible.

Another very important attribute could be easily computed is "avg. saving amount / month" where client with high saving numbers are already very high potential mortgage, which need month by month aggregated spending amounts.