INTERNSHIP – DATA ANALYTICS

PROJECT 3
CREDIT CARD CUSTOMER DATA ANALYSIS

TEAM MEMBERS

HRISHIKESH KALITA (EI0123)

AMLAN RANJAN SAHOO (EI0126)

ISHAAN WASSAN (EI0122)

COLLEGE

NATIONAL INSTITUTE OF TECHNOLOGY ROURKELA

GOAL

- Checking the Customer's eligibility to get an approval for Credit Card.
- Building logic in Python for various metrics to check the performance of acquisition strategy adopted by the firm.

DATA STATS

- Shape of the Data: (1000, 18)
- Columns of the Data:
  'Application id', 'first_name', 'last_name', 'email', 'gender', 'address', 'age', 'tdecision', 'empstaus', 'ExCus (Customer in Past)', 'Source', 'Salary', 'ExDebt (Liability)', 'Booking', 'INT_ID', 'Prev_ID', 'AGT_ID', 'Booking_Amt'

- Data Info:

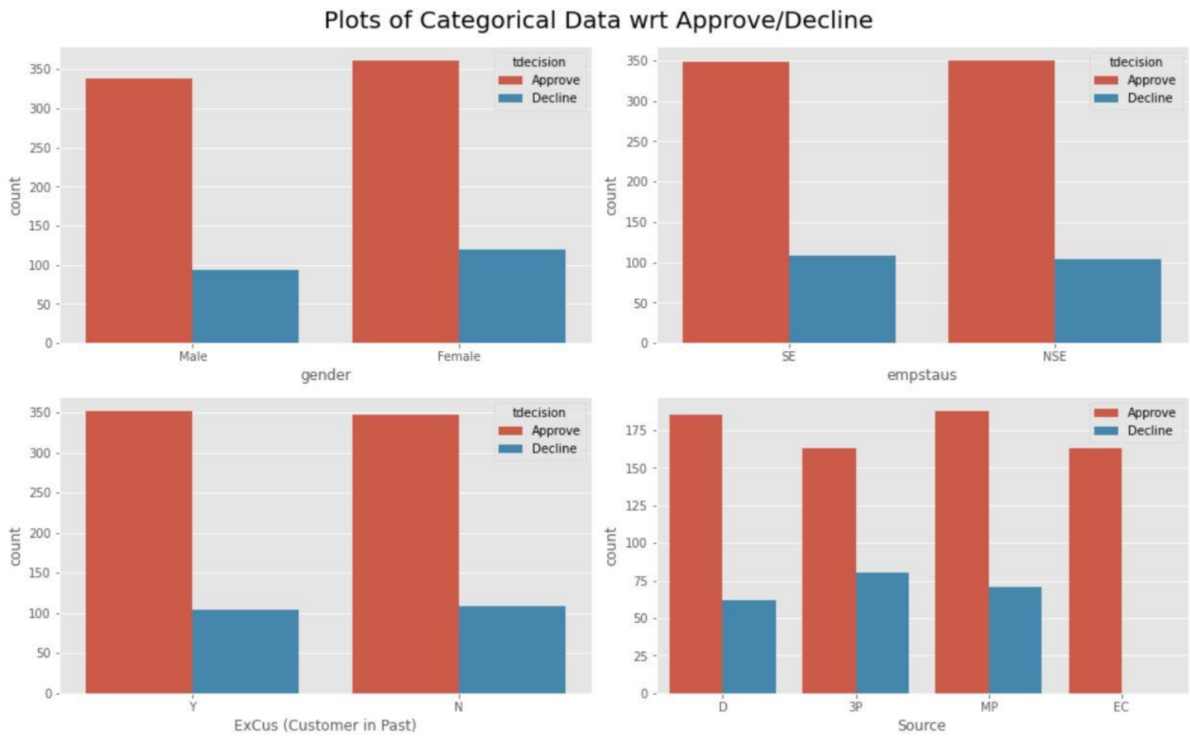
	Application id	age	Salary	ExDebt (Liability)	INT_ID	Booking_Amt
count	1000.000000	1000.000000	1000.000000	1000.000000	1.000000e+03	699.000000
mean	500.500000	43.993000	300965.584000	25719.898000	4.991688e+09	363096.037339
std	288.819436	12.199827	174484.016951	2728.736685	2.902731e+09	192897.386823
min	1.000000	20.000000	1473.000000	21002.000000	1.788664e+07	47470.500000
25%	250.750000	33.000000	147630.000000	23461.000000	2.500389e+09	212743.200000
50%	500.500000	47.000000	299657.500000	25782.500000	5.018401e+09	326617.200000
75%	750.250000	55.000000	452421.500000	28073.500000	7.439011e+09	490002.000000
max	1000.000000	60.000000	597399.000000	30453.000000	9.995180e+09	894333.000000

COMMENTS

- The Booking_Amt columns has only 699 valid entries out of 1000. This suggests that there must be some Null values in that column, which is supported by the fact that not all cards were being approved by the bank.
- We have observed that though the MEAN SALARY is greater than MEAN LIABILITY, but MINIMUM SALARY is lesser than MINIMUM LIABILITY by around 14 times. This gives a prior heads-up on the fact that the people belonging to the low-salaried section are likely to be declined.

EXPLORATORY DATA ANALYSIS

- CATEGORICAL DATA



COMMENTS

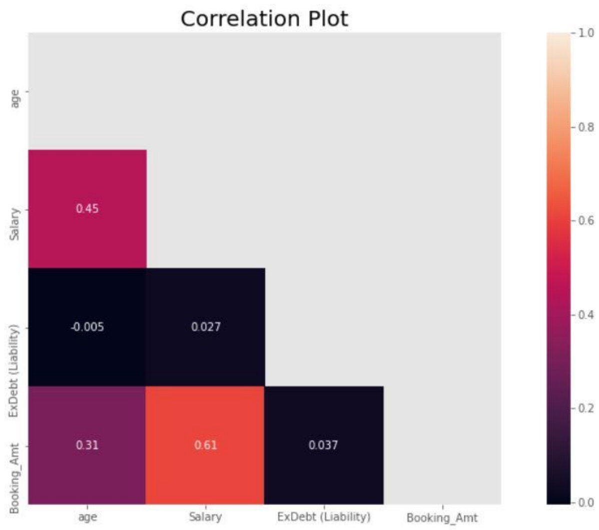
- **Females** are having a **greater** number of **Approvals** as well as **Rejections** than **Males**.
- For both Self-Employed and Non-Self-Employed, the number of Approvals is almost same.
- But, it's interesting to note that, **no credit card of an Existing Customer is being Declined**. Further analysis shows that **all** the customers in the **Pending** section are **Existing Customers**. Perhaps the bank has a little leniency over their existing customers and didn't directly reject their application.


```
pending_cases['Source'].value_counts()
```

```
EC     88
Name: Source, dtype: int64
```

This shows that all the Pending cases are Existing Customers.

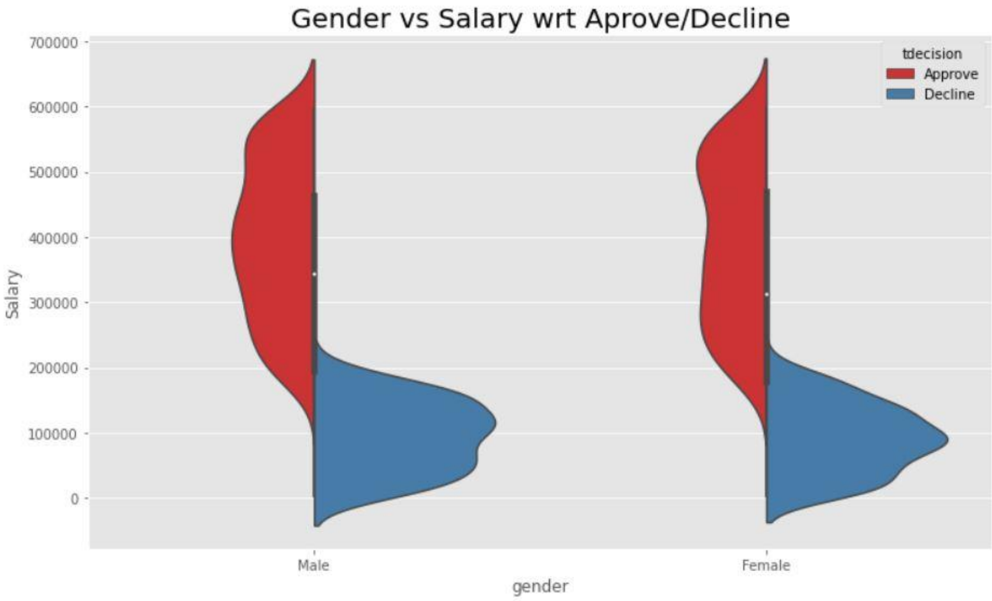
- NUMERICAL DATA

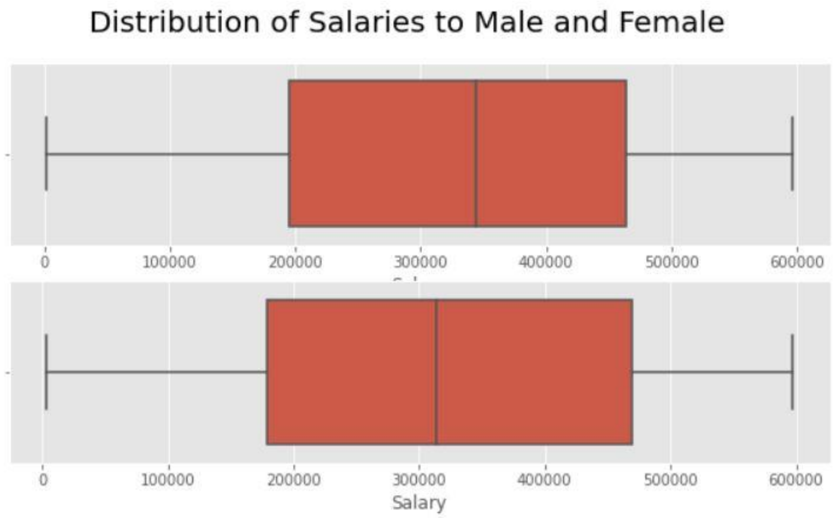


COMMENTS

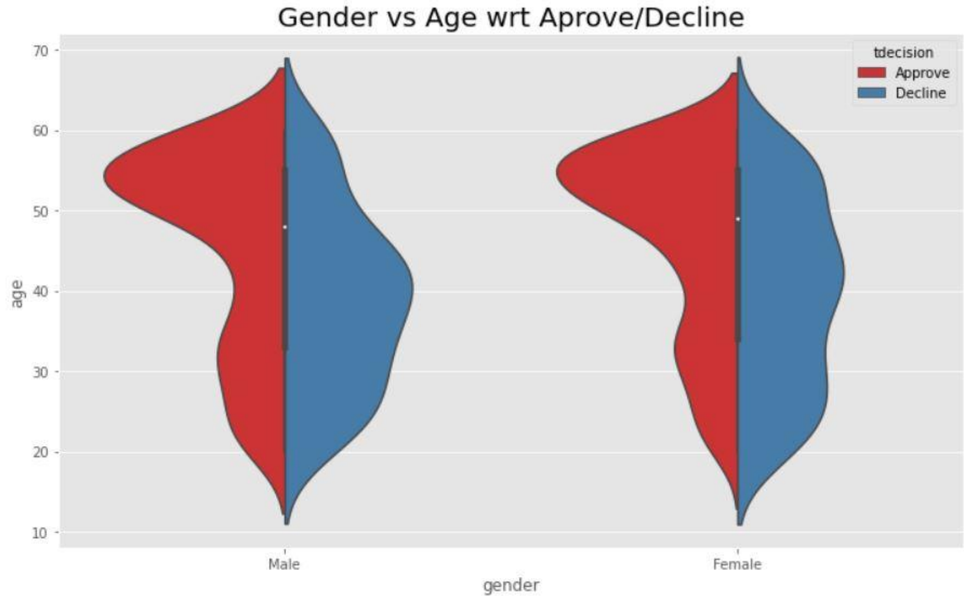
- The intuition that **low-salaried** people will mostly be **Declined** has been show in the above plot.

- In the salary range of **0-50000**, **Males** have a **greater** number of **Declined** credit cards than **Females**.
- However, in salary range of **2nd to 3rd Quartile**, **Males** have **greater** number of **Approvals** than **Females**.





Observed from the tiny box-plot above, that the median of the salaries of the Females is less than that of the Males, but their 3rd Quartile is almost at equal levels. This suggests that there are more Females having salaries in the higher bracket than males. This can be show by plotting their box-plots individually. We observe a bigger 2nd – 3rd Quartile range in case of Females.



COMMENTS

- A higher number of **Declines** are observed for **Males** in the **age** range **25 – 45** (approx..) than **Females**.

MODELLING

MODEL DESCRIPTION

MODEL USED: RANDOM FOREST CLASSIFIER

MODEL HYPERPARAMETERS: {n_estimators = 300, random_state = 3}

MODEL RESULTS

Cross Validation Scores: [98.91, 98.91, 98.9, 99.45, 98.35]

Maximum Accuracy obtained: 99.45%

Standard Deviation among Cross Validation Scores: 0.0035

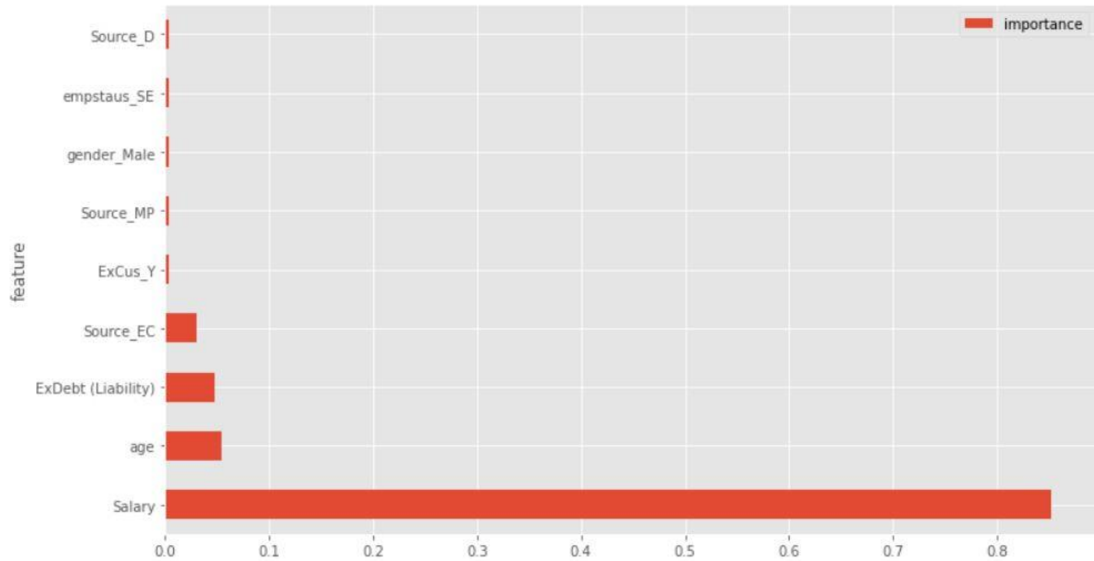
COMMENT: A **low Standard Deviation** among the scores refers that the model is **STABLE**.

feature	importance
Salary	0.852
age	0.054
ExDebt (Liability)	0.047
Source_EC	0.030
ExCus_Y	0.004
Source_MP	0.004
gender_Male	0.003
empstaus_SE	0.003
Source_D	0.003

MODEL INFERENCE

Importance of Columns:

Importance or Weightage of the columns are being calculated using Random Forest Classifier's inbuilt command *feature_importance.* This could be further used in further modelling. In the next step, a Decision Tree classifier is being implemented on the important columns as specified by Random Forest Classifier.



Feature Importance

IMPROVING MODEL

MODEL USED: DECISION TREE CLASSIFIER

COLUMNS USED: ['Salary', 'age', 'ExDebt (Liability)', 'Source_EC']

CLASSIFICATION REPORT:

```
              precision    recall  f1-score   support

     Approve       0.98      1.00      0.99        57
     Decline       1.00      1.00      1.00       217

    accuracy                           1.00       274
   macro avg       0.99      1.00      0.99       274
weighted avg       1.00      1.00      1.00       274
```

COMMENT: The classification report does provide with some good feedback about the model. However, it can also be a case of overfitting due to less amount of data.

PREDICTION OF PENDING CARDS

MODEL USED: RANDOM FOREST CLASSIFIER

```
prediction.Prediction_tdecision.value_counts()
```

```
Approve    78
Decline    10
Name: Prediction_tdecision, dtype: int64
```

File saved as **Predictions_PendingClass_using_RandomForest.xlsx**

MODEL USED: DECISION TREE CLASSIFIER

```
dtc_predictions.tdecision.value_counts()
```

```
Decline    87
Approve     1
Name: tdecision, dtype: int64
```

File saved as **Predictions_PendingClass_using_DecisionTree.xlsx**

METRICS CALCULATION

- Total Applications: **1000**
- Approved Applications: **699**

- Booked Applications: **581**
- Approval Rate: **0.699**
- Booking Rate: **0.831**

The Metric "New Booking Amount" has been appended as a new column and saved as **Data_with_New_Booking_Amount.xlsx**

---THE END---