Predicting Customer Default Payments

Introduction:

1. Problem Identification
2. Data Preparation
3. Data Modelling
4. Evaluation
5. Summary of Data

Project Objective:

The primary objective of the study was to evaluate and compare the effectiveness of different data mining methods in predicting the probability of customer default payments. This is particularly significant in the context of risk management where understanding the likelihood of default can help in making informed decisions.

1)Problem Identification :

Approach of the Project:

- Data Preparation
- Finding answer in the data
- Building supervised Learning models
- Finally, Predict default payment.

Background of the Project:

- Focus on Predictive Accuracy:- Rather than simply classifying customers as either credible or not credible, the study emphasizes the importance of predicting the actual probability of default. This approach provides a more nuanced understanding of risk, allowing for better risk management and more precise decision-making.
- Sorting Smoothing Method:- To overcome the challenge of estimating the real probability of default.
- Default on CC Bills payment can result in great financial loss
- In order to reduce or even prevent loss of this kinds, bank need to determine appropriate given credit for each specific client based on their informations.
-

Data Preparation :

21 attributes in total including education, sex, age , ...etc

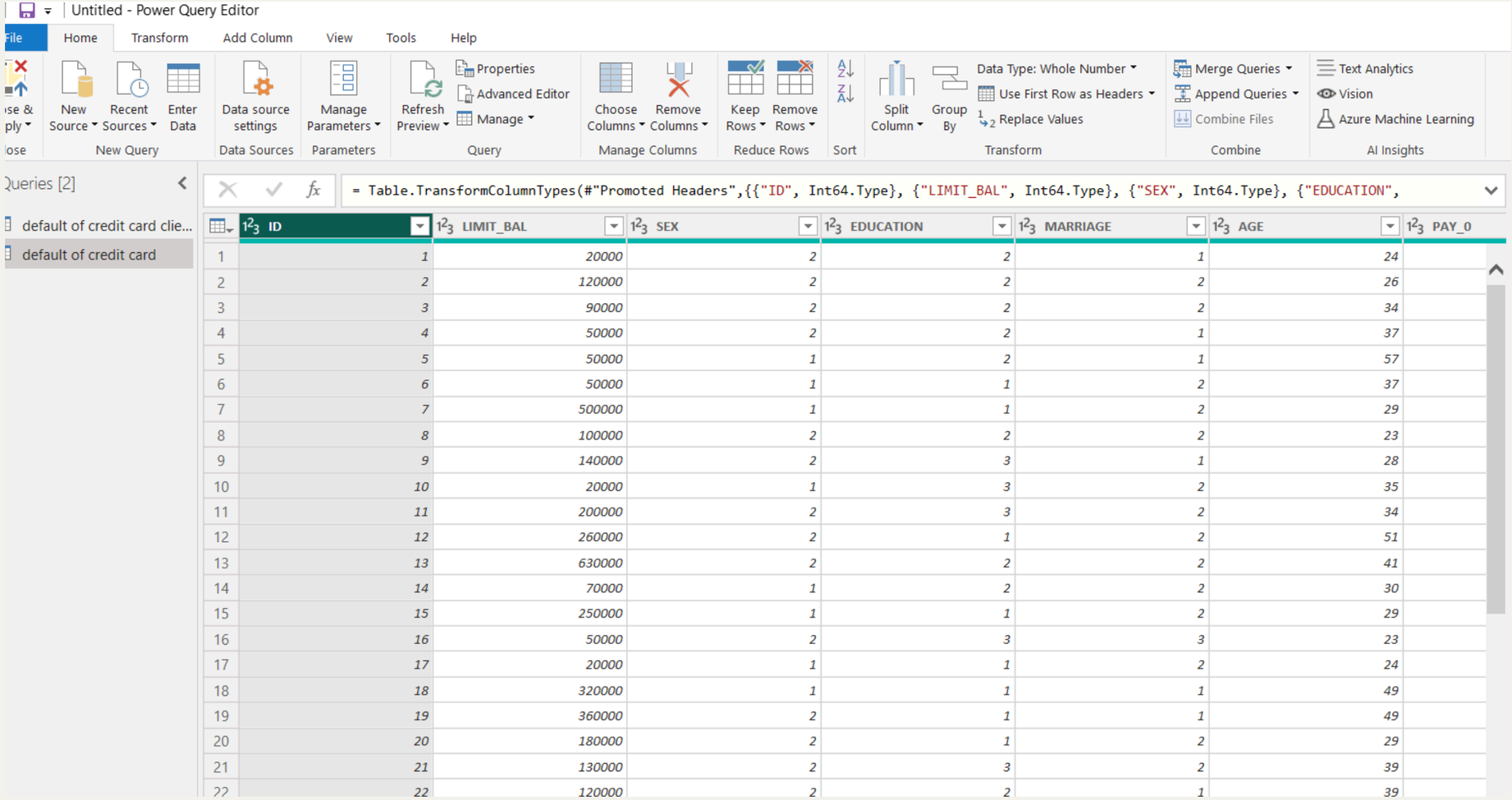
30,000 records

Explorer | default of credit card clients.csv

	A	B	C	D	E	F
1	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE
2	1	20000	2	2	1	24
3	2	120000	2	2	2	26
4	3	90000	2	2	2	34
5	4	50000	2	2	1	37
6	5	50000	1	2	1	57
7	6	50000	1	1	2	37
8	7	500000	1	1	2	29
9	8	100000	2	2	2	23
10	9	140000	2	3	1	28
11	10	20000	1	3	2	35
12	11	200000	2	3	2	34

Dealing with dummy var:-

Original dataset contains categorical variables “education”, “sex”, “marriage”, that have numerical values ranging from 1-4 . We converted them into dummy variables to fit models on them.



	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0
1	1	20000	2	2	1	24	
2	2	120000	2	2	2	26	
3	3	90000	2	2	2	34	
4	4	50000	2	2	1	37	
5	5	50000	1	2	1	57	
6	6	50000	1	1	2	37	
7	7	500000	1	1	2	29	
8	8	100000	2	2	2	23	
9	9	140000	2	3	1	28	
10	10	20000	1	3	2	35	
11	11	200000	2	3	2	34	
12	12	260000	2	1	2	51	
13	13	630000	2	2	2	41	
14	14	70000	1	2	2	30	
15	15	250000	1	1	2	29	
16	16	50000	2	3	3	23	
17	17	20000	1	1	2	24	
18	18	320000	1	1	1	49	
19	19	360000	2	1	1	49	
20	20	180000	2	1	2	29	
21	21	130000	2	3	2	39	
22	22	120000	2	2	1	39	

Data Modelling:-

Objective:
Predict if the given client will default on their next payments

Notes:-

- 30000 observations , 22% base rate.
- For all the models, 5-fold cross validation is applied
- We tried different parameters associated with each algorithm and picked the one will the best evaluation performance.

Algorithms used:

- SVM
- Random Forest
- Neural Networks

Evaluation:-

Basically,
SVM draws a boundary to separate different classes of data and tries to maximize the margins between the class and boundaries.

Neural Networks

- Multi-Layers perceptron , using propagation to Learn.
- 3 hidden layers (5,10,2 units), learning rate, epoch 500

Random Forest

- Ensemble of classification trees
- The models builds many classification trees by taking in different orders of features.
- The forest chooses the classification having the most votes.

Evaluations:

Accuracy

- Base rate of our data set given is 22%
- If the model blindly predicts all the clients not to default, we will be able to generalize as well as 78% accuracy.
- Accuracy is not a cost-sensitive measure.

Accuracy

SVM	Neural Networks	Random Forest
77.88%	77.88%	78.76%

Exploratory Data Analysis (EDA):

- Distribution Plots:

Plot histograms and box plots for numerical features (credit amount, age, bill amounts, previous payments) to understand their distribution.

- Categorical Analysis:

Use bar charts to analyze the distribution of categorical features (gender, education, marital status) and their default rates.

- Correlation Analysis:

Compute and visualize correlations between numerical features and the target variable. For example, calculate correlations between credit amount and default, or between previous payment amounts and default.

- Group Comparisons:

Use bar charts or box plots to compare default rates across different categories (e.g., education levels or marital status).

Preprocessing Steps:

- Handling Missing Values: Impute or handle missing values in features. Common strategies include mean imputation for numerical values and mode imputation for categorical values.
- Normalization/Standardization: Scale numerical features such as credit amount, age, bill amounts, and previous payments to ensure that all features contribute equally to the model training..

Feature Engineering:

- Aggregating Payment History Create summary features from the repayment history, such as average delay, maximum delay, or a count of delayed months.
- Debt-to-Income Ratio: Calculate ratios like the debt-to-income ratio if income data is available or can be inferred.
- Predictive Modeling:
- Data Splitting: Split the dataset into training and testing sets (e.g., 70% training and 30% testing) to evaluate model performance

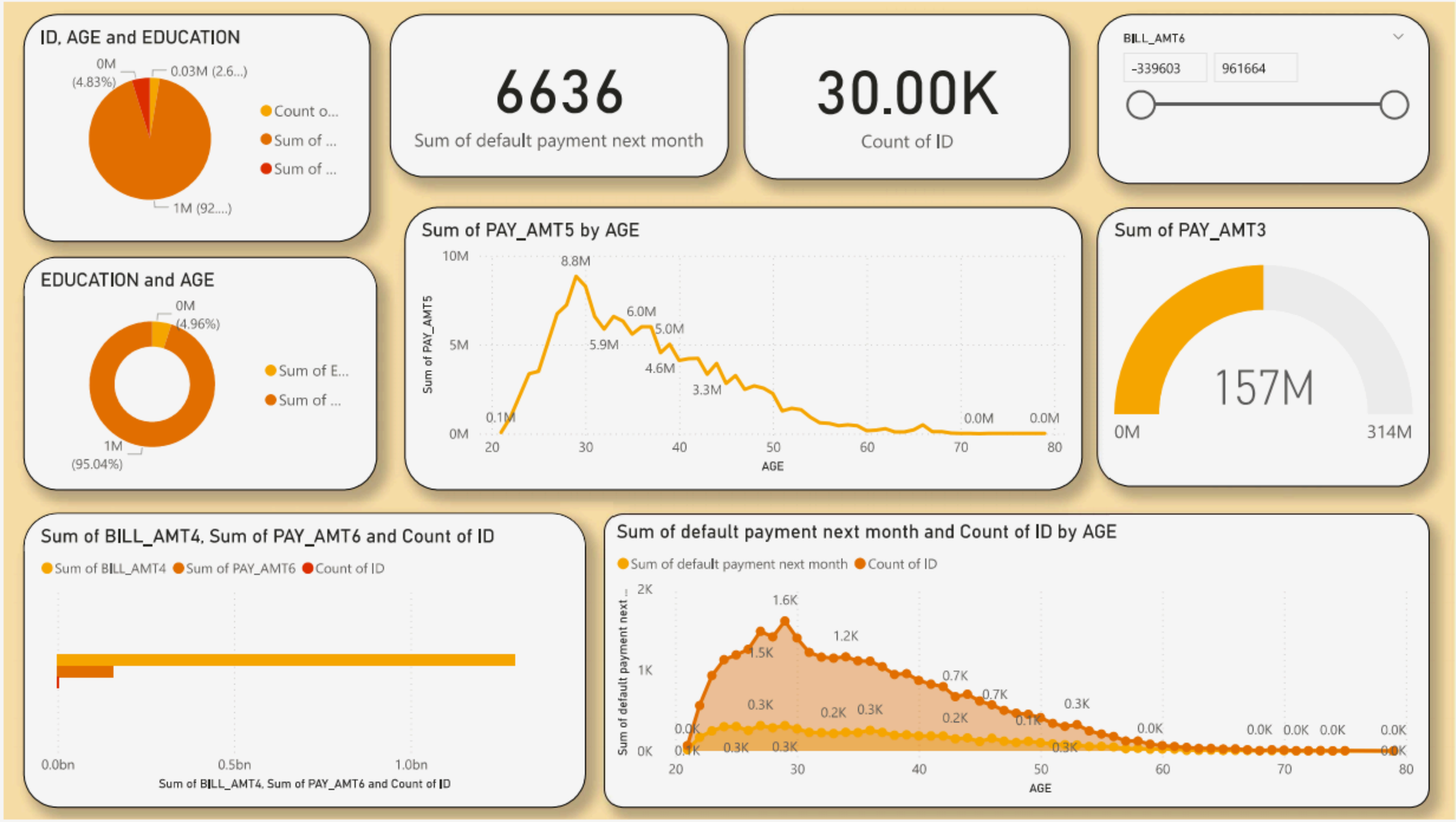
Interactive Visualizations:

- Tools:
- Use interactive dashboards in tools Power BI to allow users to explore the data, filter by different features, and drill down into specifics.

Conclusion:

- Accuracies yield from all the models are no better than simply predicting all client not to default.
- Instead , we are AUC as evaluating measure, which is more directly and naturally related to cost/benefit analysis
- Based on AUC, random forest has best performance
-

Dummy Report



Thank You