Bureau_Assignment

Colab link :-
https://colab.research.google.com/drive/1cyXBAflLo3g2iTVxLphl4BDp8obYs_yu?usp=sharing

Approach Taken :-

The approach involves data preprocessing, feature selection, model training, and evaluation using different classifiers. The final model is used to make predictions on the test dataset.

Data Preprocessing :-

Handling Missing Values:

The 'Cibil Score' column was converted to numeric, with non-numeric entries coerced to NaN.

Missing values in categorical columns were filled with 'Unknown', while missing values in boolean columns (e.g., phone social premium features) were filled with 0 and converted to integers

Missing values in numeric columns were imputed using the mean strategy

Rationale:

This approach ensures that all missing values are handled appropriately,either by filling them with a placeholder or imputing with a meaningful statistic (mean), preventing potential bias in model training

Label Encoding:

Categorical variables were converted to numeric values using LabelEncoder

Rationale:

Machine learning models require numeric input; thus, label encoding was necessary to transform categorical features into an appropriate format

Feature Selection:

A correlation matrix was computed with respect to the 'Application Status' to understand the relationship between features and the target variable

	Application Status
Application Status	1.000000e+00
MARITAL STATUS	6.439848e-01
EMPLOY CONSTITUTION	6.163364e-01
EMPLOYER TYPE	5.919047e-01
ASSET CTG	5.571758e-01
ADDRESS TYPE	3.142440e-01
EMPLOYER NAME	2.077450e-01
DEALER ID	4.677303e-02
Pan Name	4.611008e-02
vpa	4.037244e-02
upi_name	2.985155e-02
ASSET MODEL NO	2.957792e-02
Phone Social Premium.housing	2.505566e-02
Phone Social Premium.instagram	2.123996e-02
phone_phoneFootprintStrengthOverall	2.074310e-02
Personal Email Address	1.687788e-02
Phone Social Premium.indiamart	1.378950e-02
Phone Social Premium.shaadi	1.273930e-02
Phone Social Premium.flipkart	1.247774e-02
Phone Social Premium.jiomart	1.083152e-02
DEALER NAME	9.666145e-03
DOB	9.555398e-03
Phone Social Premium.jeevansaathi	7.915630e-03
Phone Social Premium.amazon	4.808518e-03
PRIMARY ASSET MAKE	2.781778e-03
MIDDLE NAME	8.865459e-04

Phone Social Premium.skype	5.983130e-04
TOTAL ASSET COST	-3.122314e-16
Phone Social Premium.toi	-8.344039e-04
Phone Social Premium.whatsapp	-1.002849e-03
APPLIED AMOUNT	-1.119298e-03
FIRST NAME	-1.401646e-03
Phone Social Premium.byjus	-1.579127e-03
name	-1.593272e-03
Phone Social Premium.microsoft	-2.428611e-03
Primary Asset Model No	-3.240593e-03
Phone Social Premium.isWABusiness	-5.653751e-03
mobile	-9.348624e-03
Phone Social Premium.paytm	-1.195215e-02
HDB BRANCH STATE	-1.261846e-02
Phone Social Premium.zoho	-1.577859e-02
HDB BRANCH NAME	-1.643720e-02
phone_nameMatchScore	-1.879666e-02
GENDER	-2.493885e-02
phone_digitalage	-3.741528e-02
AGE	-4.407300e-02
LAST NAME	-8.786085e-02
Cibil Score	-5.001575e-01

Redundant and irrelevant features such as 'APPLICATION LOGIN DATE', 'AADHAR VERIFIED', 'MOBILE VERIFICATION', and specific 'Phone Social Premium' features were dropped

Rationale:

Dropping irrelevant or redundant features reduces the complexity of the model and improves its performance by focusing on the most important features

Splitting Data:

The dataset was split into training and validation sets with an 70-30 split

Rationale:

Splitting the data ensures that the model is trained on one subset and validated on another, preventing overfitting and providing an unbiased evaluation of the model's performance

Model Training and Evaluation

Four classifiers were initially chosen for training and evaluation:

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest
- K-Nearest Neighbors (KNN)

Each classifier was trained on the training set and evaluated using accuracy, precision, recall, F1-score, and a classification report.

Rationale for Classifiers:

A diverse set of classifiers was chosen to explore different approaches, including linear models (Logistic Regression), non-linear models (SVM, Random Forest), and instance-based learning (KNN). This helps in identifying the best performing model for the given data

Result :-

```
Logistic Regression:

Accuracy: 0.6650
              precision    recall  f1-score   support

           0       0.67      1.00      0.80      1995
           1       0.00      0.00      0.00      1005
```

```
SVM:

Accuracy: 0.6650
              precision    recall  f1-score   support

           0       0.67      1.00      0.80      1995
           1       0.00      0.00      0.00      1005
```

```
Random Forest:

Accuracy: 0.8857
              precision    recall  f1-score   support

           0       0.92      0.91      0.91      1995
           1       0.82      0.84      0.83      1005
```

```
KNN:

Accuracy: 0.6203
              precision    recall  f1-score   support

           0       0.68      0.82      0.74      1995
           1       0.39      0.23      0.29      1005
```

Final Model Selection

Random Forest Classifier:

- The Random Forest model was chosen for final predictions due to its balanced performance in handling imbalanced classes, robustness, and ability to model complex relationships between features.

Model Parameters:

- `n_estimators=100`: A sufficient number of trees to ensure diversity in the forest.
- `max_depth=None`: Allows trees to grow fully, capturing more complex patterns.
- `class_weight='balanced'`: Ensures the model handles class imbalance effectively.

Conclusion

The Random Forest Classifier was selected as the final model after comparing the performance of multiple classifiers. The model was tuned to handle imbalanced classes and was successfully used to predict application statuses on the test set. The final predictions were saved in a CSV file for further analysis or deployment.