Technical Report

Objective

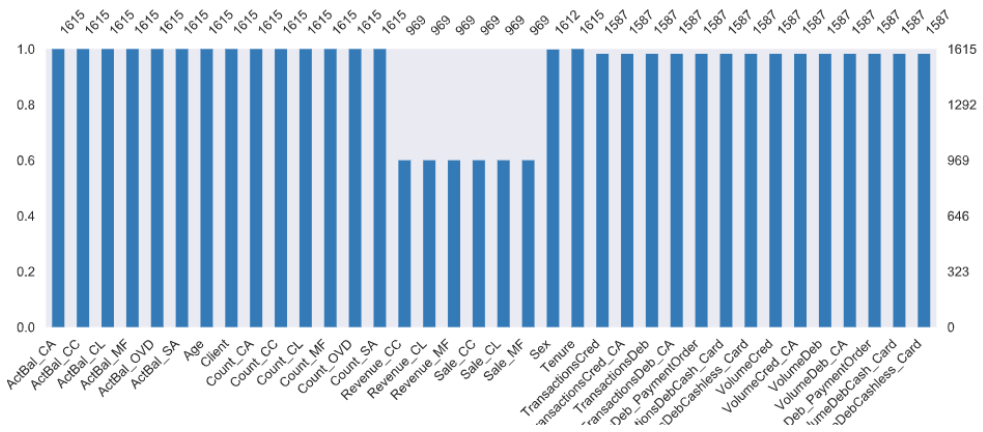
The aim of this study was to analyse and understand the insights about the customer data. The statistical models were employed to predict the Sale Propensity of Clients towards various marketing offers. And eventually optimization strategy was applied to maximize the revenue generated from the various marketing offers (Consumer Loan, Mutual Fund and Credit Card). The Classification and Regression Models were created to provide answers to the proposed questions in this Case Study.

Data Description

The dataset consists of 1615 clients having information pertaining to socio-demographic (age, tenure etc.) and Product account Balance. There were 1587 clients who had information about the inflow and outflow of the amount from their accounts. There were 969 clients having information regarding the sales and revenue on various marketing offers. The dataset consisted of three different marketing offers namely: Consumer Loan, Mutual Fund and Credit Card. In order to create the overall dataset, all the information (consisting of 1615 clients) present in the different tables was merged. The dataset further was divided into *Training Set* (consisting of 969 records i.e. 60% of data) and *Targeting Set* (consisting of 646 records) for which sales and revenues were to be predicted.

Data Handling and Pre-processing

The preliminary insights of data were studied using a 'panda profiling library' where different visualization on features, correlation between the features and missing values were determined. There were some missing values in the dataset which were later imputed. The categorical variables were converted into the integers. There were around 28 missing values present in various columns of the Inflow/Outflow table (VolumeCred, VolumeCred_CA, TransactionsCred, VolumeDeb_CA, etc.). These clients did not have any active accounts and transaction entries, hence the value was assumed to be zero. Later, it can be observed that these imputations had increased the performance of the models. Some ages of the clients were wrong as it could not be possible that 'Tenure' in any company is greater than one's 'Age'. These flaws in 'Age' were corrected by imputing the median of the client's age. Also it was evident from the histogram that range varied from 0 to 100. It is ideally not possible to hold a current savings account with age less than 15 years (assumed).



Feature Selection

In order to estimate the Sale Propensity towards various marketing offers, the proposal of classification model was made. The accuracy of the sale prediction can be improved by selecting important features from the dataset rather than taking irrelevant features. Hence feature selection could be a desirable option to filter the important features from the large number of columns. In this study, the Recursive Feature Elimination was employed to train the model with important attributes. The ranking in features would be displayed along with the details of features to be selected. The selected features will be tagged with 'True' while irrelevant ones with 'False'.

Also the Correlation Statistics was also incorporated to provide the knowledge about the most related features with the target variable (here sales of various offers). There were some features which were positively correlated whereas others were negatively correlated with the target variable. The features VolumeCred was highly correlated (0.937) with VolumeCred_CA, TransactionsDeb was highly correlated (0.918) with TransactionsDeb_CA, TransactionsCred was highly correlated (0.95) with TransactionsCred_CA. Hence one of the features was excluded during model training.

So based on these two methods, the relevant features were extracted for various marketing offers. The correlation values could be observed in the code submitted. The given below is the output representing the important features for offer 'Consumer Loan'.

```
Num Features Customer Loan: 20
Selected Features for Customer Loan: [False False False False  True  True  True  True  True  True  True  True
  True  True  True  True  True False  True  True  True  True False  True
  True False  True]
Feature Ranking Customer Loan: [8 5 4 7 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 6 1 1 2 1]
(969, 27)
```

Methodology and Evaluation Metrics

A. Classification Models for Predicting Sale Propensity of Marketing Offers

MACHINE LEARNING ALGORITHMS IMPLEMENTED

There were three different models created based of the marketing offers (Consumer Loan, Mutual Fund and Credit Card). We will further discuss in the detail about the Classification Algorithms used in training the models.

Classifiers Implemented - Logistic Regression, Random Forest Classifier**, *Stacking*** (XGBoost + Gradient Boost + Logistic Regression), and ***Deep Learning model*** created using Keras API.

The various classical machine learning models were used for the prediction of sales for different marketing offers. In this study, for classifiers Logistic Regression, XGBoost and Gradient Boosting the evaluation metrics results (using recall, precision and AUC metrics) were not satisfactory. The improved approach i.e. ***Stacking*** was applied using the three models *XGBoost, Gradient Boost and Logistic Regression as the Meta Learner*. This technique gave the better results as compared to the other classical machine learning algorithms. Eventually, the ***Deep Learning Model*** with multiple hidden layers was used for classification for which the evaluation metrics results were satisfactory. s. Hence out of the remaining two models (Stacking and Deep Learning), the Deep Learning model was finally selected based on the results and its advantages over other discussed classifiers.

REGULARIZATION

When the model is trained using iterative forward and backward passes using hyperparameter epochs, it might lead to over training of the model on a given dataset which may further lead to overfitting. To reduce the overfitting in statistical models, regularization was used while training the model. The two regularization techniques like ***Early Stopping and Dropout*** were used during the training of the model. It was found that employing Early Stopping without Dropout gave improved results. Therefore, Early Stopping method was applied in all three propensity models to predict the sale of various marketing offers.

SAMPLING IN IMBALANCED DATASET

The hybrid sampling and SMOTE were applied to balance classes in the dataset. But during the training of model using balanced dataset, the accuracy was low and loss was high. Hence the balance in dataset was later excluded from the study as it did not give the expected results.

FEATURE STANDARDIZATION

When the final dataset was created after merging all the relevant tables, it was found that there were some features which were having varying scale. To improve the prediction accuracy, Feature Scaling was performed where all the values are centered around mean with unit variance. Following are the features which were feature scaled : 'VolumeCred_CA', 'VolumeDeb_CA', 'VolumeDebCash_Card', 'VolumeDebCashless_Card', 'VolumeDeb_PaymentOrder', 'ActBal_CA', 'ActBal_SA','ActBal_MF', 'ActBal_OVD', 'ActBal_CC', 'ActBal_CL'.

MODELS DEVELOPED FOR THREE MARKETING OFFERS

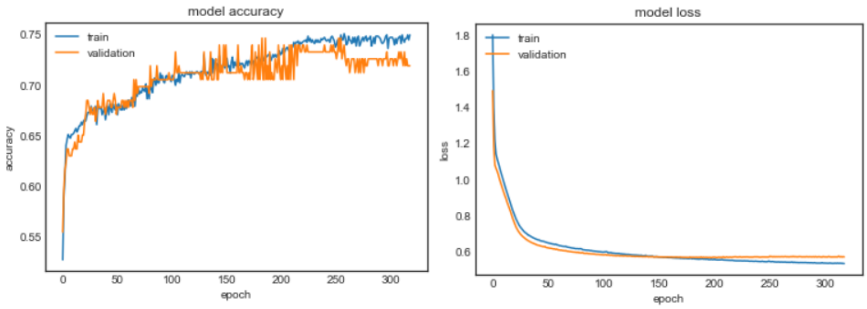
As already discussed the relevant features were selected based on the feature selection techniques as they gave improved evaluation metric results.

1. ***Propensity Model for Consumer Loan using Deep Learning***

Important Features Selected:
['Tenure', 'Sex','Count_CA', 'TransactionsCred_CA', 'TransactionsDebCash_Card', 'ActBal_MF', 'ActBal_OVD', 'TransactionsDeb_CA', 'Count_CL', 'ActBal_CL', 'VolumeDeb_PaymentOrder', 'TransactionsDeb_PaymentOrder', 'Count_MF', 'Count_OVD', 'VolumeDeb_CA', 'VolumeDebCashless_Card', 'Count_SA', 'VolumeDebCash_Card', 'TransactionsDebCashless_Card','Age']

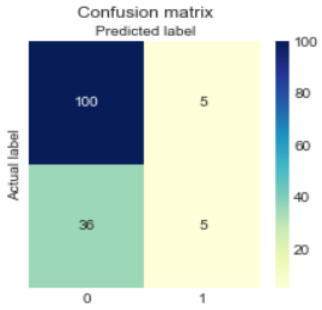
Evaluation

The evaluation was performed on the validation set which was 15% of the training data. After performing evaluation for all the classifiers, it was found that model created using Deep Learning Techniques gave the satisfactory results. Below are the details on Evaluation Metrics and graphs plotted for model accuracy and loss during validation:



	precision	recall	f1-score	support
0.0	0.735294	0.952381	0.829876	105.000000
1.0	0.500000	0.121951	0.196078	41.000000
accuracy	0.719178	0.719178	0.719178	0.719178
macro avg	0.617647	0.537166	0.512977	146.000000
weighted avg	0.669218	0.719178	0.651891	146.000000

Classification Report



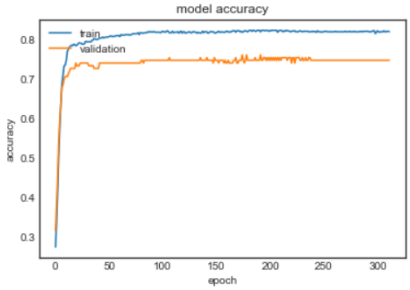
2. ***Propensity Model for Mutual Fund using Deep Learning***

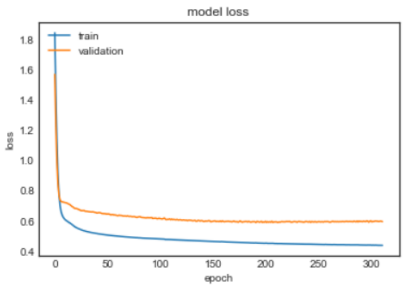
Important Features Selected:

['Tenure', 'Count_CA', 'TransactionsCred_CA', 'TransactionsDebCash_Card', 'ActBal_MF', 'ActBal_OVD', 'TransactionsDeb_CA', 'Count_CL', 'ActBal_CL', 'VolumeDeb_PaymentOrder', 'TransactionsDeb_PaymentOrder', 'Count_MF', 'Count_OVD', 'VolumeDeb_CA', 'VolumeDebCashless_Card', 'Count_SA', 'VolumeDebCash_Card', 'TransactionsDebCashless_Card','Age','Sex']

Evaluation

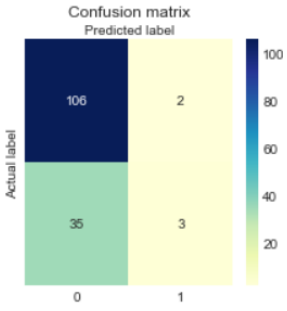
The evaluation was performed on the validation set which was 15% of the training data. After performing evaluation for all the classifiers, it was found that model created using Deep Learning Techniques gave the satisfactory results. Below are the details on Evaluation Metrics and graphs plotted for model accuracy and loss during validation:





Classification Report

	precision	recall	f1-score	support
0.0	0.751773	0.981481	0.851406	108.000000
1.0	0.600000	0.078947	0.139535	38.000000
accuracy	0.746575	0.746575	0.746575	0.746575
macro avg	0.675887	0.530214	0.495470	146.000000
weighted avg	0.712270	0.746575	0.666124	146.000000



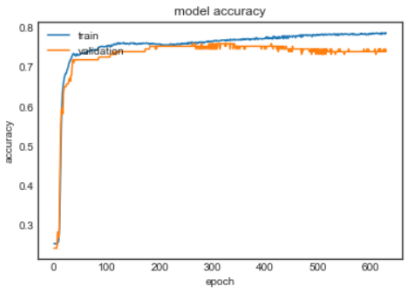
3. ***Propensity Model for Credit Card using Deep Learning***

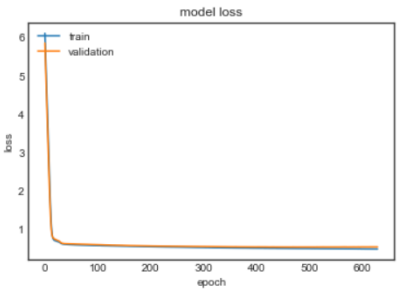
Important Features Selected

['Tenure', 'Count_CA', 'TransactionsCred_CA', 'TransactionsDebCash_Card', 'ActBal_MF', 'ActBal_OVD', 'TransactionsDeb_CA', 'Count_CL', 'ActBal_CL', 'VolumeDeb_PaymentOrder', 'TransactionsDeb_PaymentOrder', 'Count_MF', 'Count_OVD', 'VolumeDeb_CA', 'VolumeDebCashless_Card', 'Count_SA', 'VolumeDebCash_Card', 'TransactionsDebCashless_Card','Age','Sex', 'ActBal_SA']

Evaluation

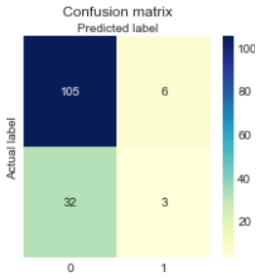
The evaluation was performed on the validation set which was 15% of the training data. After performing evaluation for all the classifiers, it was found that model created using Deep Learning Techniques gave the satisfactory results. Below are the details on Evaluation Metrics and graphs plotted for model accuracy and loss during validation





Classification Report

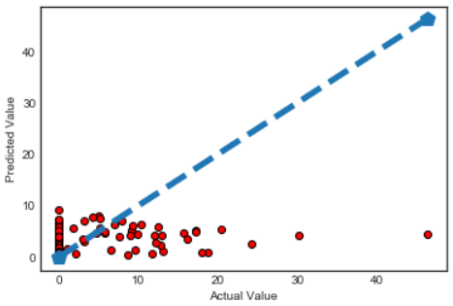
	precision	recall	f1-score	support
0.0	0.766423	0.945946	0.846774	111.000000
1.0	0.333333	0.085714	0.136364	35.000000
accuracy	0.739726	0.739726	0.739726	0.739726
macro avg	0.549878	0.515830	0.491569	146.000000
weighted avg	0.662600	0.739726	0.676470	146.000000



B. Estimation of the Expected Revenue For Various Marketing Offers

Initially, calculation of the revenues for various offers was found using the Regression Models. The linear regression was used to train the model and revenue predictions were made for these offers. The results obtained from this model were evaluated using the metrics like MSE and Adjusted R. The predictions captured from the model were used in plotting the graph to understand the relationship. But it was observed that points were scattered away from the expected line. Therefore, these results were not as accurate as expected when evaluated using metrics and graphs.

To predict the revenue for the given target data (646 data records), it was assumed that mean would be a best predictor to these revenues of different offers. Mean from the training set was calculated for each offer and used as a predicted value for the target data.



Maximize Revenue for Various Marketing Offers

The assumption was made that '**Sale Propensity is proportional to Expected Revenue**', so higher Sale Propensity would lead to a higher Expected Revenue. Based on this consideration various expected revenues were calculated by multiplying the Propensity with Predicted Revenue (mean revenue) for every offer.

The approach below was applied to every offer for calculation of Expected Revenue:

*Expected Revenue = Propensity to buy any offer * Predicted Revenue*

Results and Conclusion

- The three classification models were developed to address the Sale Propensity for various marketing offers (Consumer Loan, Mutual Fund and Credit Card). These models were trained on the training data consisting of 949 records. The Deep Learning models were selected as they gave better results when evaluated on a validation set using metrics Confusion Matrix, Precision, Recall, and Model Accuracy. The Model Loss was low during training and validation.

- The Regularization technique was used to reduce the overfitting of training data. The Feature Standardization was performed on some features which improved the accuracy during Sale Propensity prediction. The 'Adam' optimization algorithm was used to update the weights iteratively which helped in reducing the model loss.

- Below table depicts the Classification Accuracy and Loss while predicting Sale Propensity:

Marketing Offers	Model Accuracy	Model Loss
Consumer Loan	72 %	0.56
Mutual Fund	75%	0.59
Credit Card	74%	0.54

  The Sale Propensity was predicted for every offer on the target data (646 records) and highest propensity was then filtered and stored in corresponding CSV files.

- In order to maximize the revenue generated from the offers, it was assumed that higher the Sale Propensity, higher will be the revenue. Initially, the linear regression was developed to predict the revenues of various offers but results from evaluation metrics (MSE and Adjusted R score) were inaccurate. Also, using the predicted value from the regression model, when the graph was plotted it was found that points were away from the expected linear line.

- As the results from the regression model were inaccurate, the individual 'mean' of offers were used as predicted revenues. The mean is used as predicted value because in statistics, it is assumed in statistics that mean is always a good predictor. To find the Expected Revenues of various offers, the Sale Propensity predicted using classification model was multiplied by Predicted Revenue (individual means). This process maximizes the revenues of every offer pertaining to various clients.

- The list of 15% of the total clients (i.e. 100 clients) having the highest expected revenue was obtained with their respective target offers. **The table below depicts the distribution of marketing offers among different clients when revenue is maximized:**

Marketing Offer	Number of Clients in Each Offer (100 Clients)
Consumer Loan	75
Credit Card	21
Mutual Fund	4

The total expected revenue for 100 clients was estimated to be **835.50**

- **Below are the objectives and their respective results presented in csv files**:
    - *Which clients have higher propensity to buy consumer loan?*

      ClientsList_Propensity
      _Buy_Consumer_Loan

      (The list of CL clients, arranged in the descending order of the Propensity Value)

    - *Which clients have higher propensity to buy credit card?*

      ClientsList_Propensity
      _Buy_Credit_Card.csv

      (The list of CC clients, arranged in the descending order of the Propensity Value)

    - *Which clients have higher propensity to buy mutual fund?*

      ClientsList_Propensity
      _Buy_Mutual_Fund.csv

      (The list of MF clients, arranged in the descending order of the Propensity Value)

    - *Which clients are to be targeted with which offer? General description.*

      Top100_Clients_to_b
      e_Targeted_with_Vari

      (Top 100 Clients with Targeted Offers are displayed based on Expected Revenue)

    - *What would be the expected revenue based on your strategy?*

      The total expected revenue for 100 clients was estimated to be **835.50**

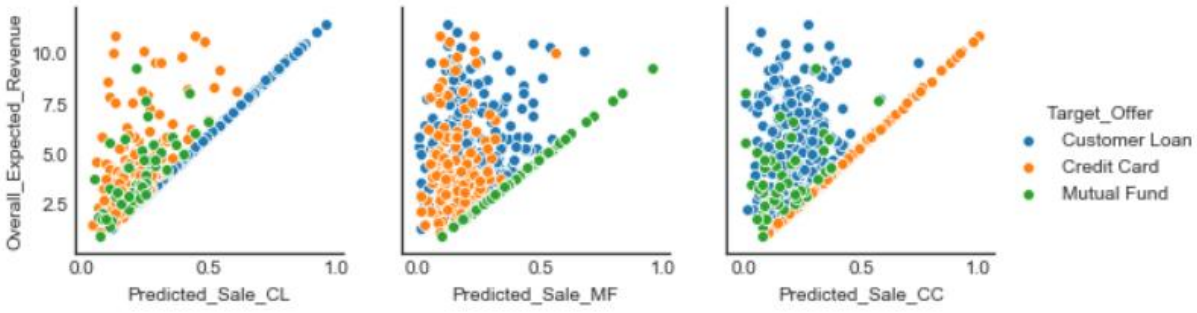
Insights from the Results obtained from the Target Data

- The Consumer Loan clients having good profit are young, working and loyal customers. They are having more number of withdrawals as compared to deposit. The have good credits in current account as compared to other offer clients (mutual fund and credit card).

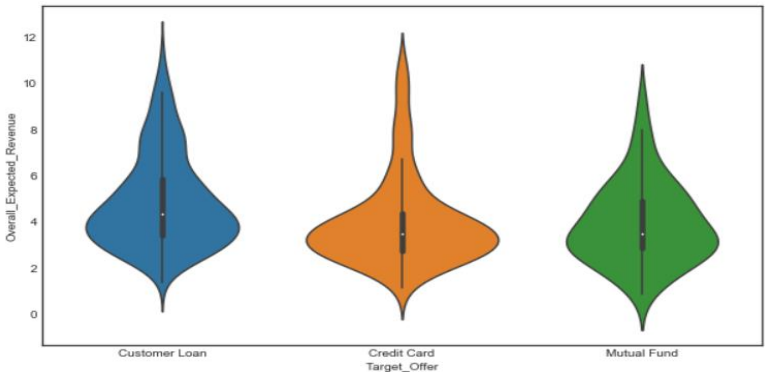
- The Mutual Fund clients having maximum profit have more number of withdrawal transactions as compared to deposit. But good amount is deposited monthly as compared to withdrawal amount. They prefer cashless transaction as compared to withdrawal using cards. The retired customers have less balance in current and saving accounts.

- The Credit Card clients having good profit has more number of withdrawal transactions as compared to deposit. The young credit card customers have more balance amount in current account as compared to retired customers. The retired customers have good amount in saving account as compared to young customers.

- The graphs represent the relationship between the Overall Revenue and Sale Propensity of various marketing offers.



- The violin plot represents the distribution of the revenue across the Target Offers.



Discussion

- Accuracy of the Classification Model could have been improved if more number of data points were present.
- Various other Feature extraction techniques can be used to improve model performance like Principal Component Analysis and Linear Discriminant Analysis.
- The Balancing dataset using Hybrid Sampling and SMOTE did not give satisfactory results.
- In order to estimate the overall revenue, the Sale Propensity was multiplied by the predicted revenues for various offers.
- Used Dropout as regularization method to reduce overfit, but prediction became inaccurate.