Capstone Project (Classification)

Credit Card Default Prediction

By: Kumari Rashmi

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Data Pre-processing: Before Modelling

D. Model Creation and Evaluation

E. Conclusion: Key Takeaways

What this presentation is about?

AI

About Credit Card Default:

- Missing credit card payments once or twice does not count as a default, instead card owner is considered as Delinquent for that period
- Default occurs when customer fails to pay the minimum due amount for a few consecutive months
- The standard period of 6 months is widely used

Why credit card default a risk?

- **For Issuers**: Issuer sees credit card default as a loss, and report as charged off to the credit bureaus. Issuer may file a lawsuit against defaulter or sell the debt to a debt collection agency
- **For Customers**: Default results in dropped credit score, late fees & increased interest, which will further increase the outstanding amount rapidly and let the debt spiral out of control

Business Goals:

- **Minimize Loss/ Risk** by correctly identifying all defaulters or as many as possible in advance
- **Maximise Business** by minimizing the misidentifications of good customers (non-defaulters) as defaulters

EDA: Insights Generation

- Identifying Underlying Patterns & spotting irregularities
- Drawing actionable insights
- Demographic & Behavioural factors and their relationships with defaults

Modelling:

- Build Classification models to predict whether a customer will default on payment next month or not
- Choose relevant evaluation metrics
- Evaluate models and choose best performing Model

Basic Information about Dataset

- This dataset contains information on default payments of credit card clients in Taiwan from April 2005 to September 2005
- There are total 30,000 observations with 25 features, each customer is identified by unique customer Id
- Out of 25 features, 15 features are numerical, and rest are categorical
- Demographical Features: Gender, Age, Marital Status & Education
- Behavioral Features: Past 6 months of paid amounts, Past 6 months of Bill amounts, Repayment status for last 6 months & Maximum credit line approved
- Maximum credit line approved ranges from 10,000 NT Dollars to 1 Million NT Dollar (NT stands for New Taiwan Dollars)

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Data Pre-processing: Before Modelling

D. Model Creation and Evaluation

E. Conclusion: Key Takeaways

Data Preparation & Cleaning

1. Formatting Inconsistent data types of columns:

- Values in each column were of object data type, which should have been either integer or float. Accordingly, values of respective columns were converted to int/float data type

2. Handling Missing Values:

- There were no values missing in any column

3. Handling Data Outliers

- There were many outliers in numerical features such age, maximum credit limit, paid amounts & bill amounts in last 6 months
- On exploring, the outliers were found to be natural and depict real-world trends
- Therefore, above outliers were not treated

4. Handling unknown classes of categorical Features:

a. **Education Level**:
  - Out of total 7 given categories, only 4 are documented
  - 1 is for Graduate School, 2 is for University level, 3 is for High School level, 4 is for Others
  - Undocumented categories (0, 5 & 6) have very few observations; to avoid overfitting these were merged with the given category "Others"

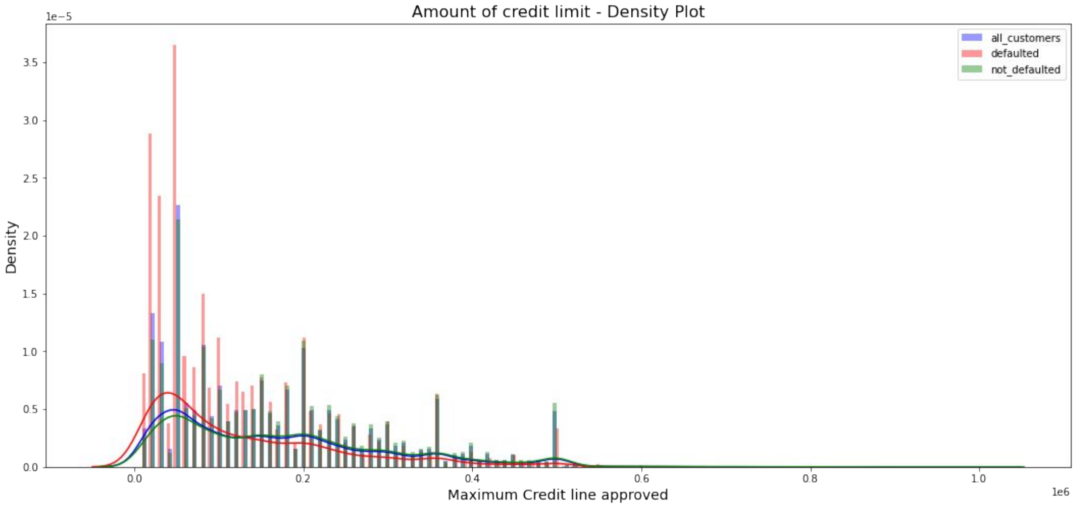
b. **Marital Status**:
  - Out of total 4 categories, 3 are documented
  - 1 is for Married, 2 is for Single, 3 is for Others
  - Undocumented category (0) has very few observations; to avoid overfitting it was merged with the given category "Others"

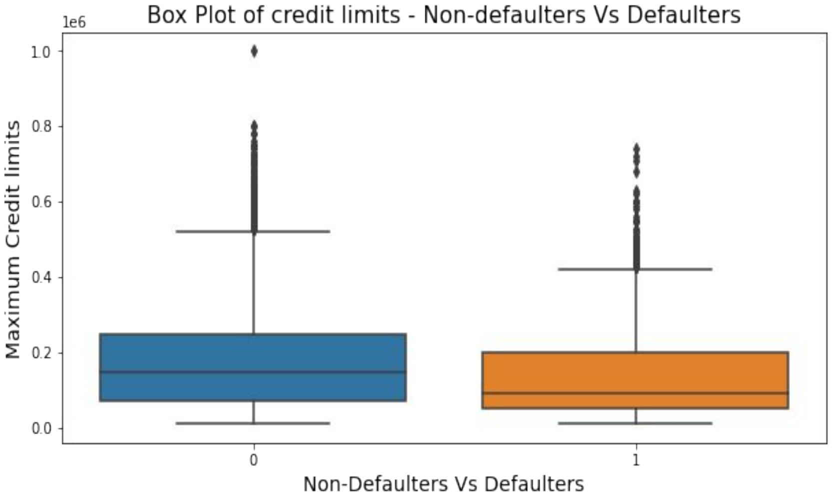
c. **Re-payment Status for last 6 months**:
  - Categories -2 and 0 are not documented
  - These categories have significant number of observations, hence were considered as given

Underlying Patterns and Trends

1. Maximum Credit Limit of Customer

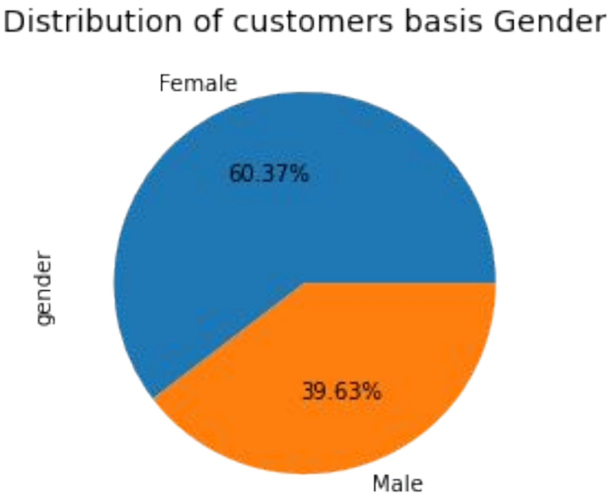
- Average of maximum credit limits approved for Defaulters is less than that of Non-defaulters
- This suggests that credit profiles of defaulters were certainly not as good as that of Non-defaulters from the beginning, and so issuing smaller credit lines to them was certainly a good decision

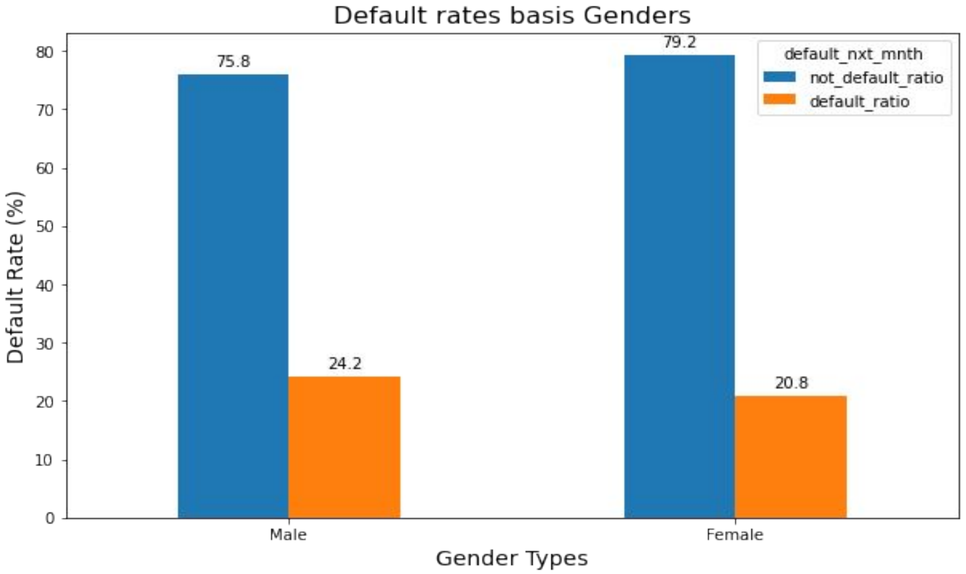




2. Male Vs Female Customers

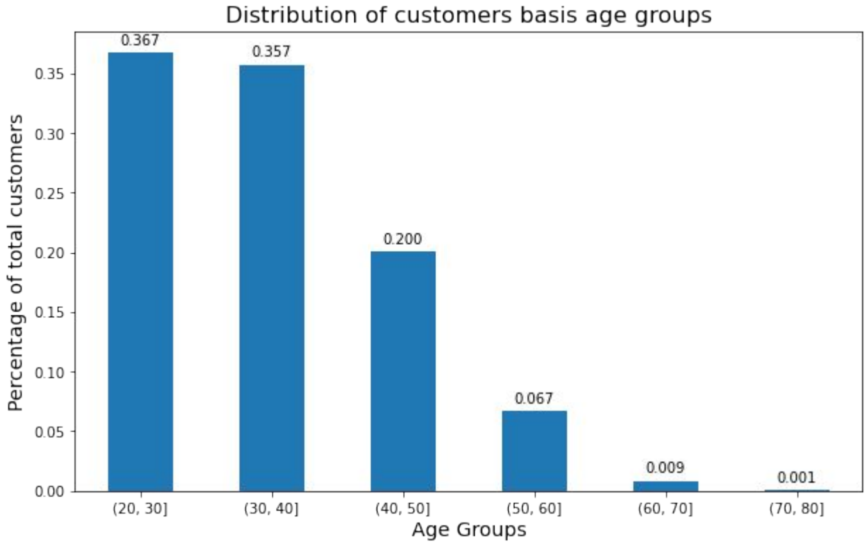
- Majority of customers are Females (~60%)
- Chances of Male customers (~24.2%) defaulting on their payments next month is higher than Female customers (~20.8%)

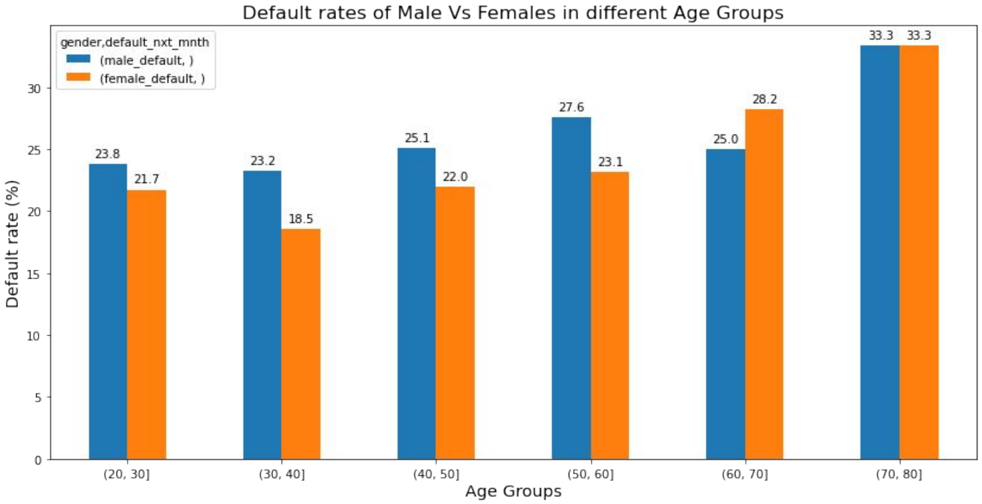




3. Age Groups of Customers

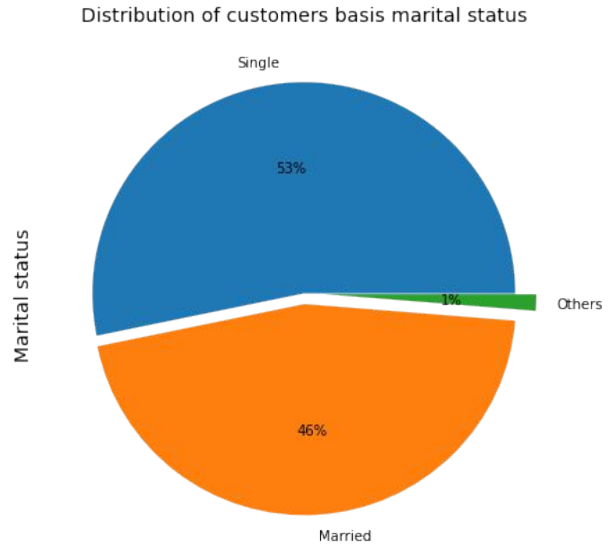
- Majority are of 21 to 40 years age group (~72%), least share of customers belong to age group 61 to 80 years (~1%)
- Customers of age 31 to 40 years are least likely to default
- Default rate gradually increases for age groups 40 years onwards and is maximum for 71 to 80 years

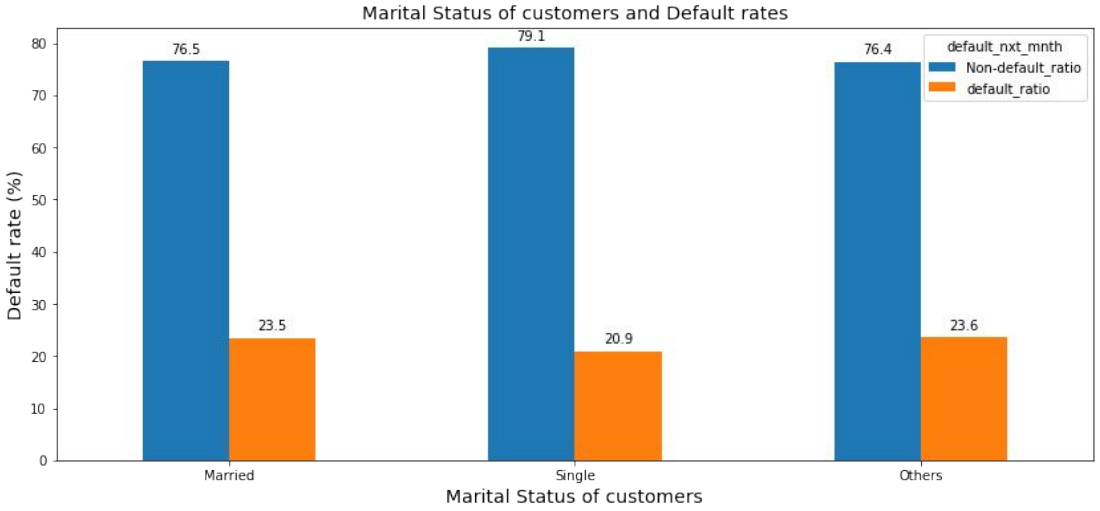




4. Marital Status of Customers

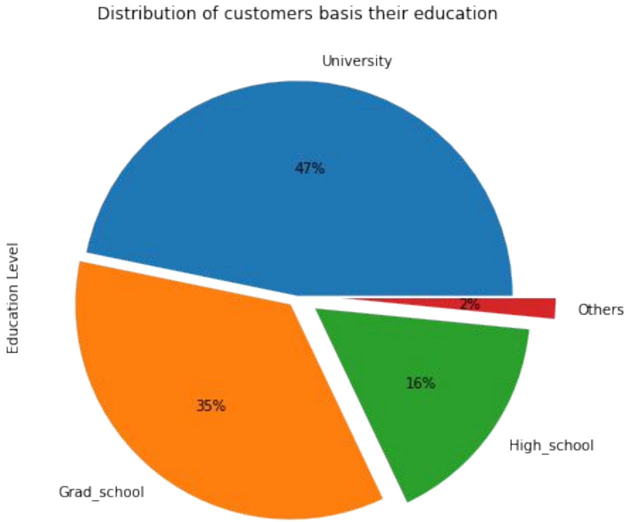
- Majority of customers (~53%) are single, followed by customers who are married (~46%)
- Customers of Other marital status are most likely to default (~23.6%), shortly followed by customers who are married (~23.5%)

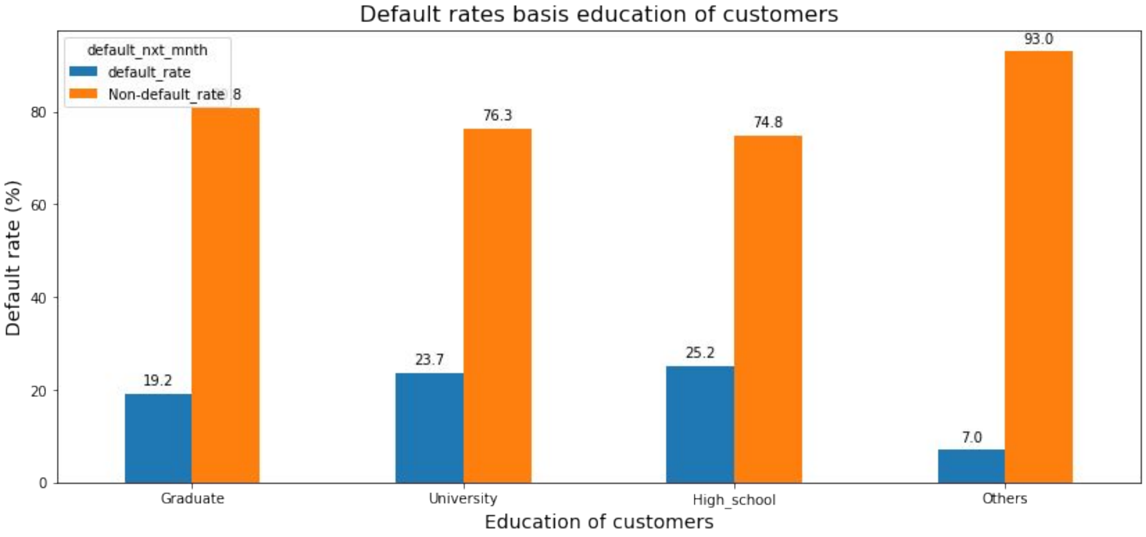




5. Education of Customers

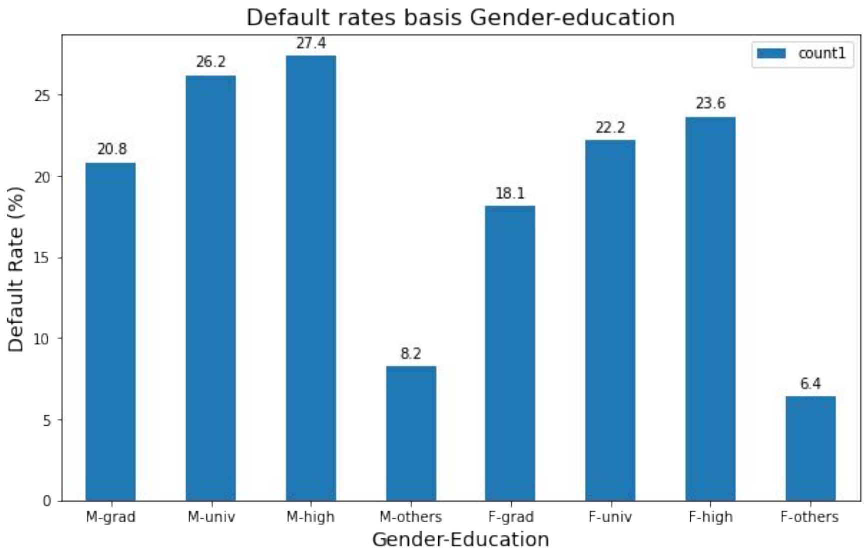
- Majority of customers have University level education (~47%), followed by Graduate school level (~35%), Only 16% have High-school level education
- Default percentage rate is highest for customers with education of high-school level (~25%), shortly followed by University level education (~24%)

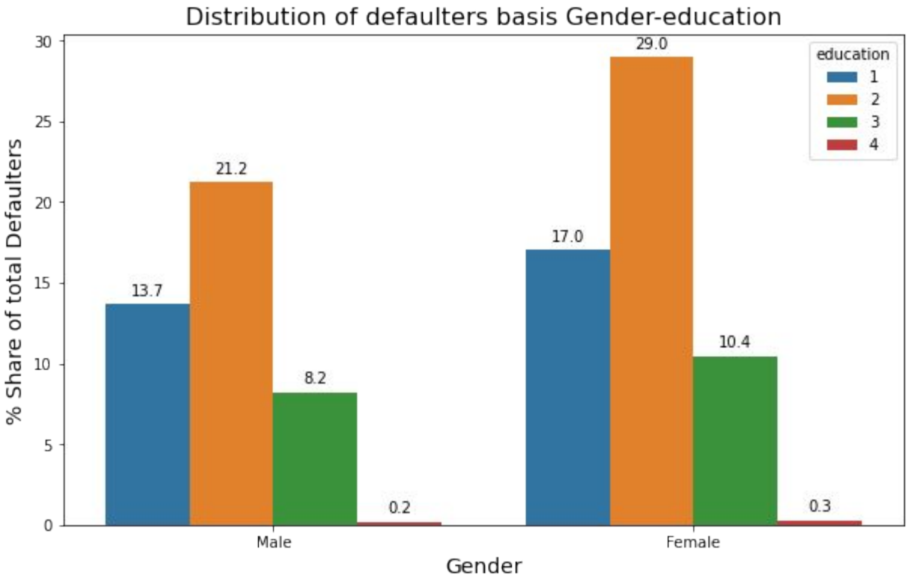




6. Gender and Education

- Basis gender and education, Males with high school level education have highest chance of defaulting (~27.4%), followed by University males (~26.2%)
- Among defaulters, “Females with university education” have the highest share (~29%), followed by “Males with university education” (~21.2%)

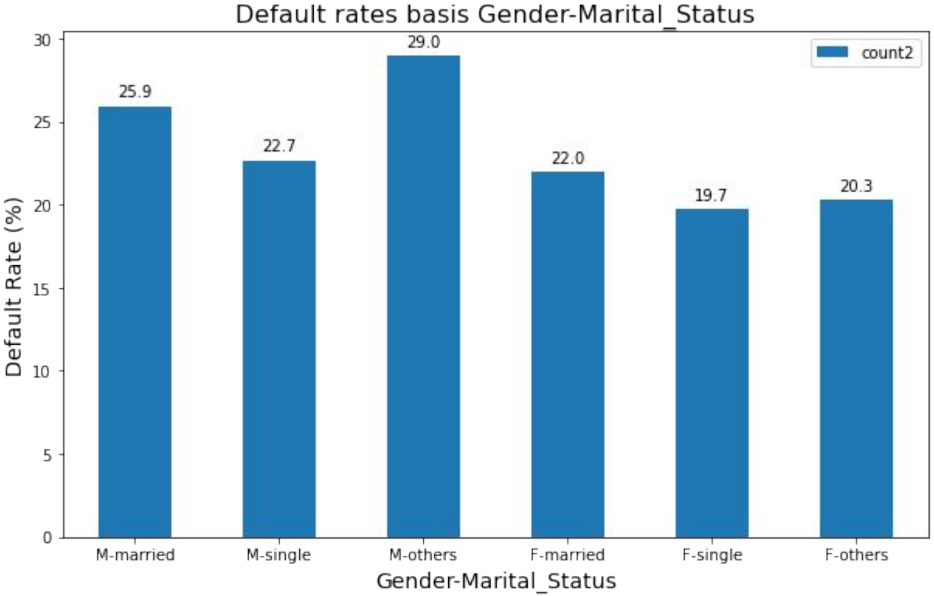


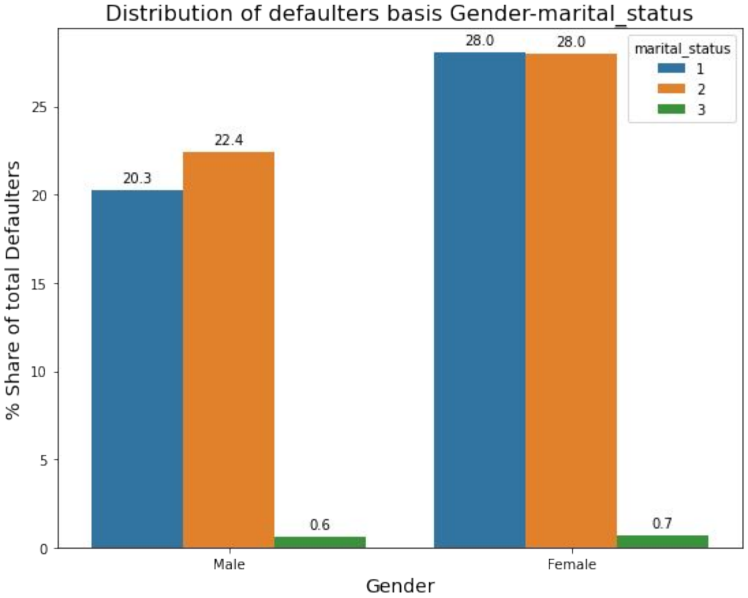


7. Gender and Marital Status

- Not considering others category, Married males have the highest chance of defaulting (~25.9%), followed by Single males (~22.7%)

- Among defaulters, Married Females have the highest share (~28%), shortly followed by Single Females





Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Data Pre-processing: Before Modelling

D. Model Creation and Evaluation

E. Conclusion: Key Takeaways

Feature Engineering

A. List of Features extracted:

- Age groups: Age of customers has been divided into 6 groups
- Average Monthly spends of a customer: This has been calculated by taking difference between Bill amounts of next & current months, and adding amount paid in next month
- Utilization rate: This is basically the share of maximum credit limit used by a customer each month
- Speed of change in utilization rate in latest month September: This has been calculated as ratio of utilization rate in September to the average utilization rate of previous 5 months

B. Encoding of Categorical Features

- Gender, Education and Marital Status were encoded

Feature Selection

1. **Correlation Matrix**:
   - It quantifies the linear relationship between two features; basis this identified highly correlated features & only kept one of them
2. **SelectKBest (Filter)** class of sklearn library:
   - It selects features according to the k highest scores basis F-statistics calculated for each input variable with the target
   - score_func used is “f_classif”
3. **Sequential Forward Selection (Wrapper)**:
   - The searching algorithm adds feature sequentially to an empty set of features until the addition of extra features does not reduce the criterion
   - RandomForestClassifier algorithm with “roc_auc” as criterion was used
4. **Recursive Feature Elimination (Wrapper)**:
   - It fits a model and removes the weakest features one-by-one basis weights assigned, until the optimal number is reached
   - RandomForestClassifier algorithm was used

Before Model Creation

- Selecting predictor variables
- Stratified Train-Test splitting: Test size of 20% data was chosen
- Feature scaling:
  - MinMaxScaler was used for normalization
  - To avoid data leakage, scaler transformation was firstly fitted on training data and then based on the statistical parameters learned from training data, the same transformation was applied on test dataset.
- Handling Imbalanced dataset (~22% defaults & ~78% non-defaults)
- Choosing Evaluation Metrics

Handling Imbalanced Classes in Dataset

1. **TomekLinks Under-sampling:**
   - It removes Tomek links, which are points in the dataset whose nearest neighbor is a member of a different class
   - This includes outlier points embedded in a point cloud from another class, and boundary points in regions
2. **CentroidClusters Under-sampling:**
   - It under samples the majority class by replacing a cluster of majority samples by the cluster centroid of a K-Means algorithm
3. **SMOTE Oversampling:**
   - It starts by picking random points from minority class and computing the K-nearest neighbors of this point
   - After this, it generates synthetic points of minority class along the lines joining chosen point and its K-nearest points
4. **SMOTE-TomekLinks:** Over-sampling followed by Under-sampling

Evaluation Metrics

1. **Recall**:
   - It aligns with our main goal, “Minimize Risk”, by correctly identifying as many defaulters as possible, and thus reducing False negatives
2. **Precision**:
   - It aligns with the other goal “Maximize Business”, by not mis-identifying a good customer as defaulter and thus reducing False Positives
3. **AUC-ROC score** (Higher the AUC, the better):
   - It is the measure of the ability of a classifier to distinguish between the positive and negative classes
4. **Brier Score** (Lower the score, the better):
   - It evaluates the accuracy of probabilistic predictions and used to check the goodness of a predicted probability score
5. **KS-chart**:
   - KS statistic for two classes is simply the highest distance between their respective CDFs (evaluates model’s ability to distinguish)

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Data Pre-processing: Before Modelling

D. Model Creation and Evaluation

E. Conclusion: Key Takeaways

Baseline Models: Logistic Regression

S. No.	Classification Models	Test Accuracy	Test F-1 score	Test Precision	Test Recall	Test ROC-AUC Score	Test Brier's score
1	LogR using Imbalanced dataset	**0.69**	**0.46**	**0.38**	**0.61**	**0.71**	0.21
2	**LogR using TomekLinks**	0.68	**0.46**	0.37	**0.62**	**0.71**	0.21
3	LogR using Centroid-Clusters dataset	0.48	0.40	0.27	0.80	0.67	0.34
4	LogR using SMOTE	0.70	0.43	0.37	0.52	0.67	0.20
5	LogR using SMOTE-TomekLinks dataset	0.70	0.44	0.38	0.53	0.67	0.20

- Basis Recall, Logistic Regression model using TomekLinks under-sampled dataset is the best performing model, with acceptable F-1, Precision, AUC-ROC and Brier's scores
- Under-sampled dataset with Centroid-Clusters has the least accuracy, F-1 and Precision, but has the highest recall. The reason could be loss of important information during under-sampling

Baseline Models: Random Forest Classifier

AI

S. No.	Classification Models	Test Accuracy	Test F-1 score	Test Precision Score	Test Recall	Test ROC-AUC Score	Test Brier's score
1	RF using Imbalanced dataset	0.81	0.43	**0.65**	0.32	**0.76**	**0.14**
2	RF using TomekLinks	**0.81**	0.46	0.63	0.37	0.76	0.14
3	RF using Centroid-Clusters dataset	0.50	0.44	0.29	0.87	0.73	0.37
4	RF using SMOTE	0.79	0.50	0.52	**0.49**	**0.75**	0.16
5	**RF using SMOTE-TomekLinks dataset**	0.79	**0.51**	0.52	**0.51**	**0.75**	0.16

- Basis Recall, Random Forest model using combined re-sampling (SMOTE+TomekLinks) dataset is the best performing model, with acceptable F-1, Precision, AUC-ROC and Brier's scores
- Random Forest model using SMOTE over-sampled dataset has the second highest Recall with acceptable F-1, precision and AUC-ROC scores

Baseline Models: XGBoost Classifier

AI

S. No.	Classification Models	Test Accuracy	Test F-1 score	Test Precision	Test Recall	Test ROC-AUC Score	Test Brier's score
1	**XGB using Imbalanced dataset**	0.81	0.44	**0.62**	0.35	**0.76**	**0.14**
2	XGB using TomekLinks	**0.81**	0.47	0.59	0.38	0.76	0.14
3	XGB using Centroid-Clusters dataset	0.45	0.42	0.27	0.89	0.73	0.48
4	XGB using SMOTE	0.77	0.47	0.47	**0.47**	**0.73**	0.16
5	**XGB using SMOTE-TomekLinks dataset**	0.76	**0.47**	0.46	**0.48**	**0.73**	0.17

- Basis Recall, XGBoost classifier model using combined re-sampling (SMOTE+ TomekLinks) is the best performing model, with acceptable Precision, F-1, AUC-ROC and Brier's scores
- Second best performing model is using SMOTE over-sampled dataset with the second highest Recall with acceptable F-1, precision AUC-ROC and Brier's scores

Baseline Models: SVM Classifier

S. No.	Classification Models	Test Accuracy	Test F-1 score	Test Precision	Test Recall	Test ROC-AUC Score	Test Brier's score
1	**SVM using Imbalanced dataset**	**0.77**	**0.51**	**0.49**	**0.54**	**0.74**	**0.14**
2	**SVM using TomekLinks**	**0.77**	**0.51**	**0.49**	**0.54**	**0.74**	**0.14**
3	SVM using Centroid-Clusters dataset	0.48	0.42	0.28	0.84	0.71	0.39
4	SVM using SMOTE	0.72	0.46	0.40	0.53	0.70	0.19
5	SVM using SMOTE-TomekLinks dataset	0.72	0.46	0.40	**0.54**	0.70	0.20

- SVM classifier models using Imbalanced dataset and TomekLinks under-sampled dataset are the best performing models, with the highest Recall, F-1 score, AUC-ROC score , Precision and lowest Brier's score
- Second best performing model is using SMOTE-TomekLinks re-sampled dataset

Models After Hyper-parameters Tuning

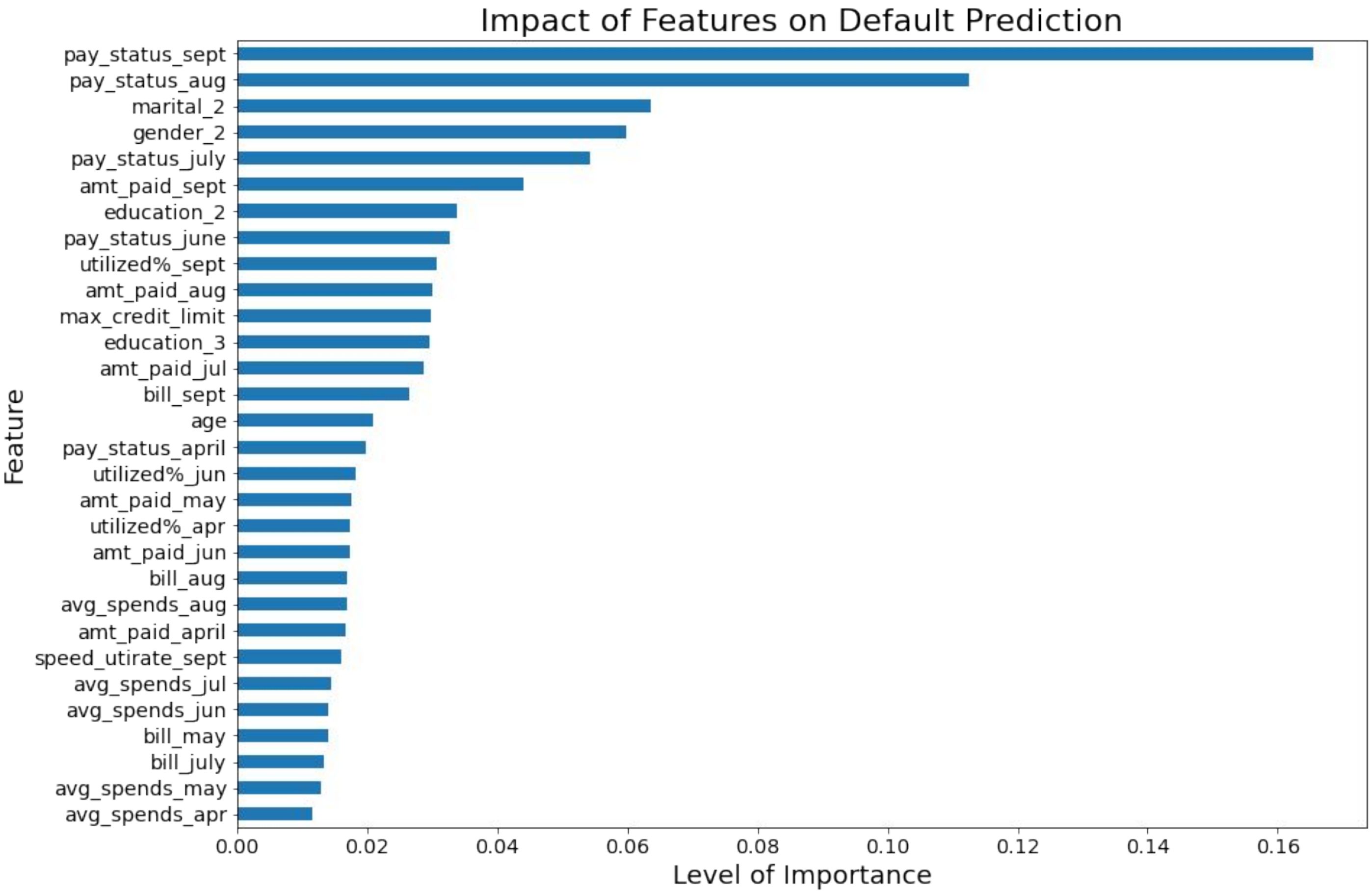
AI

S. No.	Classification Models	Test Accuracy	Test F-1 score	Test Precision	Test Recall	Test ROC-AUC Score	Test Brier's score
1	LogR using Imbalanced dataset	0.69	0.46	**0.38**	**0.61**	0.71	0.21
2	LogR using TomekLinks	0.68	0.46	**0.37**	**0.62**	0.71	0.21
3	**RF using SMOTE**	**0.77**	0.51	0.48	**0.55**	**0.76**	0.17
4	RF using SMOTE-TomekLinks dataset	0.77	0.51	0.48	0.54	0.76	0.16
5	XGB using SMOTE	**0.78**	0.49	**0.50**	0.49	0.75	0.16
6	**XGB using SMOTE-TomekLinks dataset**	0.76	0.50	0.47	**0.55**	0.75	0.17
7	**SVM using TomekLinks**	**0.77**	**0.52**	0.48	**0.57**	0.74	**0.14**
8	SVM using SMOTE-TomekLinks dataset	0.73	0.47	0.42	0.54	0.72	0.19

RF-SMOTE: Features and their Coefficients

- Top 5 predictors are as:

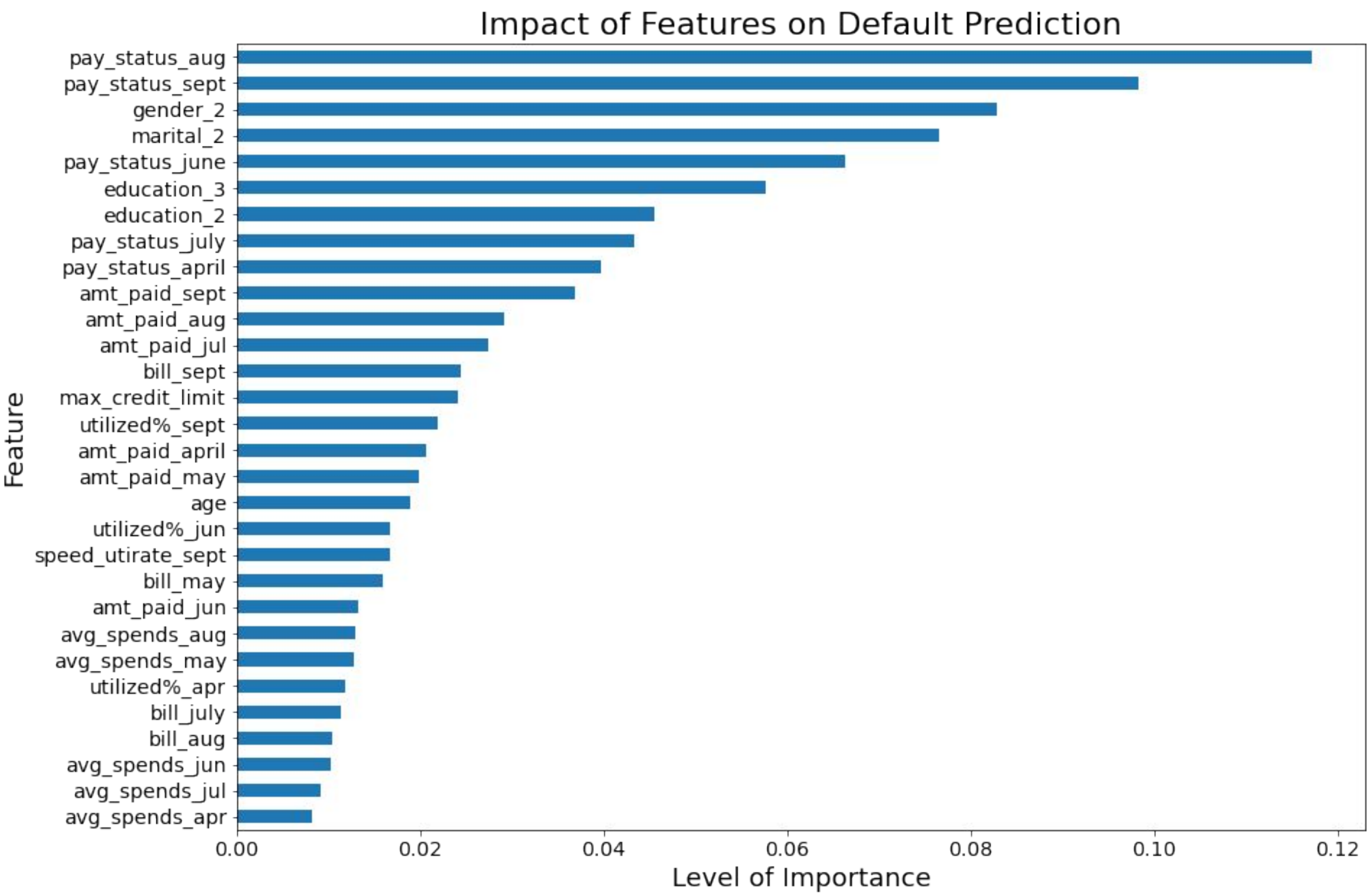
1. Re-payment status of September
2. Re-payment status of August
3. Marital Status “Single”
4. Gender
5. Re-payment status of July



XGB (SMOTE-TomekLinks): Feature Importance

- Top 5 predictors are as:

1. Re-payment status of August
2. Re-payment status of September
3. Gender
4. Marital Status "Single"
5. Re-payment status of June



Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Data Pre-processing: Before Modelling

D. Model Creation and Evaluation

E. Conclusion: Key Takeaways

Understanding the Credit card Owners:

- Majority of credit card owners are Females (~60%)
- Majority of card owners have University level education (~47%), followed by Graduate school level education (~ 35%)
- Only 16% have High-school level education
- Majority of card owners (~53%) are single, followed by married (~46%)
- Majority of card owners (~72%) are within 21 years to 40 years age group
- As per given dataset, ~22% customers have defaulted on payment next month

Defaulters and their Demographics:

- Chances of Males defaulting on their payments next month is higher than that of Females
- As education level increases (i.e., high school to university to graduate school), default rate decreases
- Basis Marital status, Chances of defaulting is highest for Married customers
- Customers of age between 31 to 40 years are least likely to default, followed by 21 to 30 years group
- Females of age between 31 to 40 years are least likely to default next month
- Chances of a "Married Male" defaulting is the highest, while that of "Single Female" is the least
- Male customer with High-school level education, has the highest chance of defaulting
- Females with graduate education are least likely to default next month

Defaulters and their Financial Behaviours:

- Average of maximum credit limit approved for Defaulters is less than that of Non-defaulters, suggesting that defaulters credit profile was certainly not as good as Non-defaulters even from the beginning, and as a result they were issued smaller credit lines
- Customers with payments pending for more than 1 month, have higher chances of defaulting
- Defaulter's overall pay-down ratio kept on decreasing each successive month, while for non-defaulters, the ratio has an overall increasing trend (Pay down ratio is the ratio of total amount paid by customer to the total Bill amount)
- Defaulter's utilization rate increased significantly in latest month of September, while for Non-defaulters the utilization rate decreased (Utilization rate is the ratio of the Bill amount to the maximum credit limit)

Top 3 Best Performing models:

- Considering Recall metric with utmost importance, followed by Precision and F-1 scores, found the following top 3 models:

  1. **SVM model built using TomekLinks** under-sampled dataset is the best performing model, with good Recall (0.57), precision (0.48), F-1 score (0.52), AUC-ROC (0.74) and least Brier score (0.14)

  2. Second best performing model is **Random Forest classifier built using SMOTE** over-sampled dataset, with good Recall (0.55), precision (0.48), F-1 score (0.51), AUC-ROC (0.76) and Brier score (0.17)

  3. Third best performing model is **XGBoost classifier built using SMOTE-TomekLinks** combined re-sampled dataset, with good Recall (0.55), precision (0.47), F-1 score (0.50), AUC-ROC (0.75) and Brier score (0.17)

END OF PRESENTATION
THANKS