Credit Card Behaviour Score

PRESENTED BY:- TEAM MAVERICKS

1. S.SHIVA SAI
2. G.MAHESH

# Table Of Contents:

1. Introduction
2. Problem Statement
3. Proposed Solution
4. Technology Stack
5. Implementation
6. Results And Evaluation
7. Impact and Future work

1. Introduction

Bank A, a leading credit card provider, aims to develop a "Behaviour Score"—a predictive model to proactively manage credit risk among existing customers. This model will use data like transaction history, payment behavior, and demographics, employing machine learning techniques. The Behaviour Score will identify high-risk customers, enabling targeted risk mitigation, optimized credit limits, and informed credit policy decisions, ultimately aiming for a more stable and profitable credit card portfolio by moving from reactive to proactive risk management.

2. Problem Statement

Bank A seeks to enhance its risk management framework by developing a predictive model, the "Behaviour Score," for its existing Credit Card customers. This model aims to estimate the probability of a customer defaulting on their credit card obligations.

To achieve this, the bank has provided a historical dataset of 96,806 Credit Card accounts, including various attributes such as:

\* On-us attributes: Information specific to the bank's internal assessment of the customer (e.g., credit limit, internal risk scores).

\* Transaction-level attributes: Transactional behavior of the customer (e.g., number of transactions, spending patterns, merchant categories).

\* Bureau tradeline-level attributes: Information from credit bureaus (e.g., credit history, loan balances, payment history).

\* Bureau enquiry-level attributes: Recent inquiries made by the customer for credit (e.g., loan applications, credit card applications).

Using this dataset, the objective is to build a robust and accurate predictive model that can effectively identify customers at high risk of default. The Behaviour Score will then be utilized for various portfolio risk management activities .

3. Proposed Solution

Given the large dataset with 96,806 instances and 1216 features, a key focus was on dimensionality reduction and addressing class imbalance to prevent overfitting and improve model performance.

1. **Data Preprocessing and Feature Engineering:**
   * **Dimensionality Reduction:** The initial dataset with 1216 features was reduced to 4 key features, excluding "account\_number." This was done to simplify the model, improve training efficiency, and potentially mitigate overfitting.
   * **Handling Class Imbalance:** The dataset likely exhibited class imbalance with a higher proportion of non-defaulted customers. To address this, the SMOTE (Synthetic Minority Over-sampling Technique) was employed to generate synthetic instances for the minority class (defaulted customers), creating a more balanced dataset for model training.
2. **Model Selection and Training:**
   * **Ensemble Learning:** An ensemble learning approach was implemented using a combination of:
     + Decision Tree Classifier
     + K-Nearest Neighbors (KNN)
     + Logistic Regression
     + Random Forest Classifier
   * **Hard Voting:** The predictions from these individual models were combined using a hard voting mechanism, where the final prediction was determined by the majority vote.
3. **Evaluation and Refinement:**
   * The model's performance was evaluated using metrics such as precision, recall, and F1-score.
   * Precision was observed to be around 87%, indicating a high proportion of correctly predicted defaulters among the predicted positive instances.
   * Recall and F1-score were also found to be near 87%, suggesting a good balance between precision and recall

4. Technology Stack

* **Programming Language:** Python
* **Machine Learning Libraries:**
  + Scikit-learn: For implementing machine learning algorithms like Decision Tree, KNN, Logistic Regression, Random Forest, and SMOTE for oversampling.
* **Data Analysis Libraries:**
  + Pandas: For data manipulation and analysis (likely used for loading and preprocessing the dataset).
  + NumPy: For numerical computations (used in conjunction with Pandas and scikit-learn).
* **Other Libraries:**
  + Matplotlib/Seaborn: For data visualization (you might have used these to create the correlation matrix).

**Example Technology Stack Section:**

The project leverages the following technologies:

* **Python:** The primary programming language used for data analysis, preprocessing, model development, and evaluation.
* **Scikit-learn:** A comprehensive machine learning library in Python, providing tools for classification, regression, clustering, dimensionality reduction, model selection, and preprocessing. It was used for implementing the Decision Tree, KNN, Logistic Regression, and Random Forest classifiers, as well as the SMOTE technique for oversampling.
* **Pandas:** A powerful data analysis library in Python, used for data manipulation, cleaning, and analysis. It facilitated efficient loading, preprocessing, and transformation of the credit card dataset.
* **NumPy:** A fundamental library for numerical computations in Python, providing support for arrays, matrices, and mathematical functions. It was used in conjunction with Pandas and scikit-learn for data manipulation and model development.
* **Matplotlib/Seaborn:** Data visualization libraries used to create the correlation matrix and potentially other visualizations for exploratory data analysis and model evaluation.

5. Implementation

**1. Project Structure:**

* The project was organized using Jupyter Notebooks for code development and documentation.
* The main notebook, credit\_card.ipynb, contains the core logic for data loading, preprocessing, model training, evaluation, and visualization.
* The libraries required for the project are mentioned in requirements.txt file .

**2. Data Loading and Preprocessing:**

* **Data Loading:** The dataset was loaded from the provided CSV files using the pandas library.
* **Data Cleaning:**
  + Missing values were imputed using the mode for numerical features
  + Outliers were identified and handled using robust scaling.
* **Feature Selection:**
  + The 1216 dimensions are reduced to 4 dimensions excluding the account\_number feature.
  + The features onus\_attribute\_ and transaction\_attribute\_ were analyzed for potential multicollinearity.
  + Feature importance scores from tree-based models were used to identify and rank the most influential features.
* **Feature Engineering:**
* **Data Scaling:** Numerical features were standardized using StandardScaler from scikit-learn to ensure they were on the same scale.

**3. Model Training and Evaluation:**

* **Model Initialization:** The following models were initialized from the scikit-learn library:
  + Logistic Regression
  + Decision Tree Classifier
  + K-Nearest Neighbors (KNN)
  + Random Forest Classifier
* **Ensemble Learning:** A hard voting ensemble was implemented, where the final prediction was determined by the majority vote among the predictions from the individual models.
* **Training:** Each model was trained on the preprocessed training data.
* **Evaluation:** The trained models were evaluated on a holdout test set using the following metrics:
  + Precision
  + Recall
  + F1-score

**4. Hyperparameter Tuning:**

* Hyperparameters for each model were tuned using grid search technique to optimize performance.
* The search space for hyperparameters was defined based on domain knowledge and preliminary experiments.

**5. Visualization and Results:**

* Matplotlib and seaborn libraries were used to visualize the data, feature importance, model performance, and other relevant insights.
* Results were summarized in tables and charts, including confusion matrices and performance metrics.

6.Results

These are the insights obtained by our model on the test data .i.e.

|  |  |
| --- | --- |
| Metrics | Score |
| ACCURACY | 87.2 |
| PRECISION | 87.6 |
| RECALL | 86.6 |
| F1\_SCORE | 87.1 |

confusion matrix:

|  |  |  |
| --- | --- | --- |
| Predicted | Actual Positive | Actual Negative |
| Positive | 20949 | 2912 |
| Negative | 3189 | 20667 |

7.Evaluation

Here are some bullet points summarizing the analysis of the loan data based on the visualizations from the given Validation dataset:

**Key Observations:**

* **Distribution of Attributes:**
  + "Onus Attribute" and "Transaction Attribute" show a wide range of values, suggesting diverse loan sizes and applicant profiles.
  + "Bureau Enquiries" follow a right-skewed distribution, with most applicants having a few inquiries, but some having a large number.
  + "Bureau" attribute displays a wide range with a long tail, indicating a mix of applicants with varying credit histories.
* **Relationship with Bad Flag:**
  + Higher "Onus Attribute" and "Transaction Attribute" values are associated with a higher likelihood of a "Bad Flag" loan.
  + A higher number of "Bureau Enquiries" might be associated with a less stable credit history and potentially higher risk.
* **Correlations:**
  + Strong positive correlation between "Onus Attribute" and "Transaction Attribute" suggests a potential link between loan size and applicant financial capacity.
  + Moderate negative correlation between "Bureau Enquiry" and "Bad Flag" suggests that a high number of inquiries might not always be a positive sign.
* **Scatter Plot Analysis:**
  + Bad loans tend to have higher values on both "Onus Attribute" and "Transaction Attribute."
  + Bad loans are often associated with higher "Onus Attribute" and a wider range of bureau inquiries.

**Overall:**

* The data suggests a complex interplay of factors in determining loan risk.
* High "Onus Attribute" and "Transaction Attribute" values seem to increase risk.
* The role of credit history and inquiry numbers is less clear-cut.
* A combination of these factors likely determines the fate of each loan application.

8.BENEFITS

Here are the key advantages of the credit card behavior score prediction software in bullet points:

**1. Enhanced Risk Management**

* Early identification of at-risk customers
* Improved portfolio health

**2. Increased Profitability**

* Optimized credit limits
* Differentiated pricing
* Reduced operating costs

**3. Improved Customer Relationships**

* Personalized service
* Early intervention and support

**4. Competitive Advantage**

* Data-driven decision making
* Innovation

**5. Compliance and Regulatory Adherence**

* Better compliance
* Reduced regulatory risk

9.Conclusion

The Behaviour Score model developed for Bank A will be a valuable tool for risk management activities. By predicting the probability of customer default, the bank can proactively identify and manage at-risk accounts, optimize credit limits, and make informed decisions regarding interest rates and other credit policies.