- **Risk Analysis**

In
Banking, Financial Services and Insurance

Agenda

- Introduction
- Data Overview
- Data Cleaning and Pre – processing steps
- Exploratory Data Analysis on Application Data Table
- Exploratory Data Analysis on Previous Application Data Table
- Dashboards from PowerBI
- Predictive Modelling

Introduction

Business Objective :

- The case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Data Overview

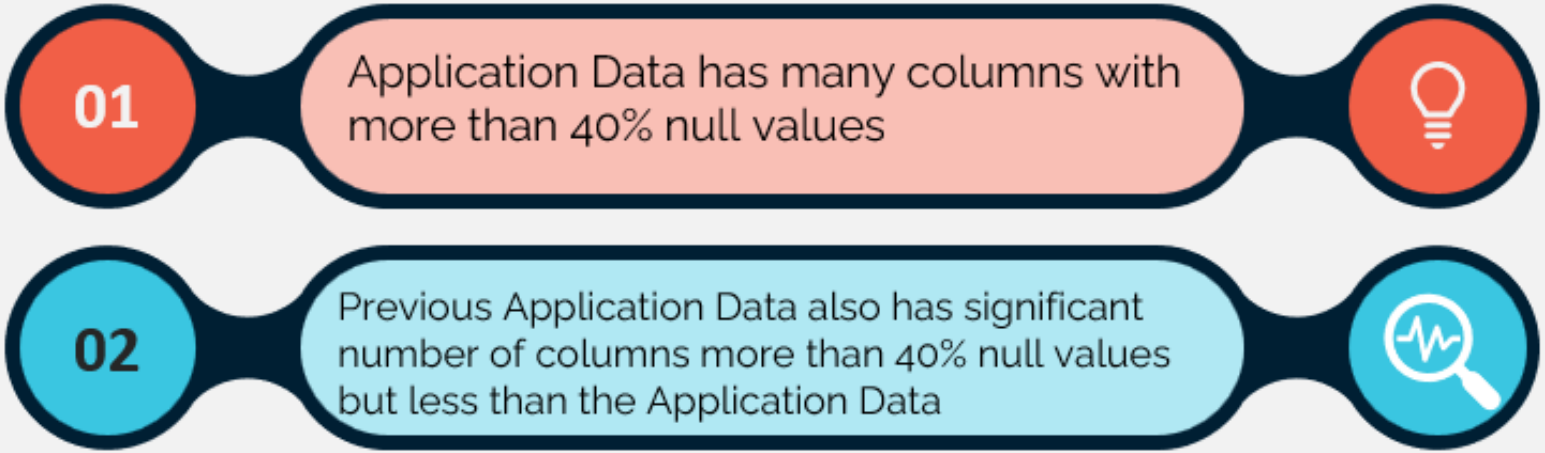
Application Table

- Rows: 3,07,511
- Columns: 122
- Primary Key: SK_ID_CURR
- Description: Contains detailed information on current loan applications, including:
- Loan ID, Contract Type, Credit Amount
- Applicant Demographics: Gender, Income Level, Education Level, Family Status
- Regional Population, Contact Information
- Real Estate Ownership, Submitted Documents
- TARGET: Indicator of potential default risk

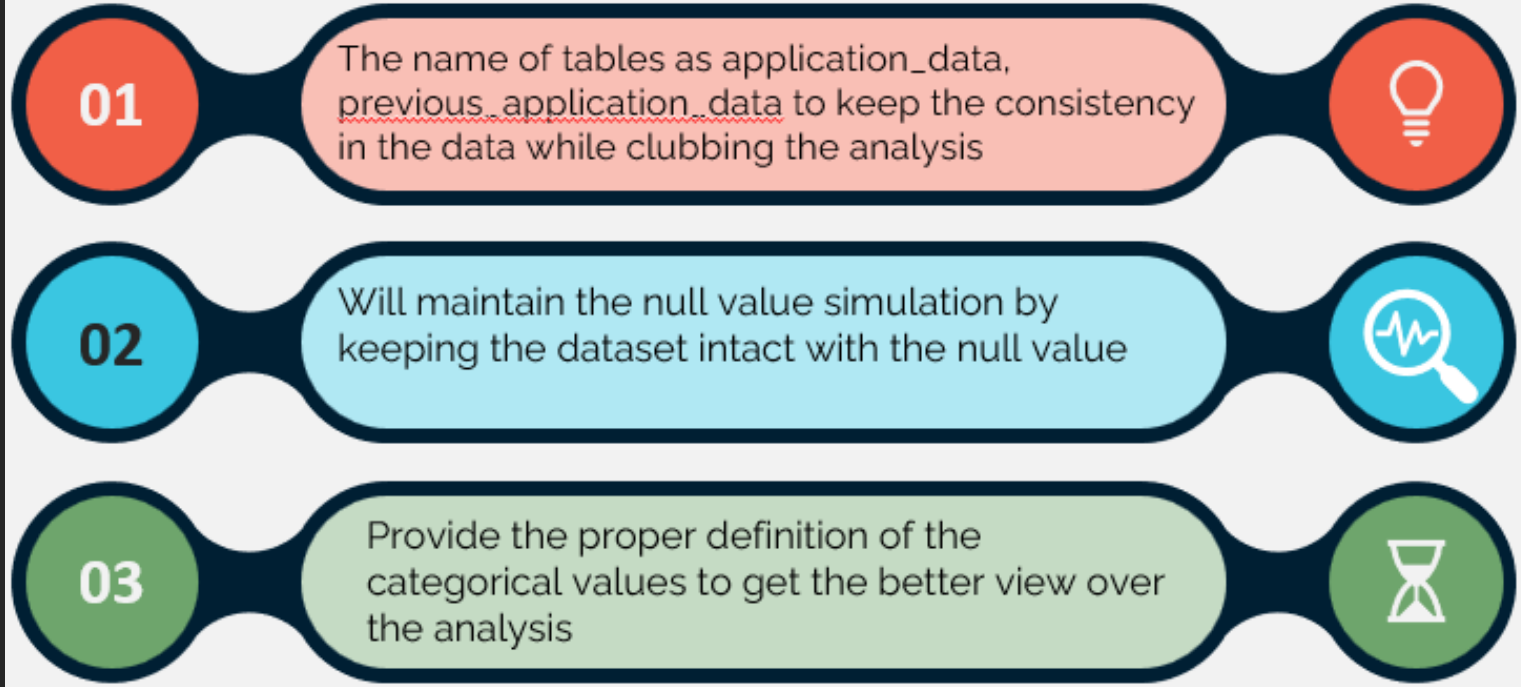
Previous Application Table

- Rows: 16,70,214
- Columns: 37
- Primary Key: SK_ID_PREV
- Description: Details of all previous home loan applications, including:
- Loan ID, Contract Type, Credit Amount
- Status of Previous Loan
- Decision Timing Relative to Current Application
- Payment Method, Portfolio Type, Yield Type

Observations made during Data Auditing:



Data Cleaning and Pre-processing



Exploratory Data Analysis

- **Understanding The Demographics**

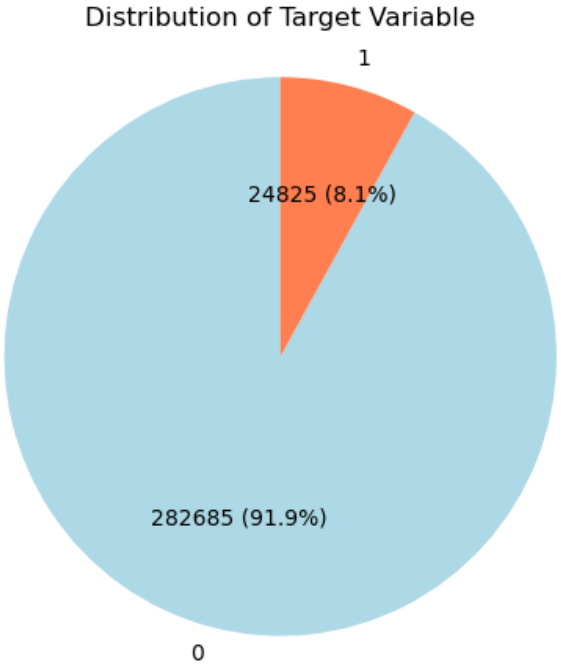
Male Vs Female
Ratio : 1.93

Minimum Income : INR 25K
Max Income : INR 11.7Cr
Average Income : INR 1.68L

Age distribution
Minimum age : 20
Maximum age: 70
35- 40 years has max applicants

Exploratory Data Analysis

Distribution Of Target Variable

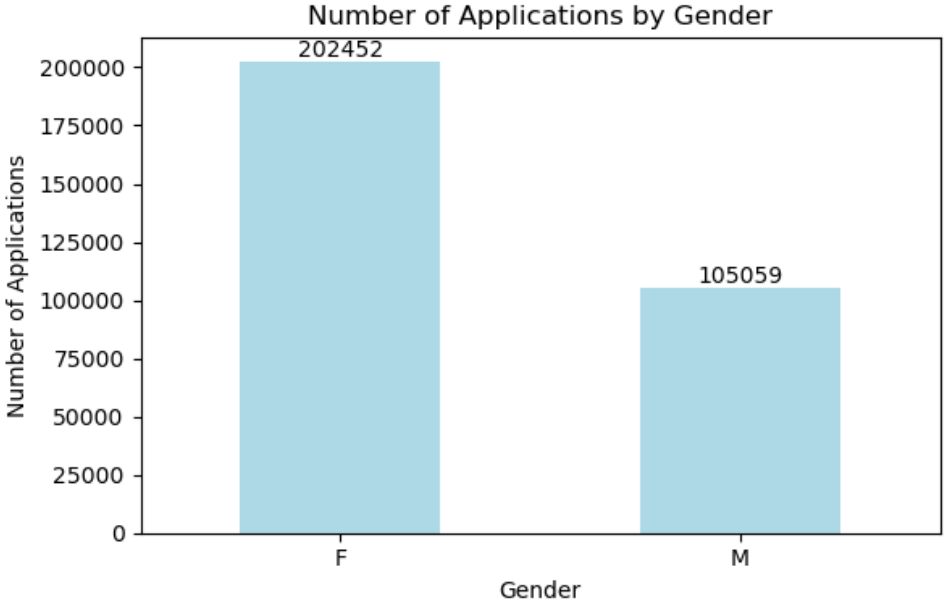


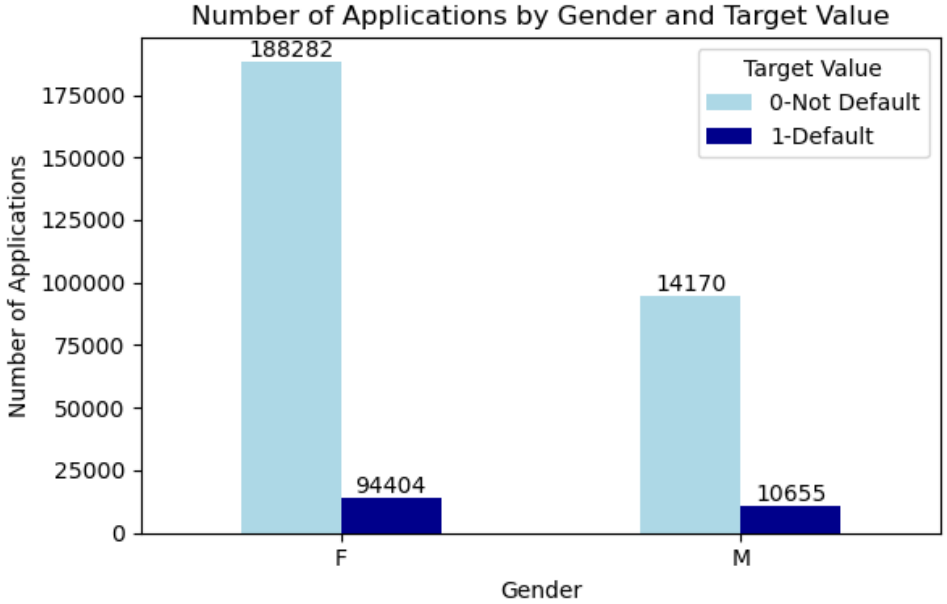
Insight :

- The majority of the individuals in the dataset are non-defaulters, making up approximately **91.9%** (282,686 out of 307,511) of the total sample
- Defaulters are only about **8.1%** of the sample
- The relatively low percentage of defaulters suggests that most individuals are managing their loans well.

Exploratory Data Analysis

Analyzing The Gender Ratio



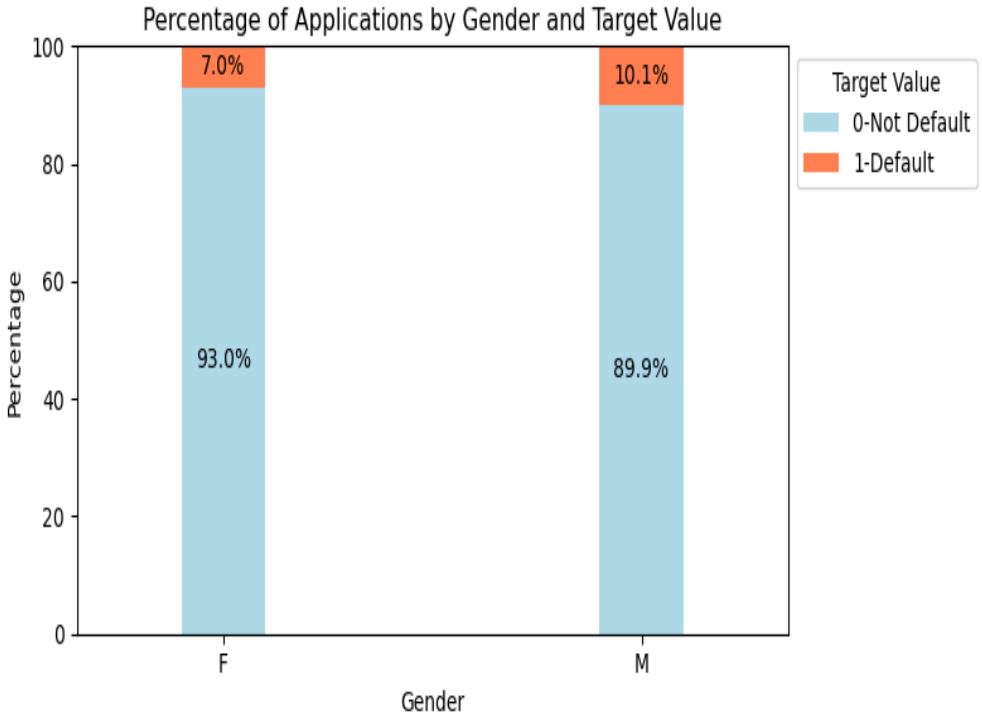


Insight :

- **Females** account for a larger share of applications **(65.8%)** compared to **males (34.2%)**.
- This suggests that female applicants are more prevalent in the dataset.

Exploratory Data Analysis

Analyzing The Gender Ratio



Insight :

Application Share:

- Females make up **65.8%** of applications.
- Males account for **34.2%** of applications.

Non-Default Rates:

- **93.0%** of female applicants are non-defaulters (188,282 out of 202,452).
- **89.9%** of male applicants are non-defaulters (94,404 out of 105,059).

Default Rates:

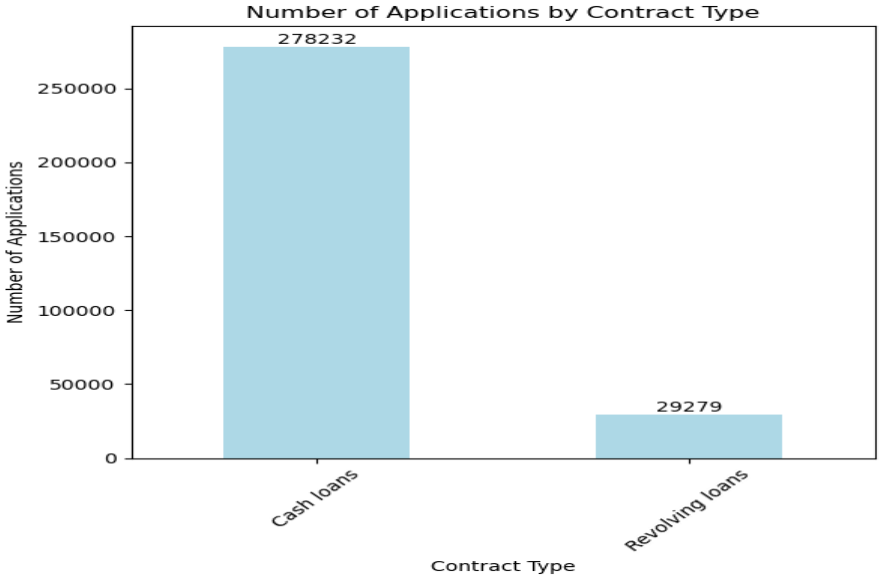
- Female default rate is **7.0%** (14,170 out of 202,452).
- Male default rate is **10.1%** (10,655 out of 105,059).

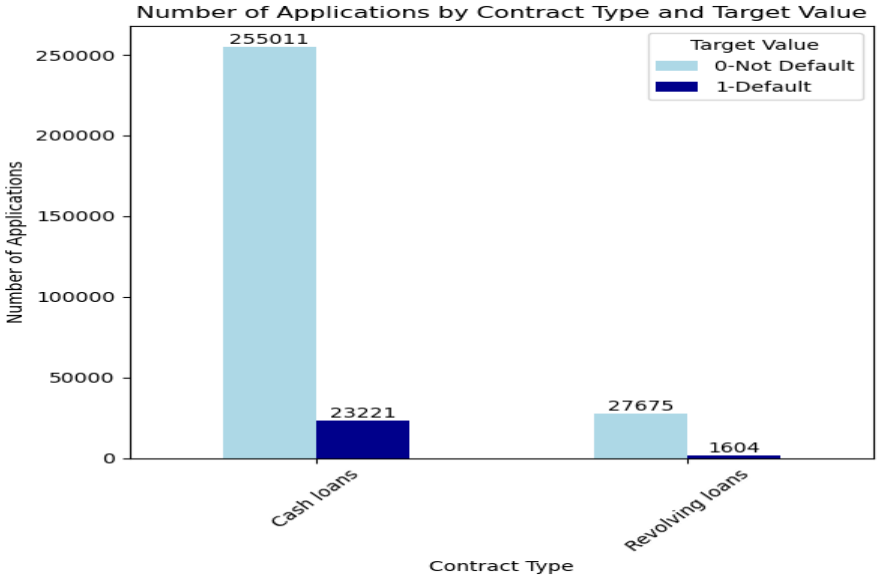
Male applicants show a higher likelihood of default compared to female applicants.

Image source: Icons8

Exploratory Data Analysis

Analyzing The Contract Type





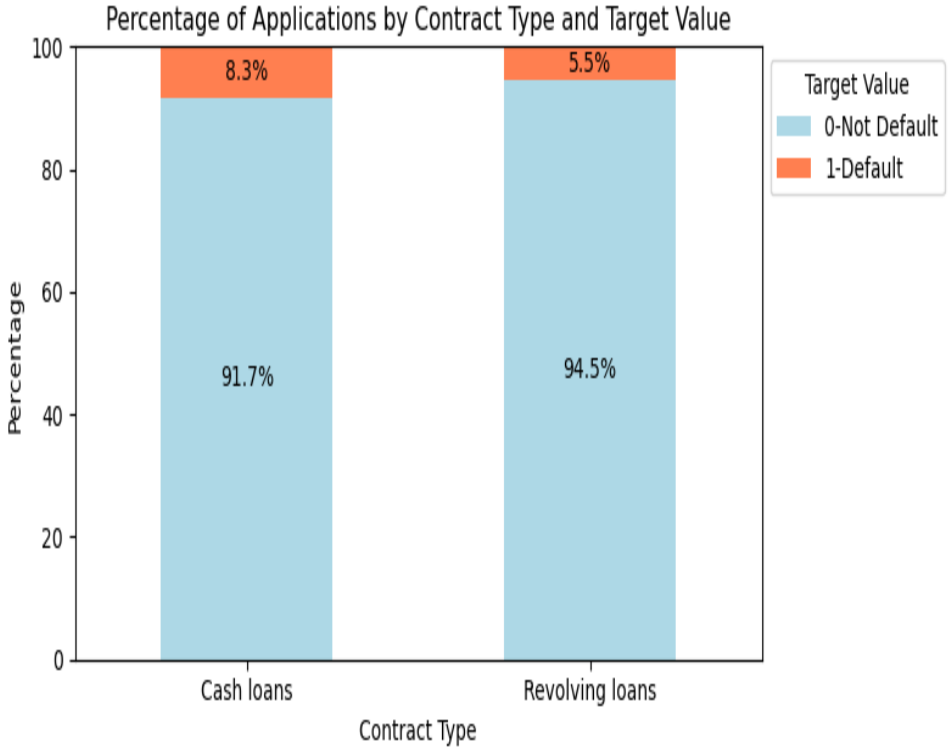
Insight :

- **Cash loans** are the dominant type of loan, making up approximately **90.5%** of total applications.
- **Revolving loans** constitute around **9.5%**.

Image source: Icons8

Exploratory Data Analysis

Analyzing The Contract Type



Insight :

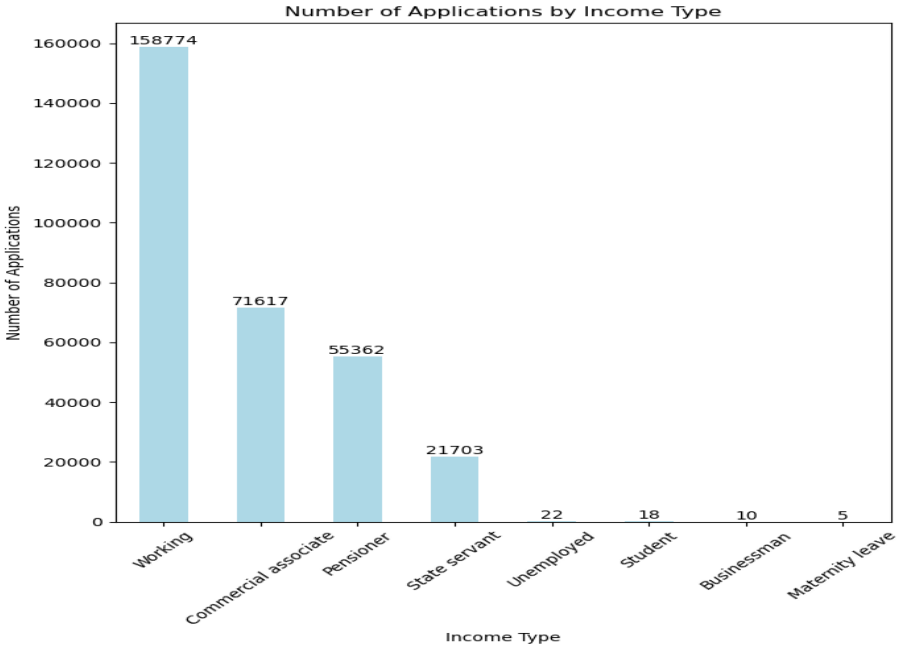
Loan Distribution:
•Cash loans make up **90.5%** of total applications.
•Revolving loans constitute **9.5%** of total applications.

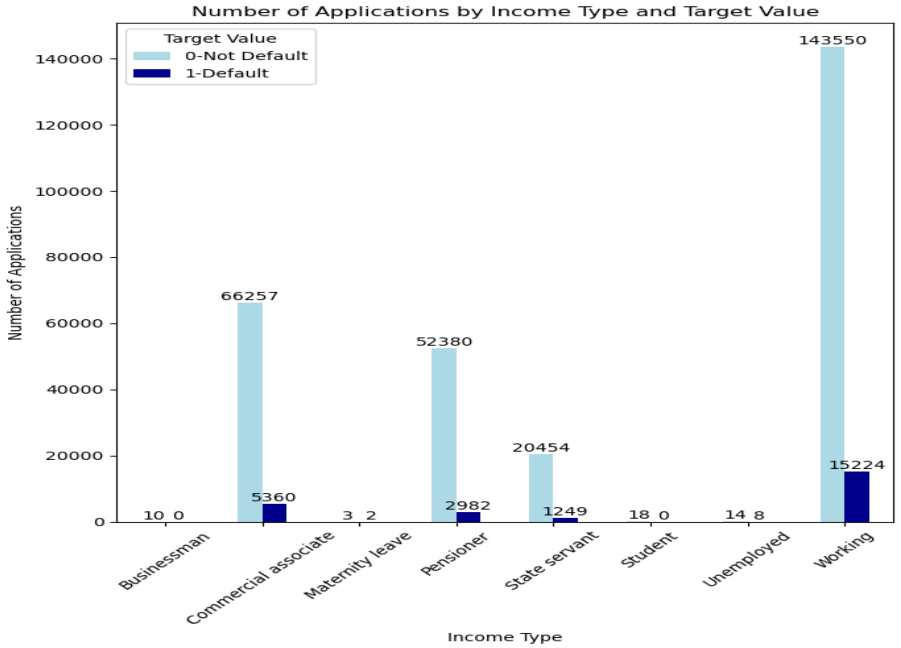
Default Rates by Contract Type:
•Cash loans have a default rate of **8.3%** (23,221 out of 278,232).
•Revolving loans have a lower default rate of **5.5%** (1,604 out of 29,279).

Non-Default Rates by Contract Type:
•Cash loans have a non-default rate of **91.7%** (255,011 out of 278,232).
•Revolving loans have a higher non-default rate of **94.5%** (27,675 out of 29,279).

Exploratory Data Analysis

Let's Look At The Income Type





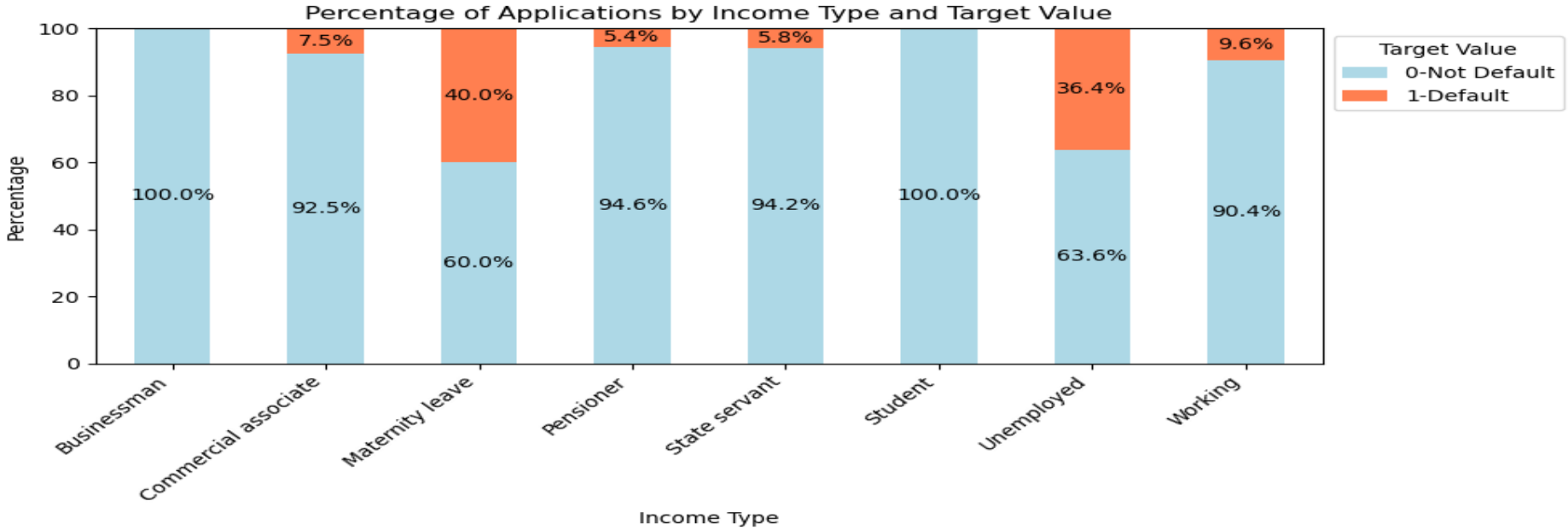
Insight :

- The **majority** of applications come from individuals classified as **"Working"** (approximately **50.5%**), followed by "Commercial Associate" (about 23.5%), and **"Pensioner"** (around **18%**). Other categories contribute very few applications.

Image source: icons8

Exploratory Data Analysis

Let's look at the income type

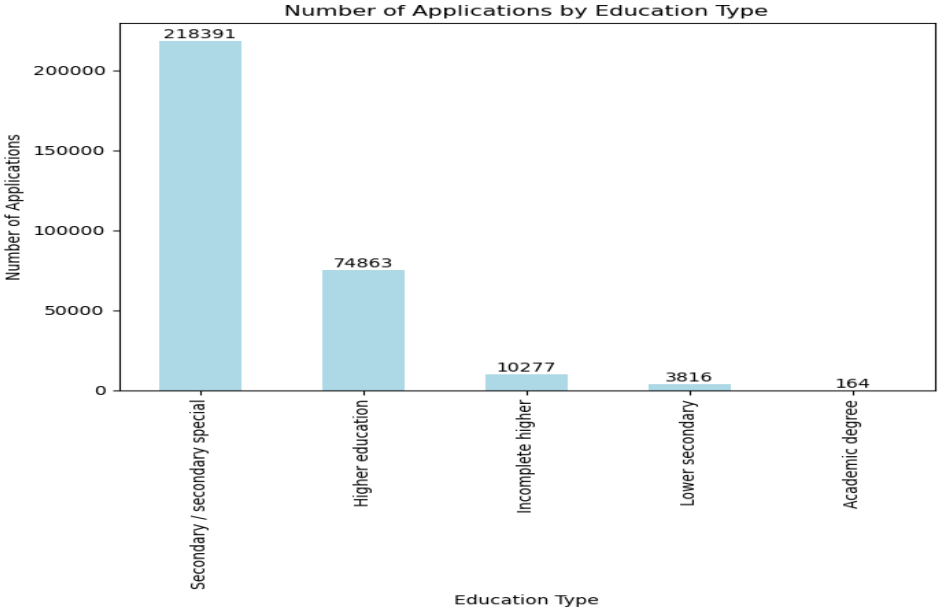


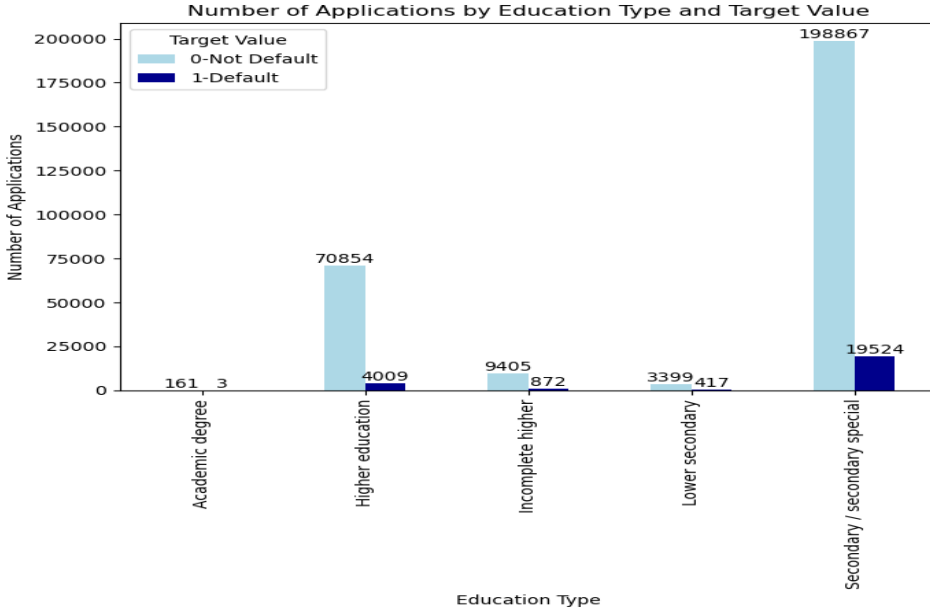
Insight : Most applications are from "Working" individuals (50.5%), followed by "Commercial Associate" (23.5%) and "Pensioner" (18%). The highest default rates are for **"Unemployed" (36.4%)** and **"Maternity Leave" (40%)**. "Working" applicants have a 9.6% default rate, indicating stable employment leads to better repayment.

Image source: icons8

Exploratory Data Analysis

Analyzing the Education Type





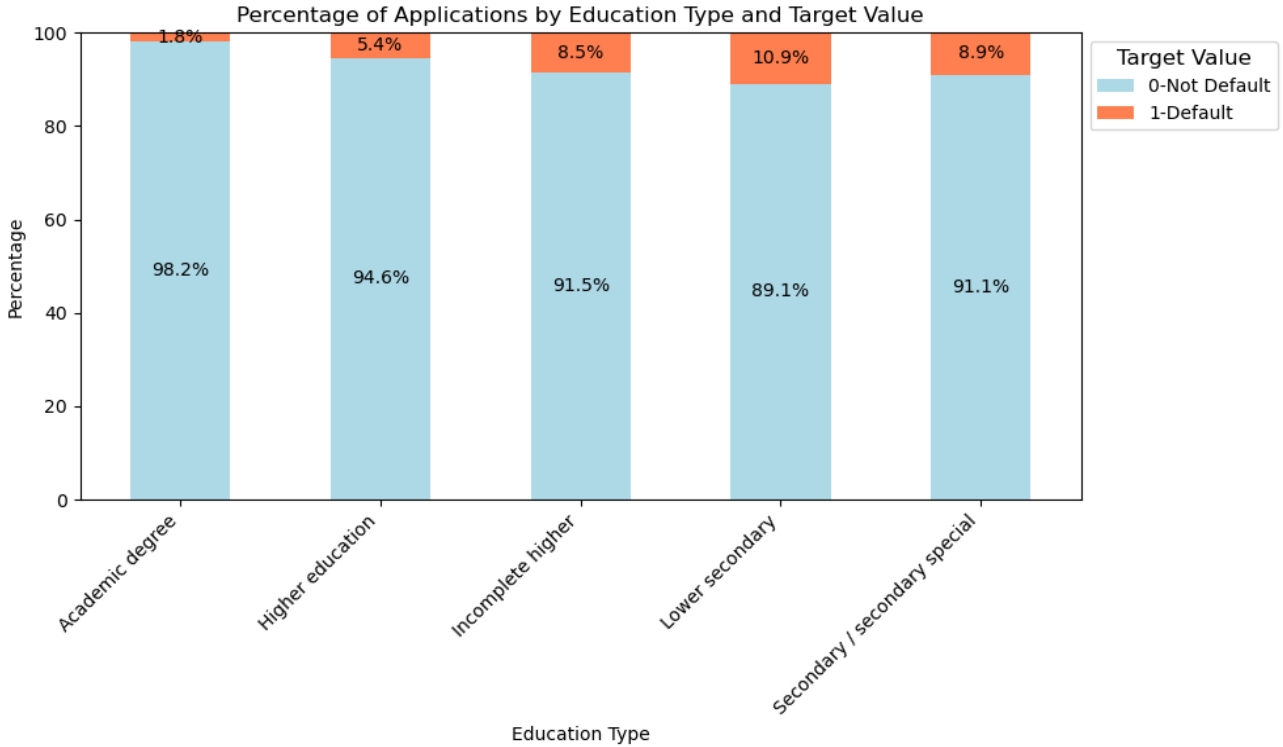
Insight:

- Most applications come from individuals with **Secondary/ Secondary Special education (218,391)**,
- followed by Higher Education (74,863). Those with an **Academic Degree have a low default rate (3 out of 161)**.

Image source: Icons8

Exploratory Data Analysis

Analyzing the Education Type



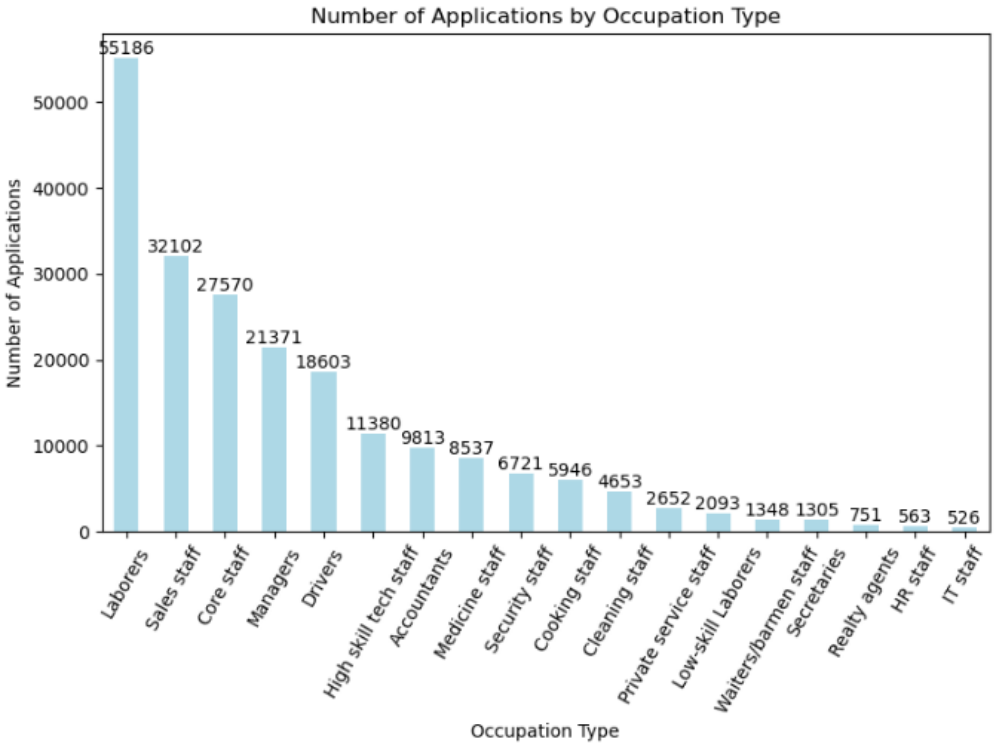
Insight:

- The Secondary group has 198,867 non-defaulters and 19,524 defaulters (8.9%),
- Lower Secondary education shows higher risk with 3,399 non-defaulters and 417 defaulters (11.0%).

Image source: Icons8

Exploratory Data Analysis

Analyzing the Occupation Type

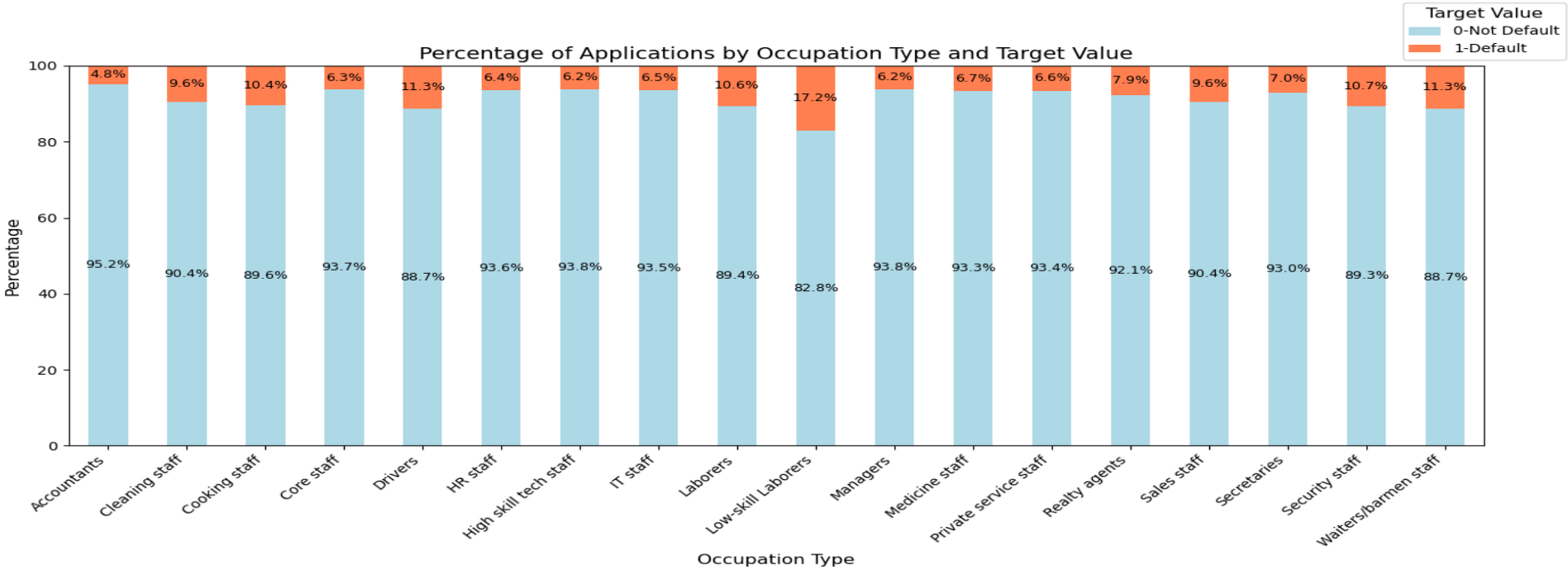


Insight :

- **Laborers:** Largest group, ~22.4% of total applications
- **Other Significant Segments:** Sales Staff and Core Staff Indicates strong interest among lower and middle-skilled workers
- **Managers:** Notable application numbers
- **High-Skill Roles (e.g., IT Staff, HR Staff):** Less represented, possibly due to more stable financial situations

Image source: Icons8

Exploratory Data Analysis ➤ Analyzing the Occupation Type

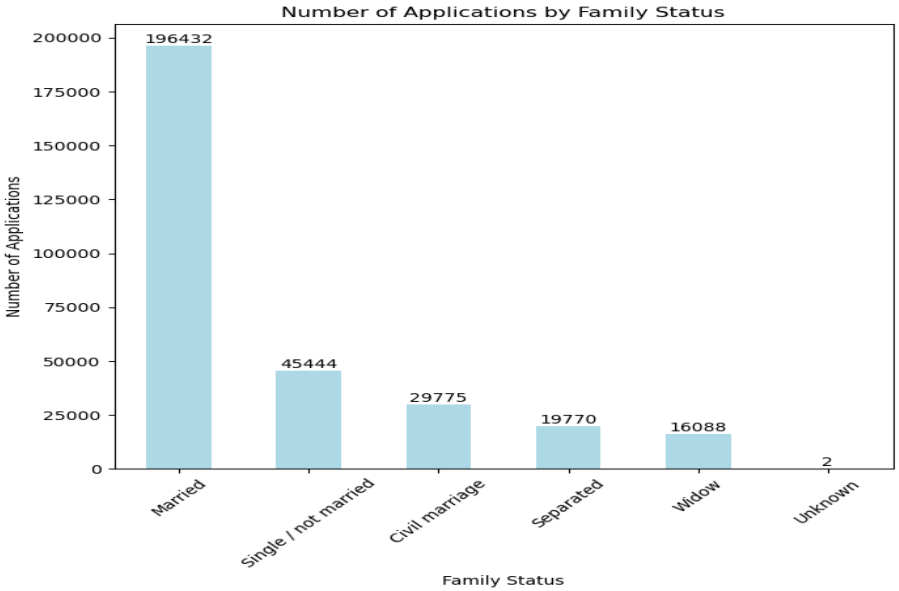


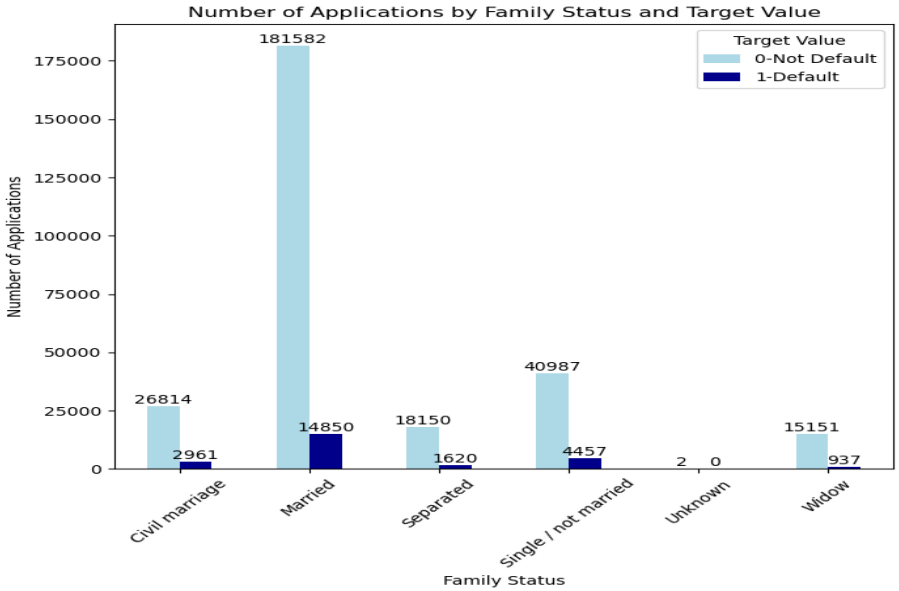
Insight : **Laborers** make up **22.4%** of **loan applications,** with the **highest defaults** (5,838) and a 10.6% default rate. **Sales Staff** (9.6%) and **Drivers** (11.3%) also show **instability**. Core Staff (6.3%) and High Skill Tech Staff (6.2%) have moderate rates, while Accountants have the lowest at 4.8%, indicating better financial stability.

Image source: Icons8

Exploratory Data Analysis

Analyzing the Family Status



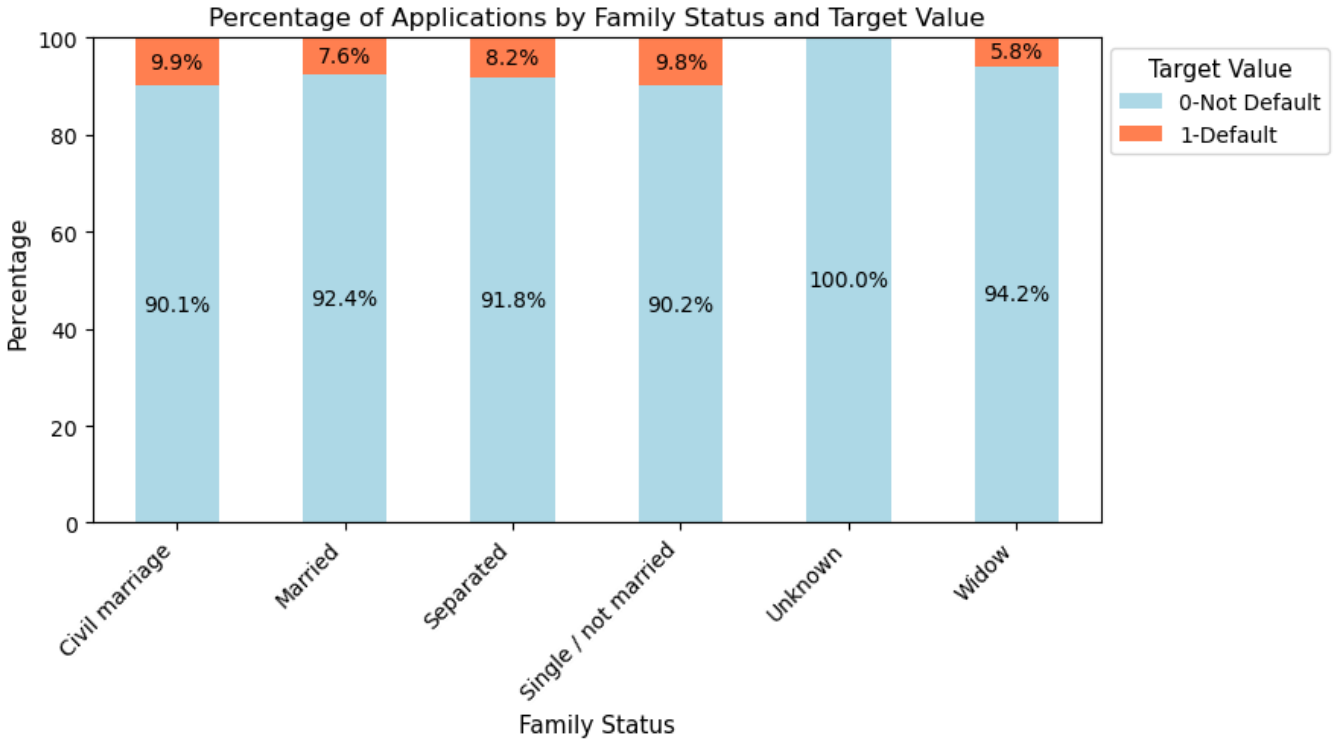


Insight :

- **Married** individuals account for the **largest share** of applications (196,432), representing about **62.5%** of total applications.

Exploratory Data Analysis

Analyzing the Family Status



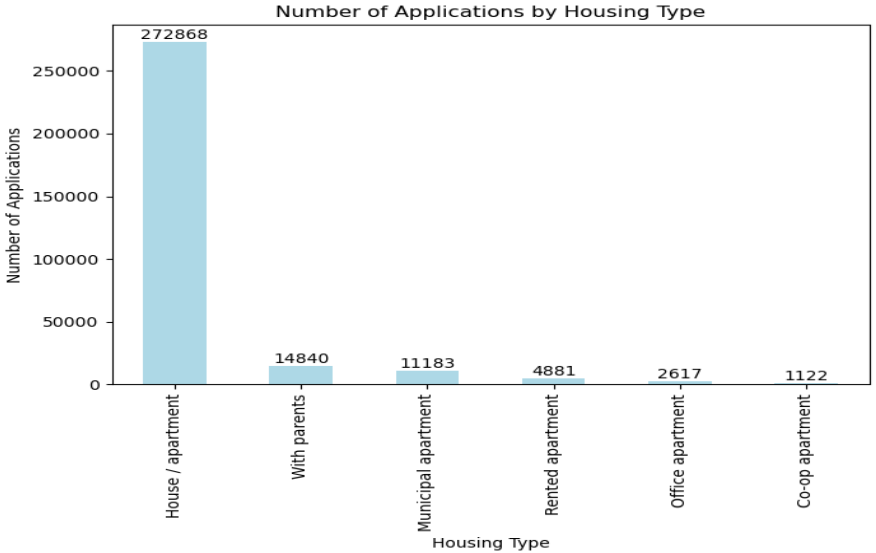
Insight:

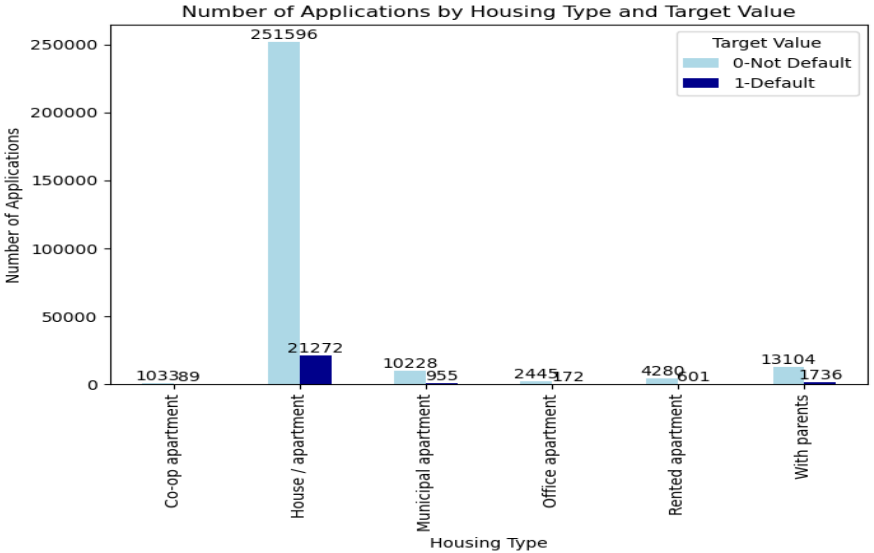
- Married individuals make up 62.5% of applications (196,432).
- Civil Marriage default rate is 9.9%, and Separated individuals have an 8.2% rate.
- Widows have a lower default rate of 5.8% (16,088 applications).
- Single/Not Married applicants (45,444) have a higher default rate of 9.0%.
- Married applicants show a low default rate of 7.6%.

Image source: Icons8

Exploratory Data Analysis

Analyzing the Housing Type





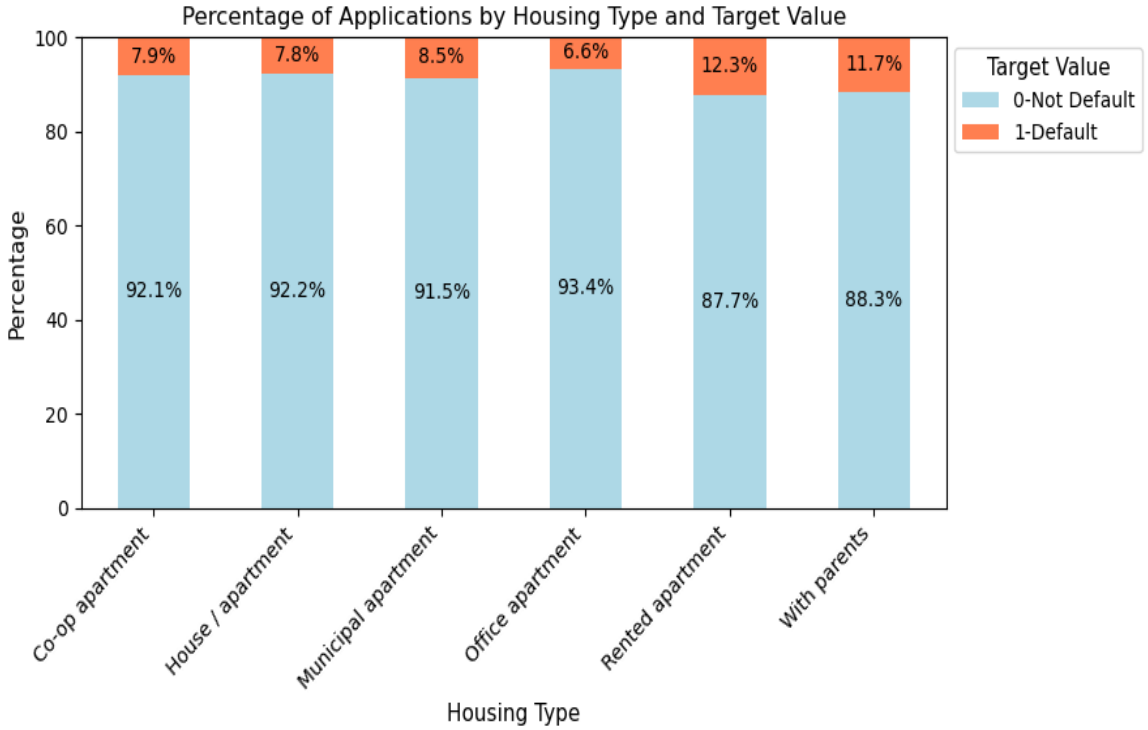
Insight :

- The data shows that 70% (272,868) of loan applications come from individuals in stable housing, indicating a strong link betwee housing stability and loan-seeking behavior. Rented apartments account for 4,881 applications (1.3%), office apartments for 2,617 (0.7%), and co-op apartments for 1,122 (0.3%), highlighting their limited presence among applicants.

Image source: icons8

Exploratory Data Analysis

Analyzing the Housing Type



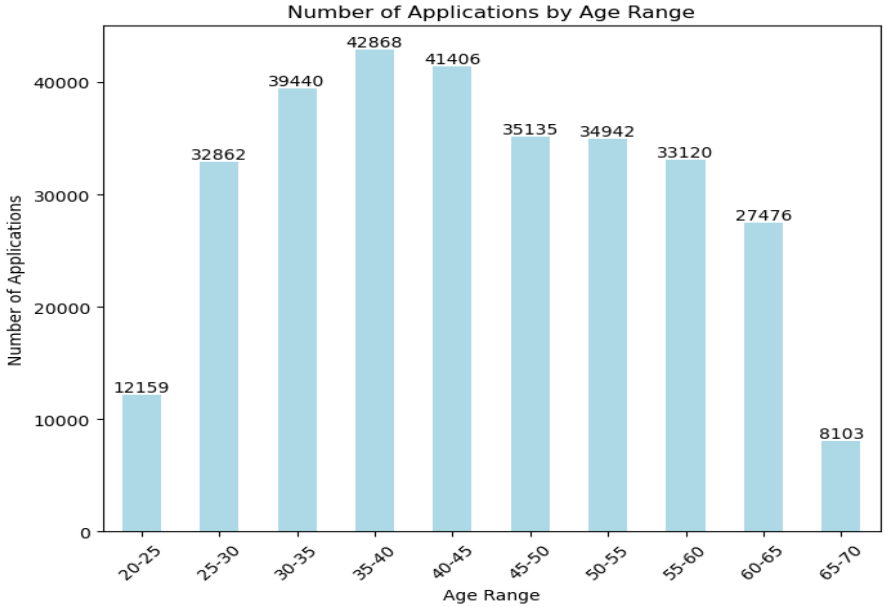
Insight:

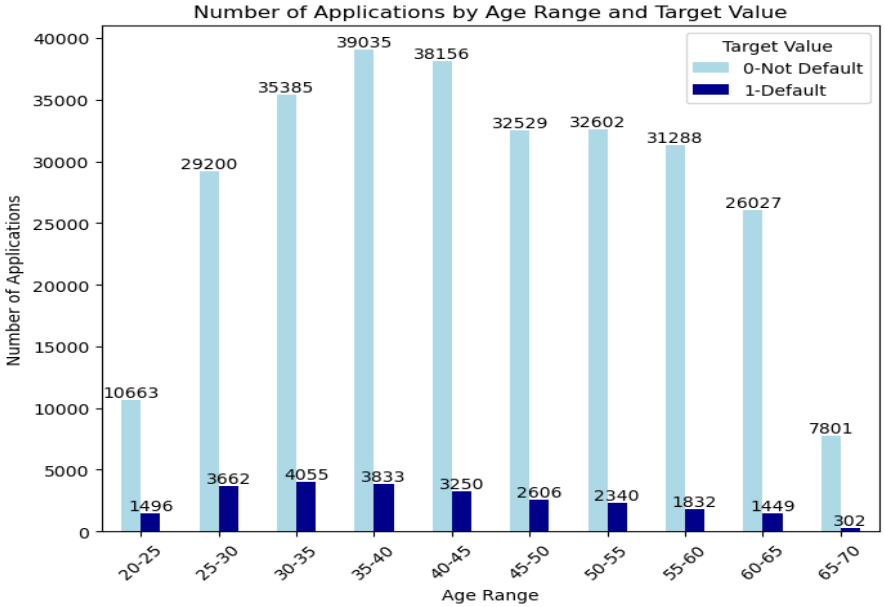
- **Majority of Applications**: 272,868 (about 70%) from stable housing.
- **Rented Apartments**: 4,881 applications (1.3%).
- **Office Apartments**: 2,617 applications (0.7%).
- **Co-op Apartments**: 1,122 applications (0.3%).
- **Default Rates**: House/apartment applicants at 7.8%, while living with parents is 11.7%.
- **Renters**: Highest default rate at 12.3% (601 defaulters).

Image source: Icons8

Exploratory Data Analysis

Analyzing the Age Range



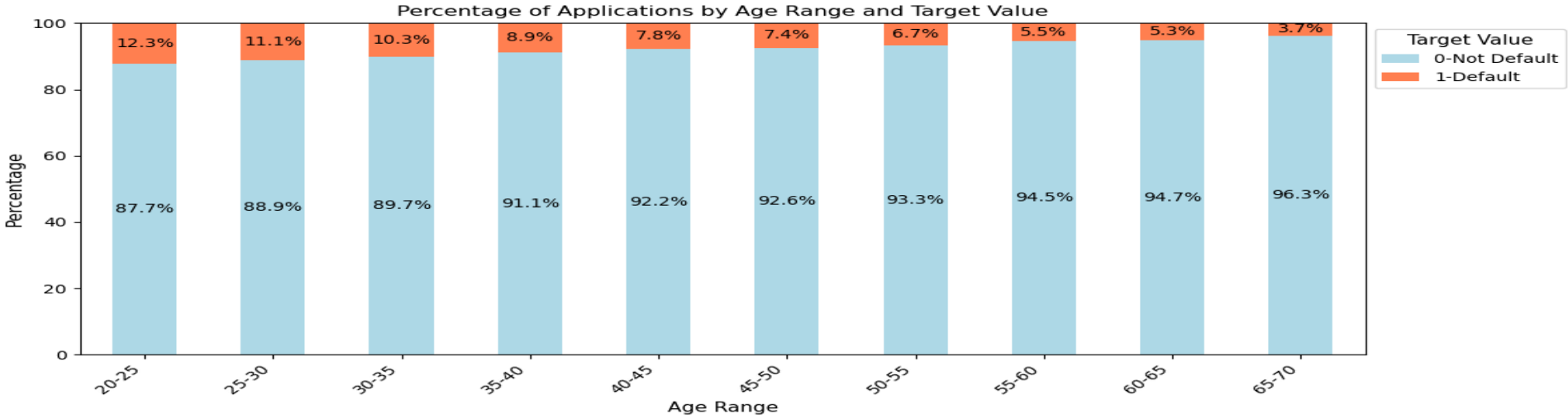


Insight :

- The **35-40 age group leads** with 42,868 applications, about **16.2%** of the total, The **30-35 age group follows** closely with 39,440 applications (**15.0%**), indicating financial activity related to career and family investments.
- The **25-30 age group** has 32,862 applications, making up around **12.4%** of total applications.

Exploratory Data Analysis

Analyzing the Age Range

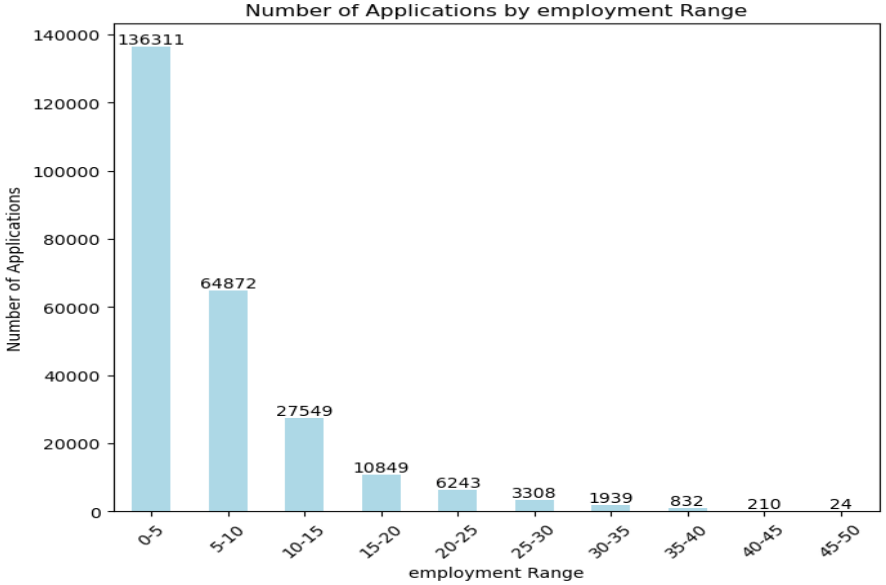


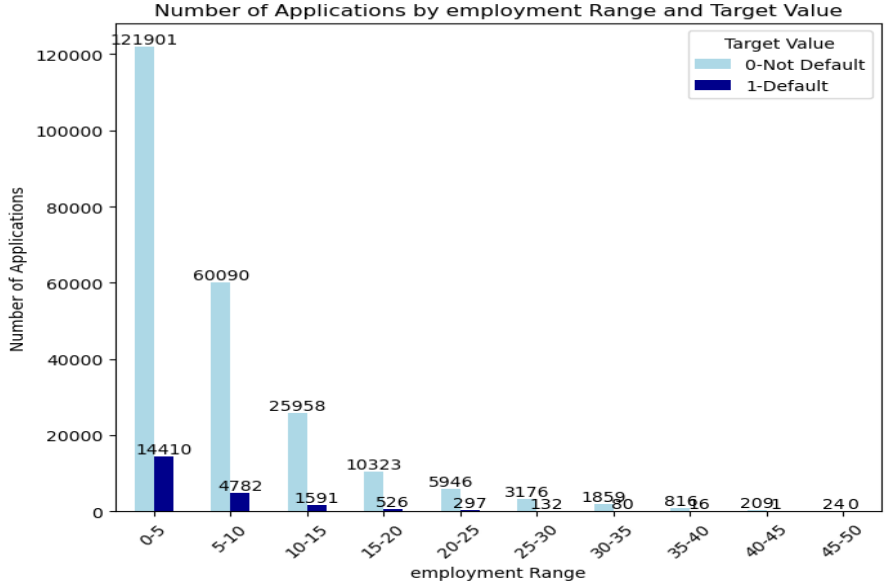
Insight :

- The **35-40 age group** has the most applications at **42,868** (16.2%), followed by **30-35** with **39,440** (15.0%), and **25-30** with **32,862** (12.4%).
- The **20-25 age group** shows the highest **default rate** at **12.3%**, indicating repayment challenges.
- **Default rates decrease** with age, falling to **3.7%** for the **65-70 age group**, suggesting greater financial stability among older borrowers.

Exploratory Data Analysis

Analyzing the Employment Range





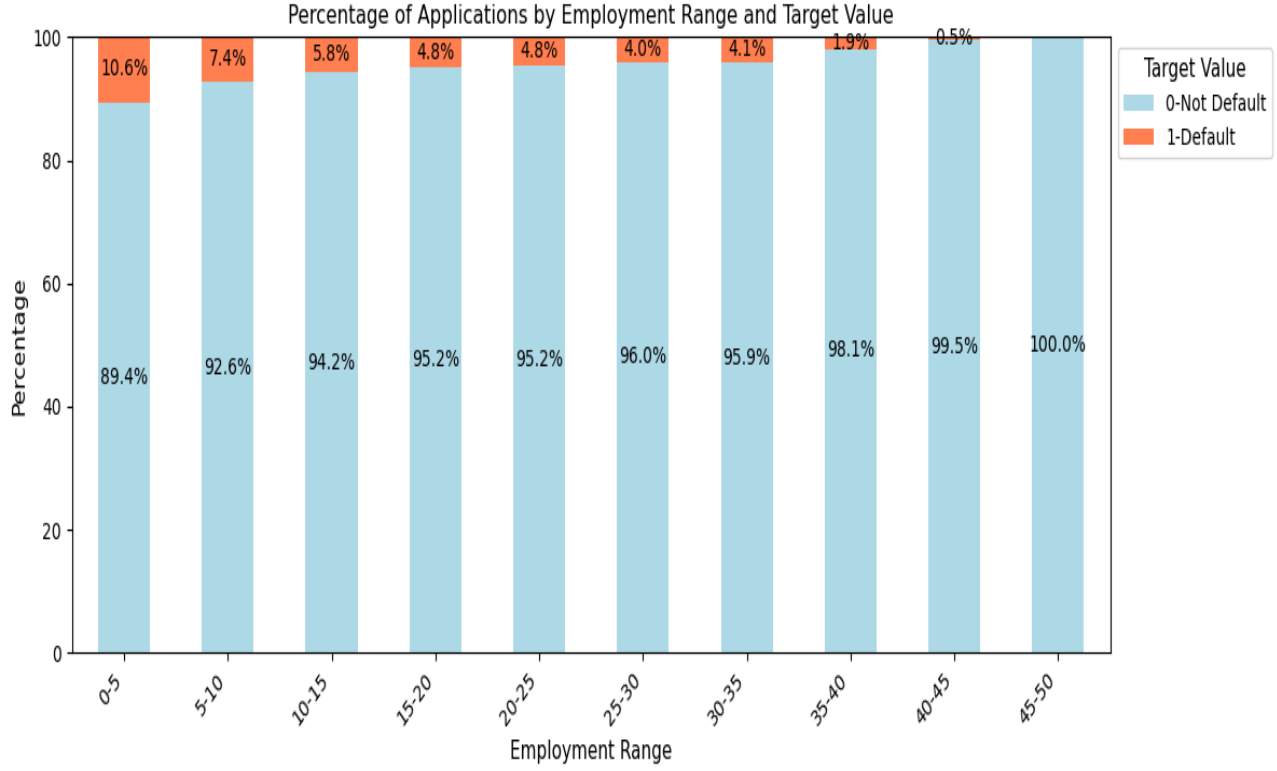
Insight:

- Most applications are from individuals with **0-5 years** (136,311) and **5-10 years** of experience (64,872).
- **Higher default rates** are seen in **less experienced borrowers**.
- Default rates drop with experience; **10-15 years** has **5.9%** and **45-50 years** has no defaults.
- Less experience is linked to higher default risk, while more experience indicates stability.

Image source: Icons8

Exploratory Data Analysis

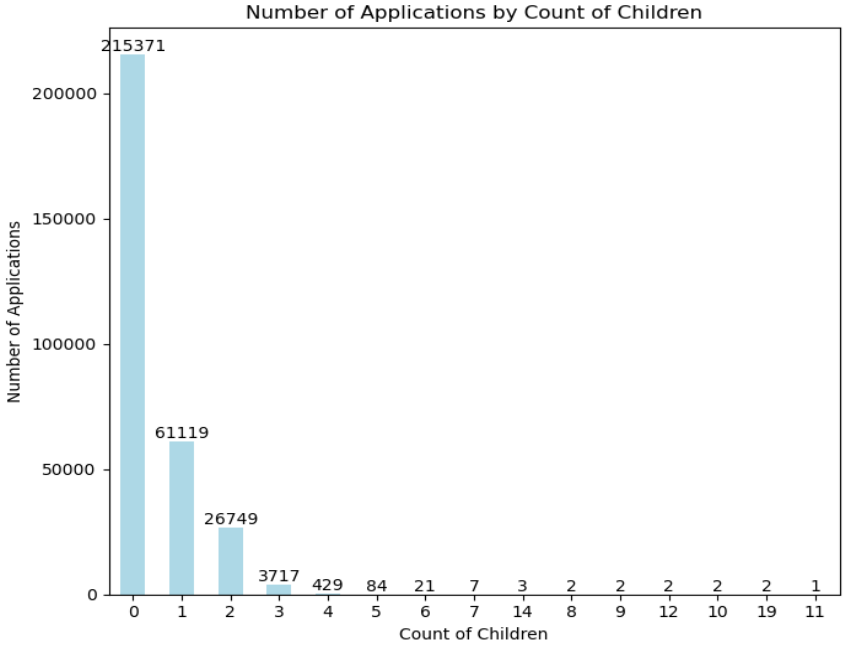
Analyzing the Employment Range

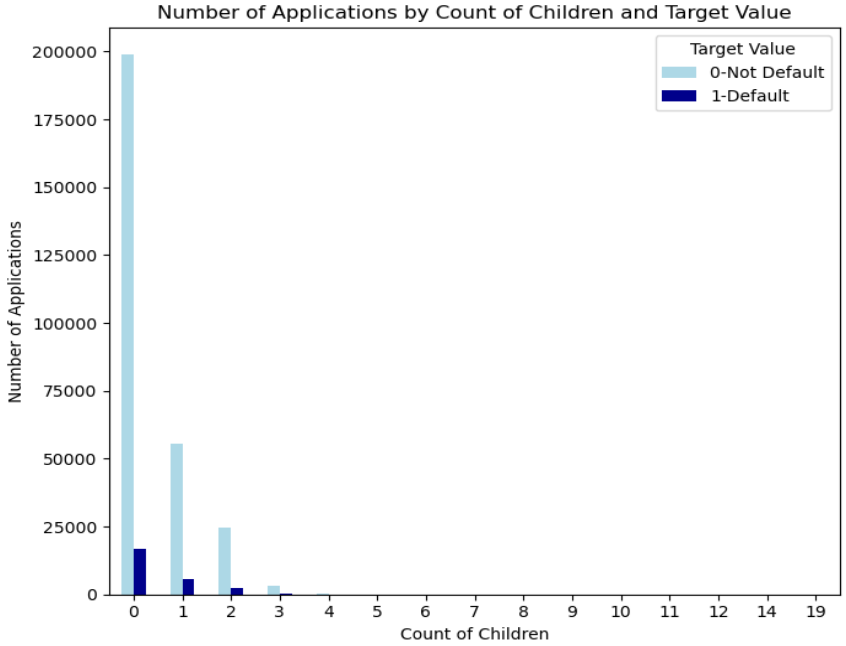


- Most applications (136,311) are from individuals with 0-5 years of experience, followed by 64,872 from those with 5-10 years. Default rates are higher among less experienced applicants, decreasing significantly with more experience. For example, the 10-15 years group has a 5.9% default rate, while the 45-50 years group has none. This suggests that limited employment experience is linked to higher default risk, whereas longer employment histories indicate greater stability.

Exploratory Data Analysis

Analysis on number of children

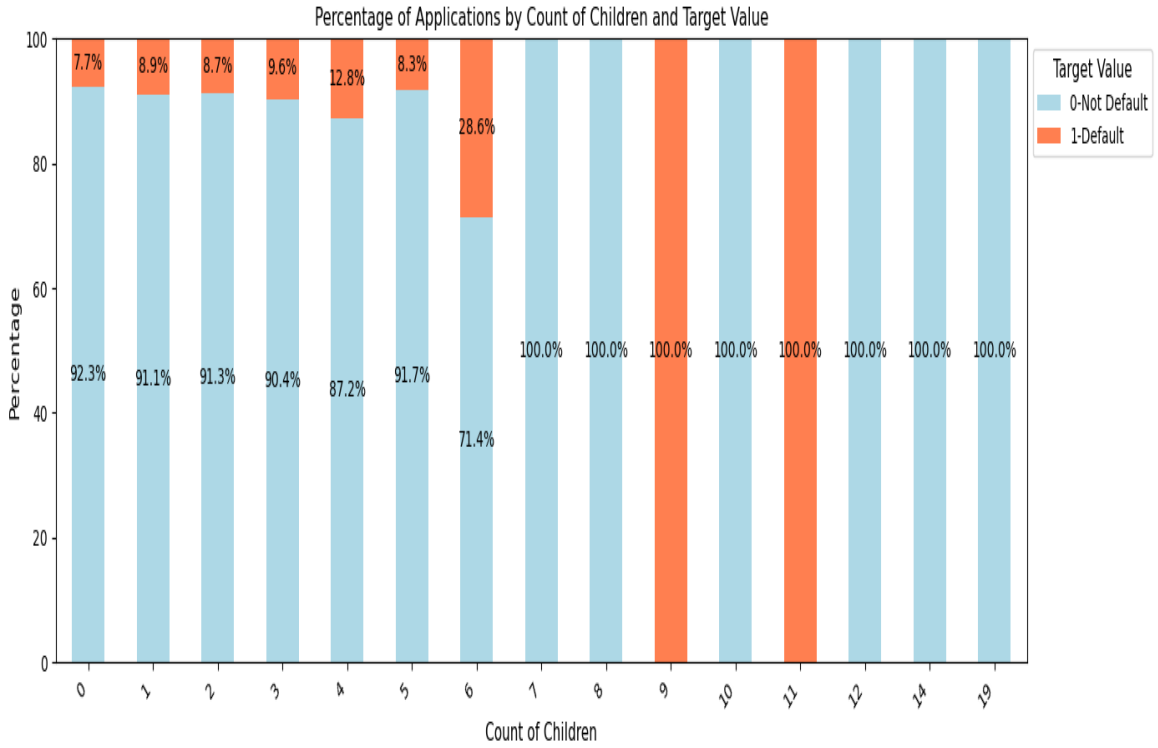




- The majority of applicants (215,371) have no children, with numbers decreasing sharply as the count of children increases.
- Notably, as the number of children increases, the proportion of defaults relative to total applications also rises, pointing to a potential correlation between family size and financial risk.

Exploratory Data Analysis

Analysis on number of children



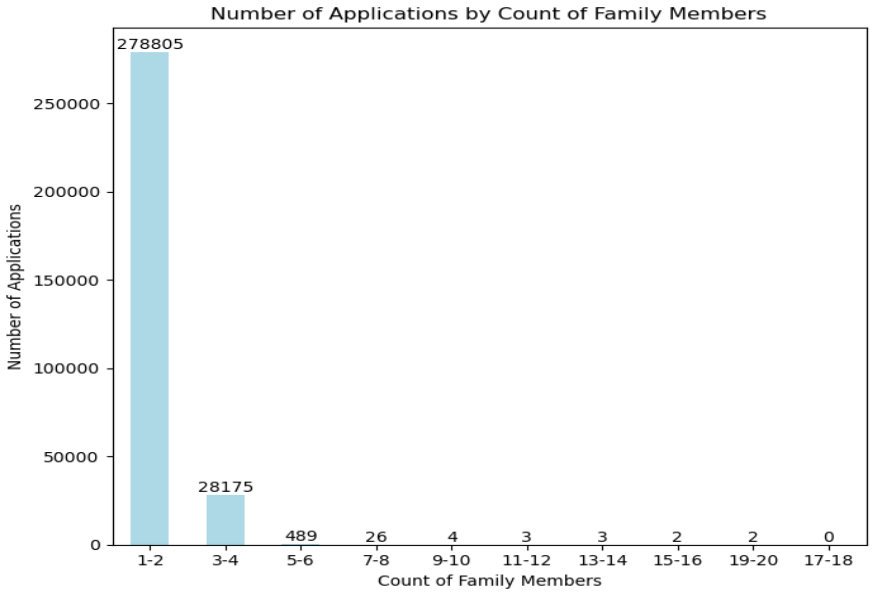
Insight:

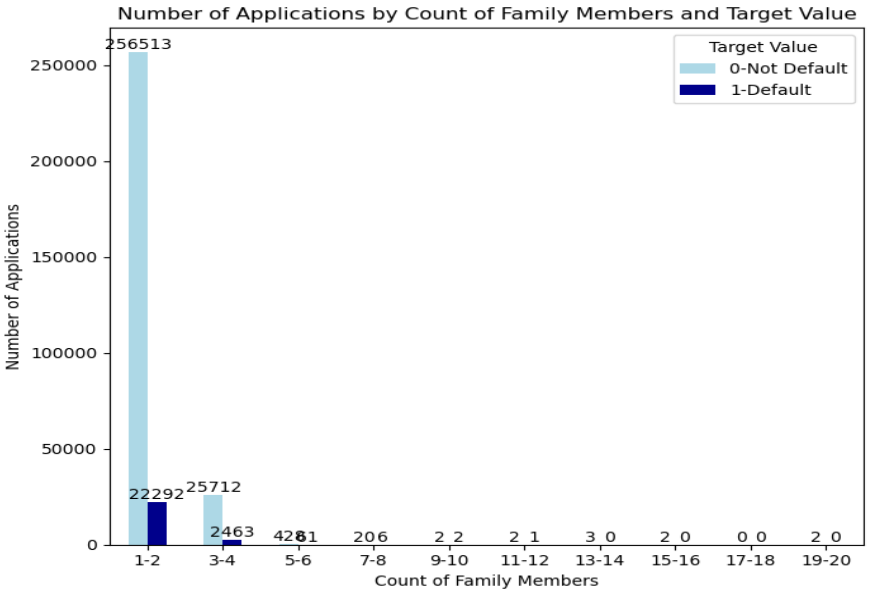
- When examining defaults, families with no children have 198,762(92.3%) non-defaults compared to 16,609(7.7%) defaults, indicating a strong likelihood of stability among childless applicants.

Image source: Icons8

Exploratory Data Analysis

Analysis on count of family members





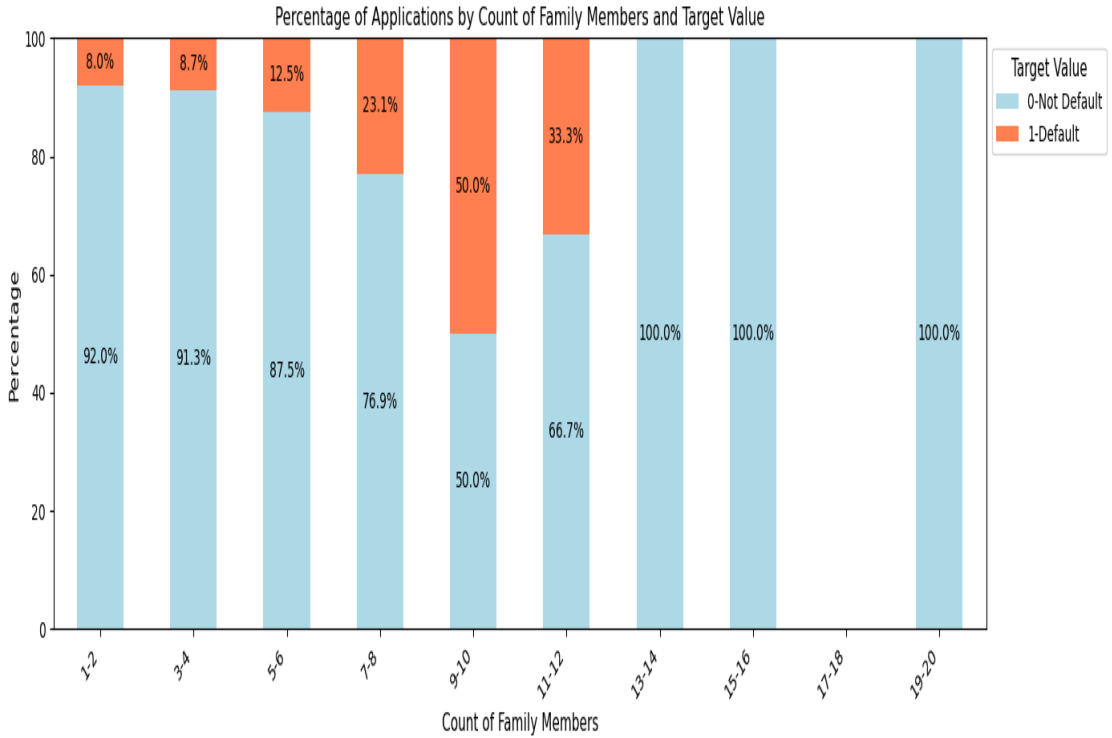
Insight:

- The majority of applications (278,805) come from families in the 1-2 member range, highlighting a preferer for smaller households.
- Applications drop sharply as family size increases, with only 28,175 from families with 3-4 members and even fewer from larger families.

Image source: Icons8

Exploratory Data Analysis

Analysis on count of family members



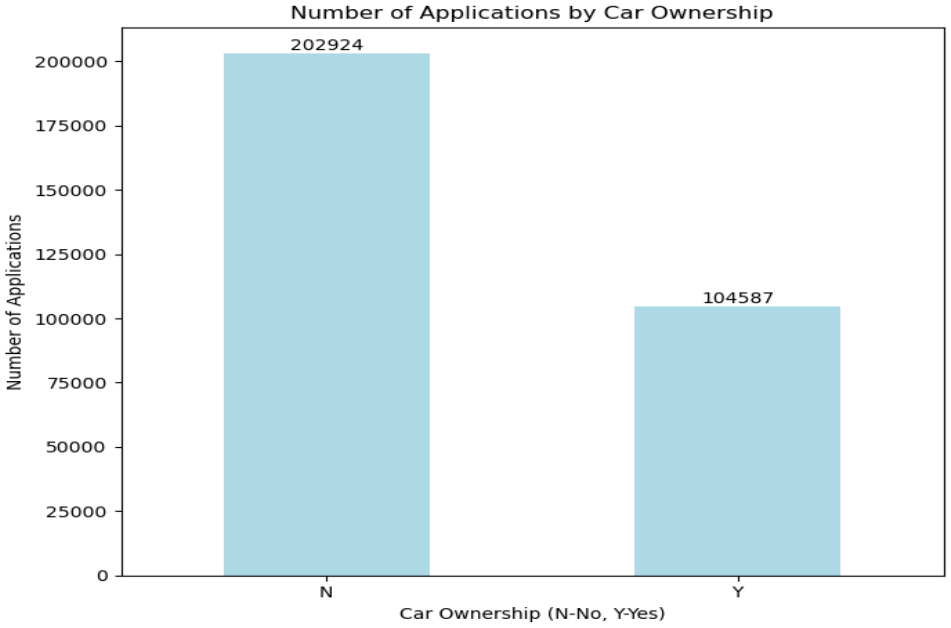
Insight:

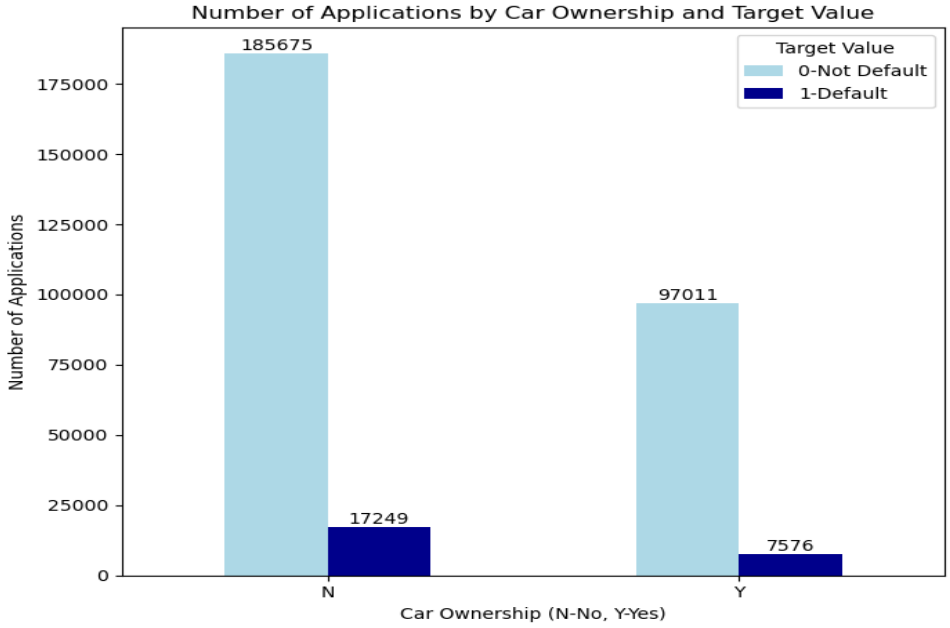
- Non-default rates are significantly higher among smaller families, with 256,513(92%) non-defaults and 22,292(8%) defaults in the 1-2 member category.
- Larger families exhibit increased financial risk, as evidenced by lower application counts and minimal default data.

Image source: Icons8

Exploratory Data Analysis

Analyzing the Car Ownership





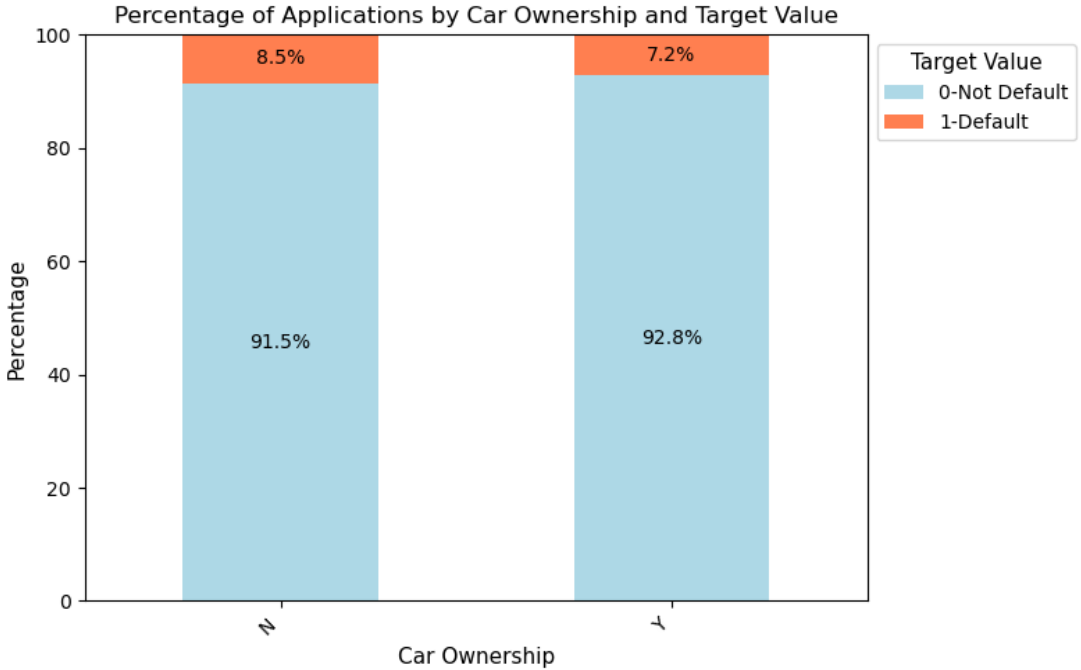
Insight :

- The **"No Car"** group comprises 202,924 applications, representing about **66% of the total**. In contrast, the **"Own Car"** group has 104,587 applications, accounting for around **34%** of the total.
- Car ownership suggests greater financial stability and the capacity to invest in assets.

Image source: icons8

Exploratory Data Analysis

Analyzing the Car Ownership



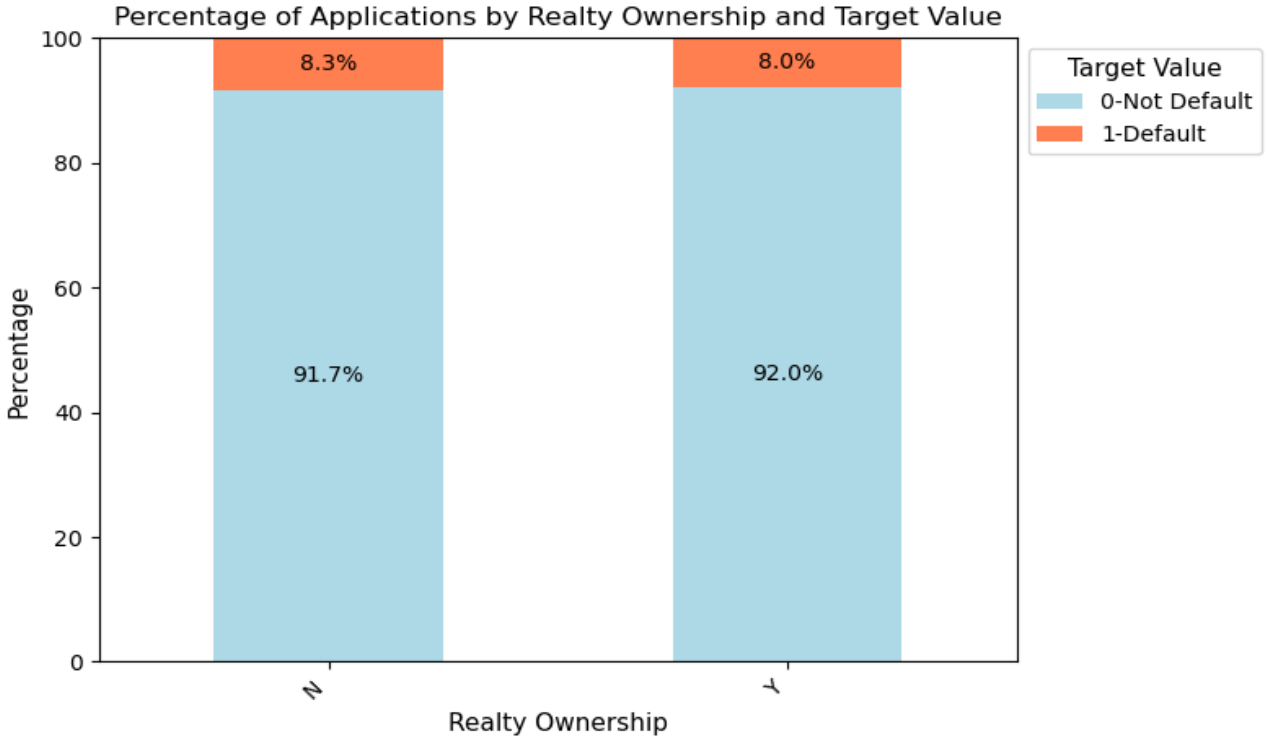
Insight :

- The **"No Car"** group has **202,924 applications** (about **66%** of total), while the **"Own Car"** group has **104,587 applications** (around **34%**).
- The **default rate** for non-car owners is **8.5%**, compared to **7.2%** for car owners, indicating that car owners may have greater financial stability and reliability in loan repayment.

Image source: Icons8

Exploratory Data Analysis

Analyzing the Reality Ownership



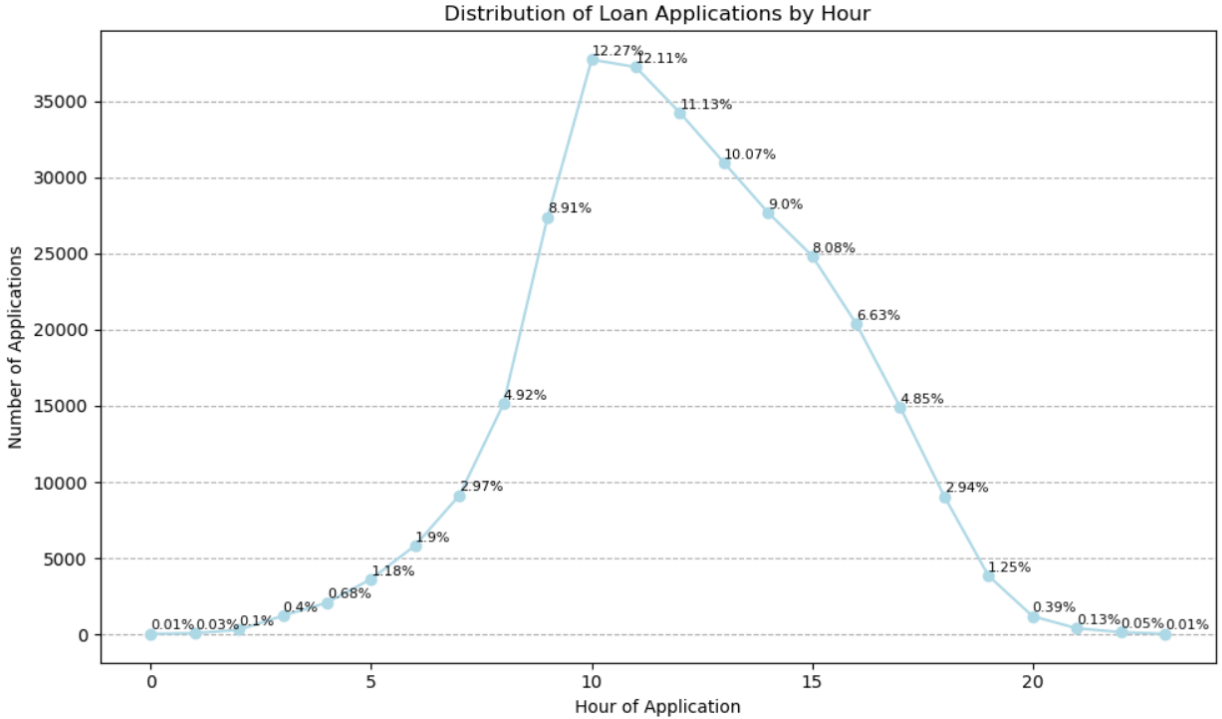
Insight :

- The **Own Realty** group has **213,312 applications** (69%), indicating financial stability, while the **No Realty** group has **94,199 applications (31%),** suggesting less financial establishment.
- The **default rate** is **8.3%** for **non-realty applicants** and 7.9% for realty owners, showing that owning realty correlates with lower financial risk.

Image source: Icons8

Exploratory Data Analysis

Analyzing the loan applications by the hour of the day



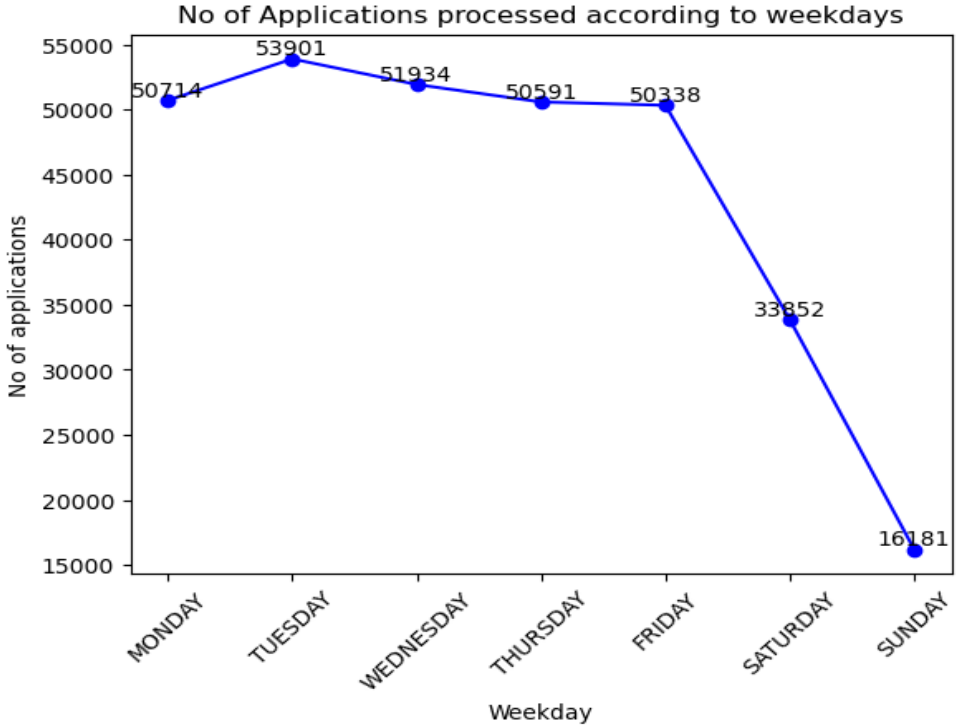
Insight :

- Application volume exhibits a clear peak between **10 AM and 1 PM**, with the **highest activity at 10 AM** (37,722 applications, **12.27%**) and **11 AM** (37,229 applications, **12.11%**).
- Applications begin to rise significantly from 8 AM, reaching 15,127 at that hour (4.92%), indicating a preference for applying early in the day. **After** the peak around **11 AM**, application numbers **gradually decline.**
- Afternoon activity remains steady but decreases further into the evening, with only 1,196 applications at 8 PM (0.39%) and just 41 at 11 PM (0.01%).

Image source: Icons8

Exploratory Data Analysis

Analyzing the loan applications by the weekday



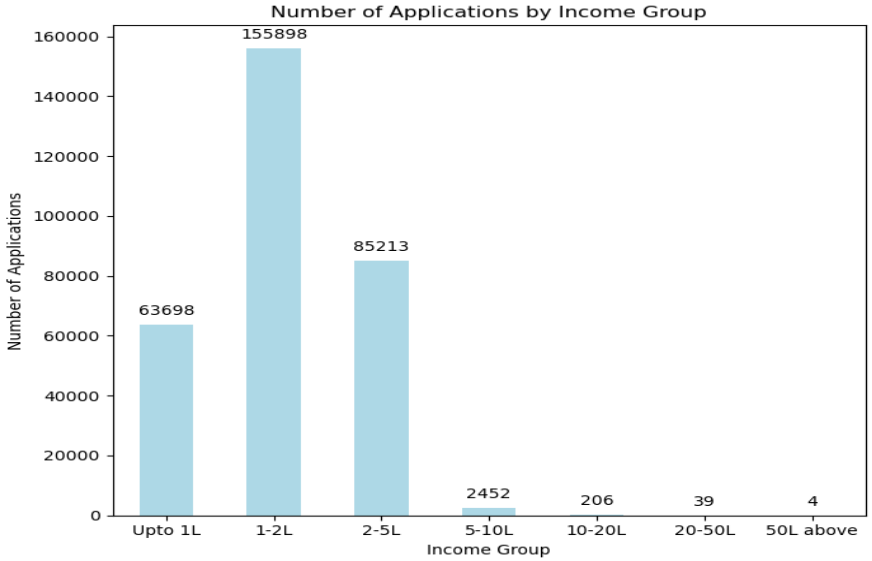
Insight :

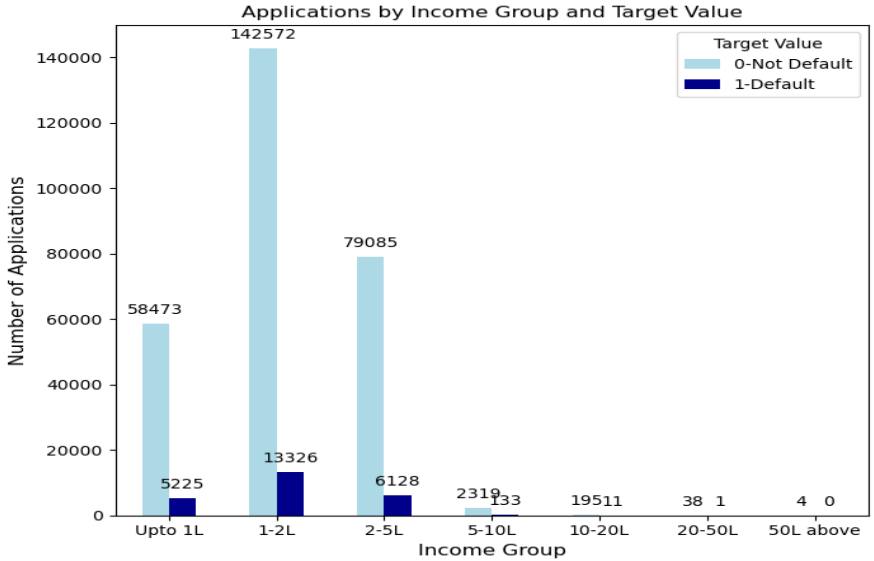
- **Tuesday** is the **peak** day **for loan applications**, with 53,901 submissions (**17.53%**), indicating that borrowers are most active early in the week, **Midweek shows steady application counts**, with **Wednesday** at 51,934 (**16.89%)** and **Thursday** at 50,591 (**16.45%**), suggesting that individuals are actively assessing their financial situations.
- However, **application volume drops over the weekend**, with **Saturday** at 33,852 (**11.01%**) and **Sunday** at just 16,181 (**5.26%**), indicating that potential borrowers are less engaged with financial decisions during this time.

Image source: Icons8

Exploratory Data Analysis

Analyzing the Income group





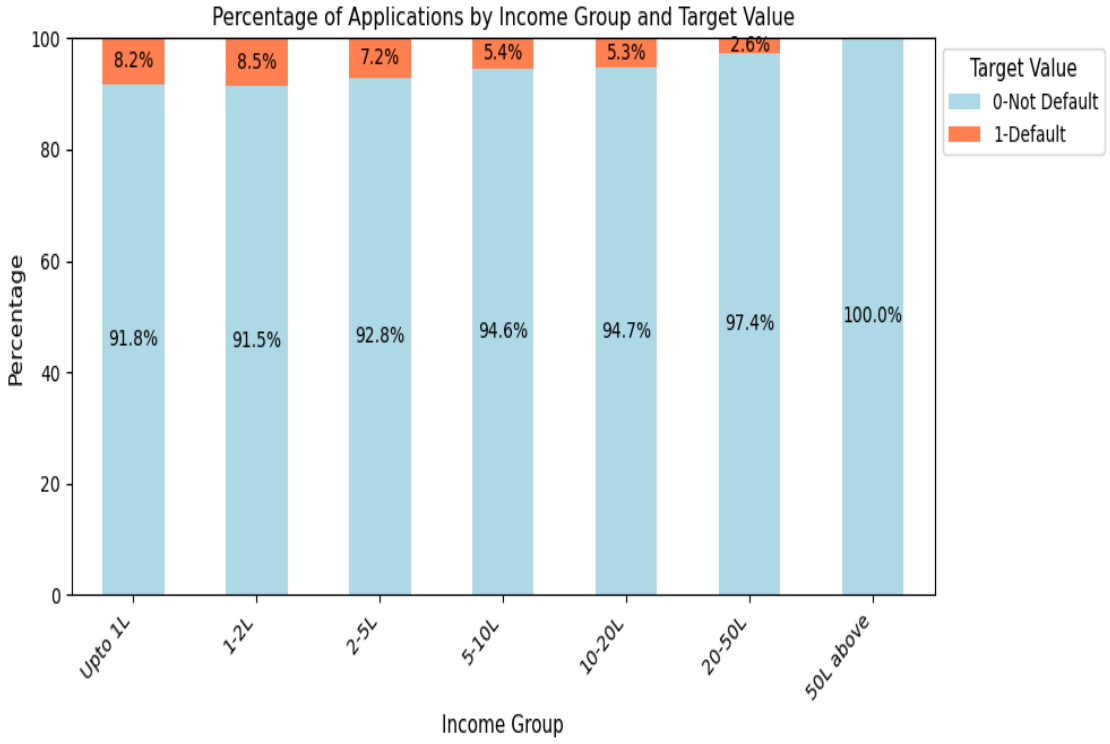
Insight :

- The data shows notable trends across income brackets. The "Upto 1 lakh" group has 63,698 applications (17.3% of total), highlighting engagement from lower-income individuals. The "1-2 lakh" group leads with 155,898 applications (42.5%), suggesting strong activity for major purchases or investments. Engagement drops sharply in higher brackets, with just 0.7% (2,452) from the "5-10 lakh" range and even fewer in "10-20 lakh" (206), "20-50 lakh" (39), and "50 lakh and above" (4) categories. This suggests most loan applicants fall within the lower to middle-income ranges.

Image source: Icons8

Exploratory Data Analysis

Analyzing the Income group



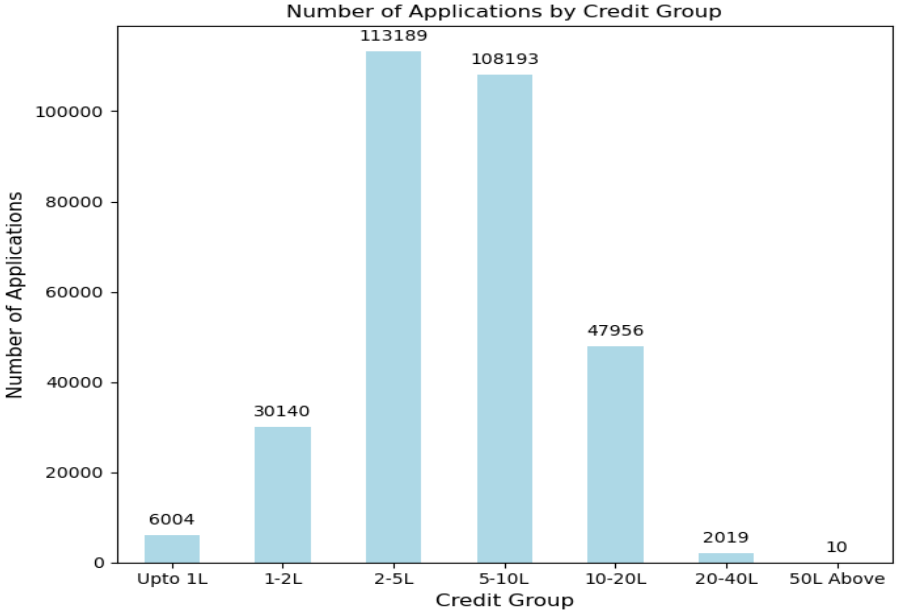
Insight:

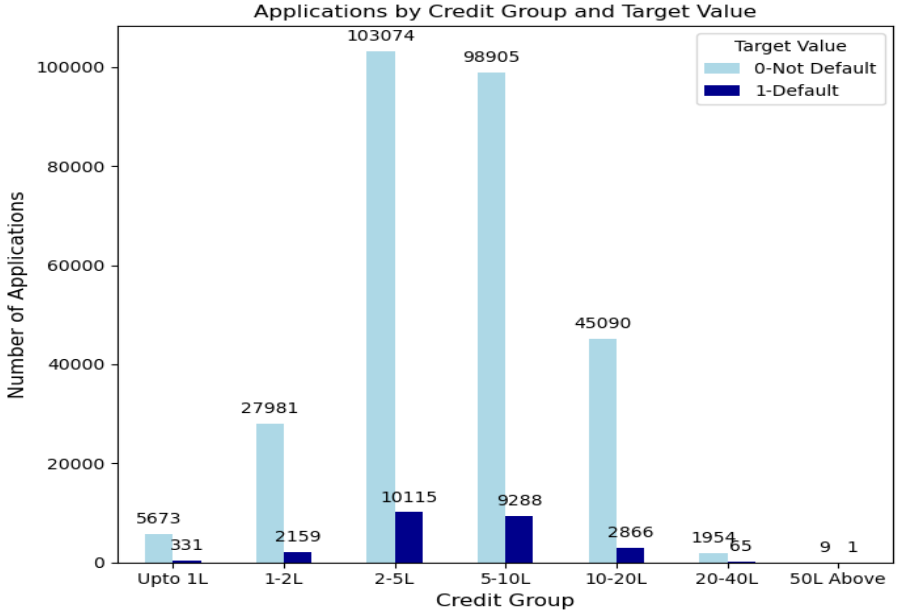
- The **"Upto 1 lakh"** group has **63,698** applications (17.3%), while **"1-2 lakh"** borrowers account for **155,898** (42.5%), indicating strong engagement from lower to middle-income applicants.
- Higher income brackets show limited interest, with only **0.7%** (2,452) from the **"5-10 lakh"** group.
- Default rates are higher in the lower-income groups, at **8.2%** for **"Upto 1 lakh"** and **8.5%** for **"1-2 lakh,"** decreasing to **7.2%** for **"2-5 lakh"** and **5.4%** for **"5-10 lakh,"** reflecting greater repayment reliability among higher-income borrowers.

Image source: Icons8

Exploratory Data Analysis

Analyzing the Credit group



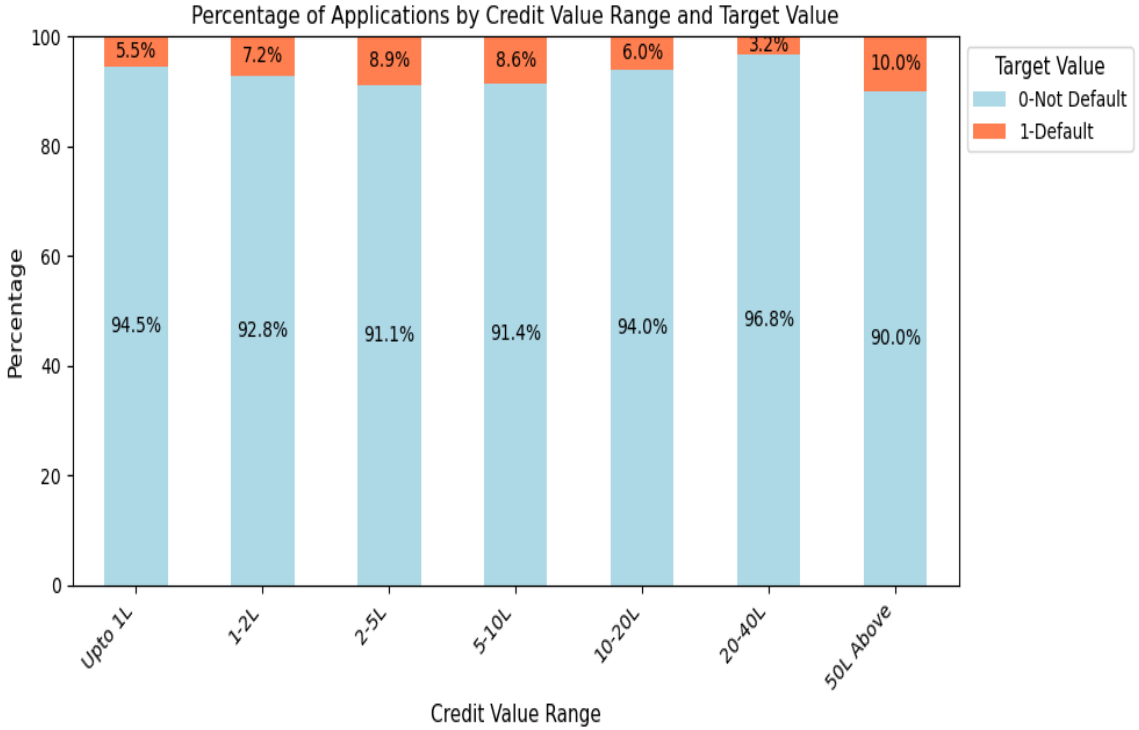


Insight :

- The **2-5 lakh credit group has the highest application volume**, suggesting a significant demand for loans within this bracket, likely for major purchases or investments.

Exploratory Data Analysis

Analyzing the Credit group



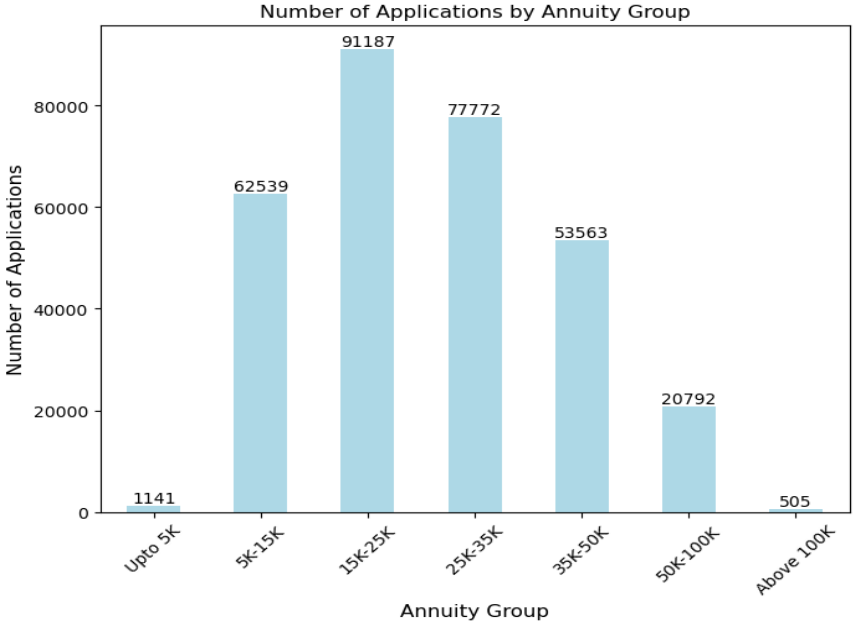
Insight :

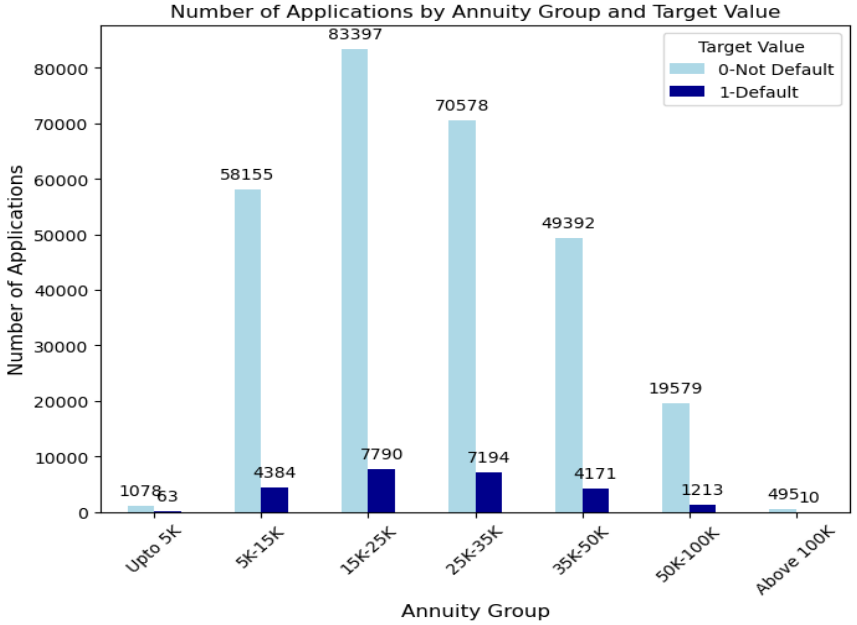
- Default rates rise with credit amounts, peaking at **8.9%** for **2-5 lakh** and falling to **3.2%** for **20-40 lakh**.
- Very high credit groups (20 lakh and above) attract fewer applicants, suggesting a preference for smaller loans or alternative financing.

Image source: Icons8

Exploratory Data Analysis

Analyzing the Annuity group





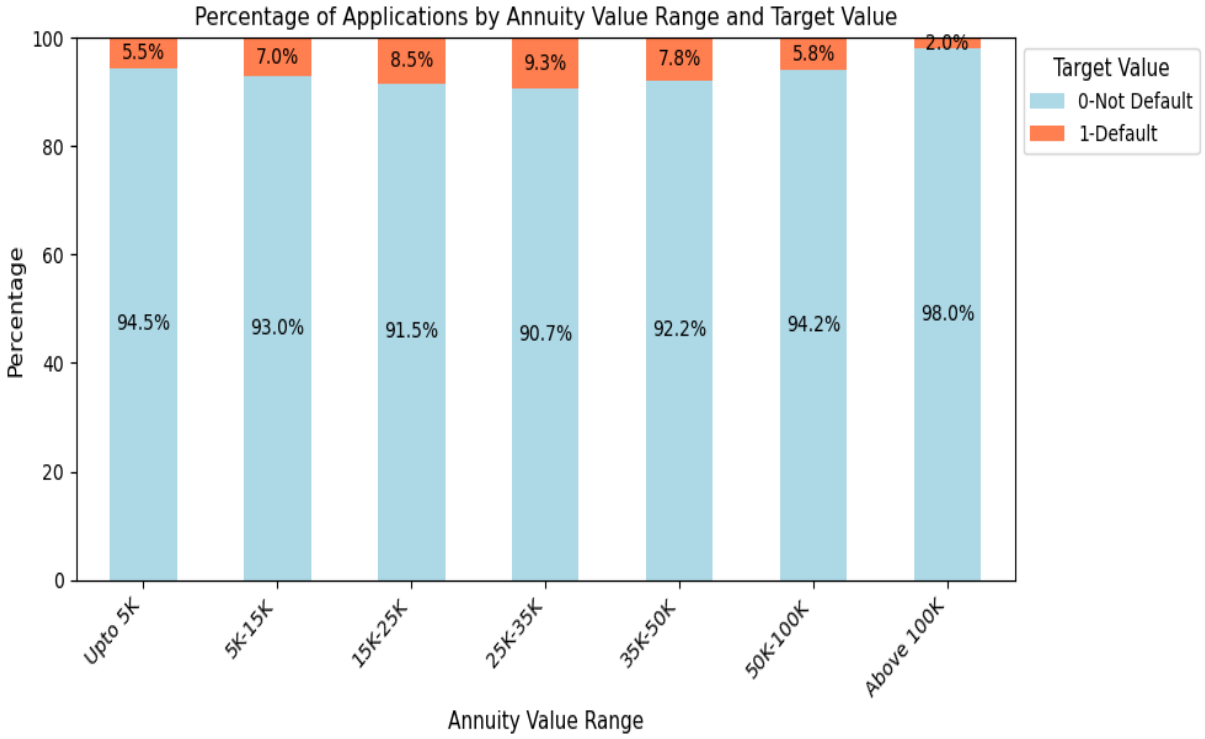
Insight :

- The 15K-25K annuity group is the most popular, with 91,187 applications (approximately 30.7% of all applications), indicating strong demand for loans in this monthly payment range. The 5K-15K group also attracts significant interest, with 62,539 applications (about 21.3%).

Image source: Icons8

Exploratory Data Analysis

Analyzing the Annuity group



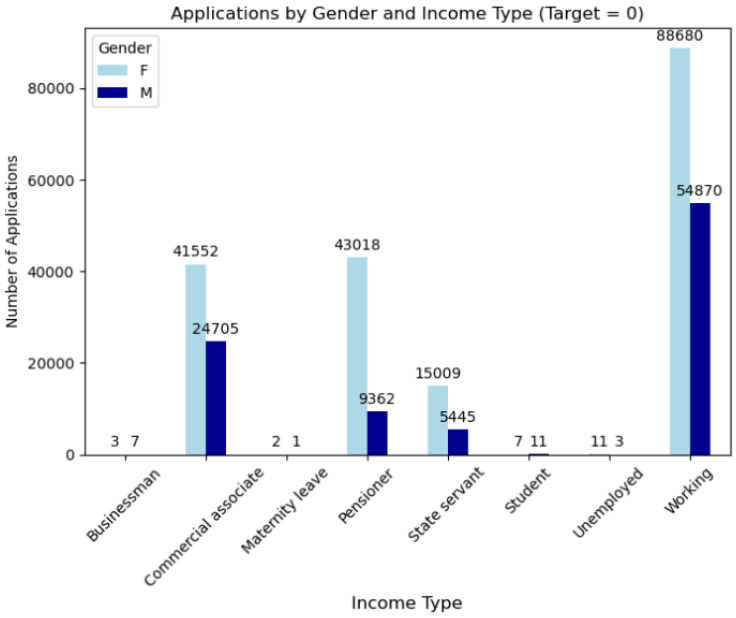
Insight :

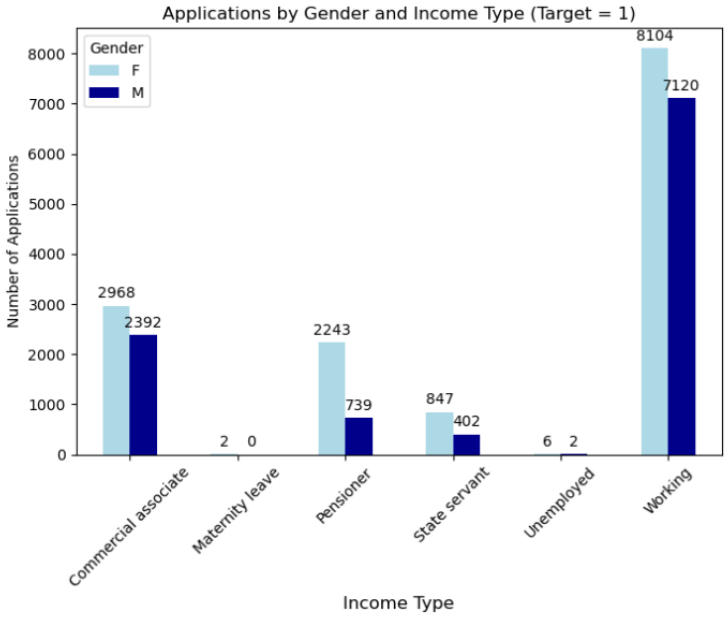
- The **25K-35K**group has **77,772** applications (26.5%), while **interest drops to 505** applications (0.2%) for **Above 100K**
- Default rates increase with payment amounts: **5.5% (Up to 5K)**, **7.0% (5K-15K)**, **8.4% (15K-25K)**, and **9.3% (25K-35K).**
- The 50K-100K group has a lower default rate of 5.8%, indicating greater financial stability.

Image source: Icons8

Exploratory Data Analysis

Analyzing the Income type by gender





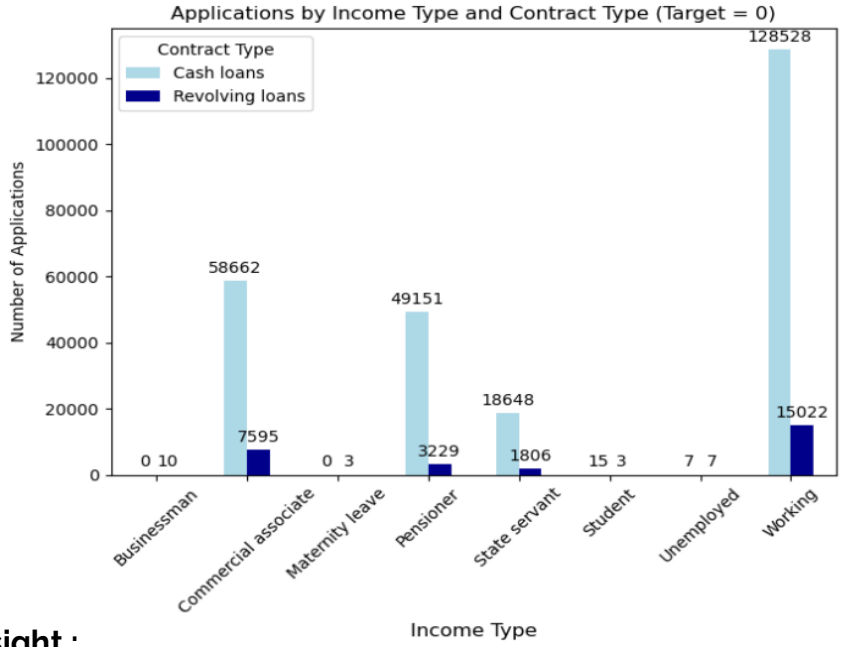
Insight :

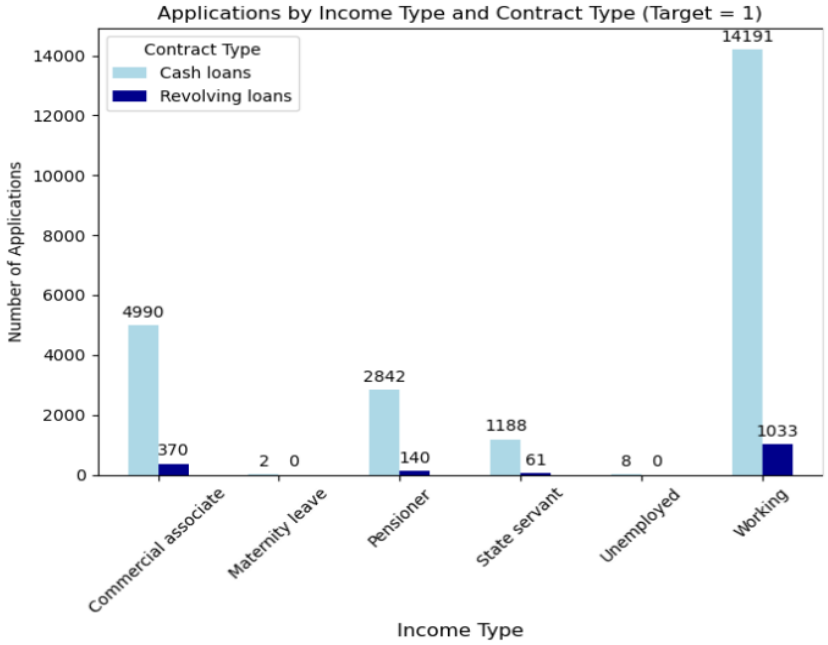
- **Both genders** face challenges, but men show **higher defaults** in the **Working category**, while low defaults among Pensioners suggest that retired individuals are generally better positioned to manage their financial obligations.

Image source: Icons8

Exploratory Data Analysis

Analyzing the Contract type and income type as per target





Insight :

- The **"Working"** category has for approximately **70% of total non-defaulter** applications, while the "Commercial associate" contributes about 25% of these applications. Among defaulters, both the **"Commercial associate" and "Working" categories represent roughly 20%** of their total applications, highlighting a higher risk associated with these income types.

Image source: Icons8

Exploratory Data Analysis

Average Income type by gender according to Target

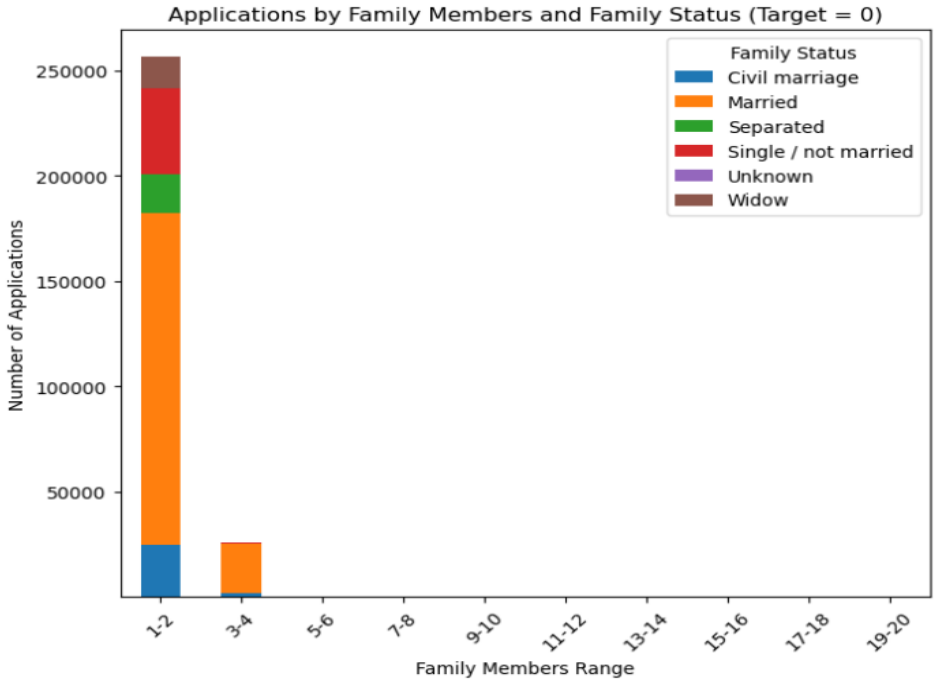
Insight :

- Overall, **males have a higher average income than females in both target categories**, indicating a gender income gap.
- While female incomes are closer in range between defaulters and non-defaulters, male incomes show a more significant disparity, suggesting that income stability could be more critical for male borrowers in avoiding defaults.

Image source: Icons8

Exploratory Data Analysis

Analyzing family member and family status

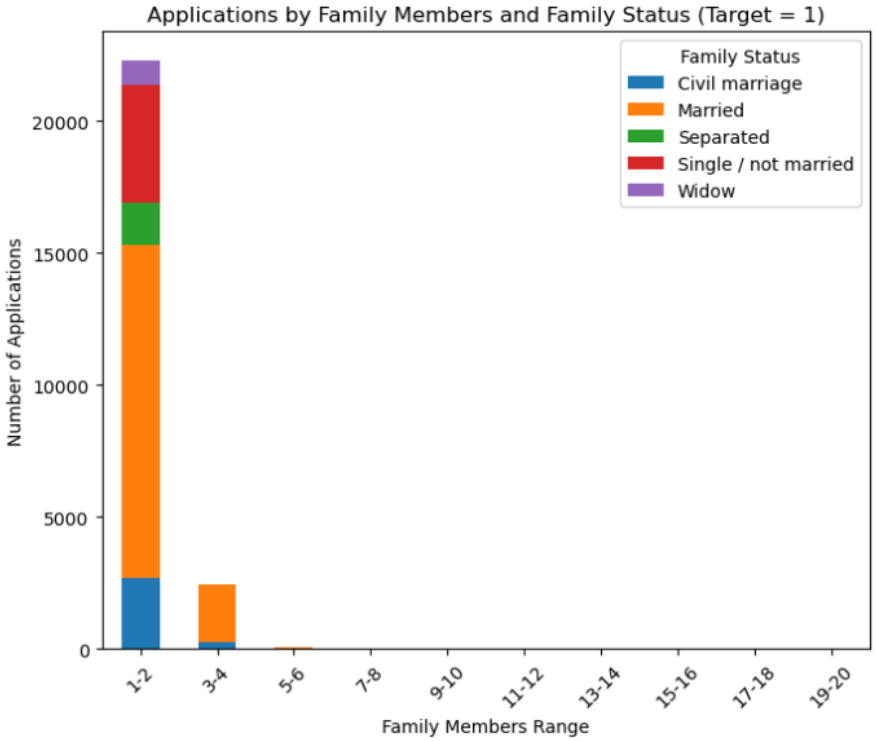


Insight :

- **1-2 Member Families:** Majority of non-defaulters, suggesting smaller units manage finances better.
- **Married Applicants:** 157,646 applications, indicating strong repayment rates among married individuals.
- **Single/Unmarried Applicants:** 40,904 applications, also showing high repayment capability.
- **3-4 Member Families:** 23,517 applications from married couples; repayment rates still significant.
- **5+ Member Families:** Sharp decline in applications, suggesting larger families face more financial challenges.

Exploratory Data Analysis

Analyzing family member and family status



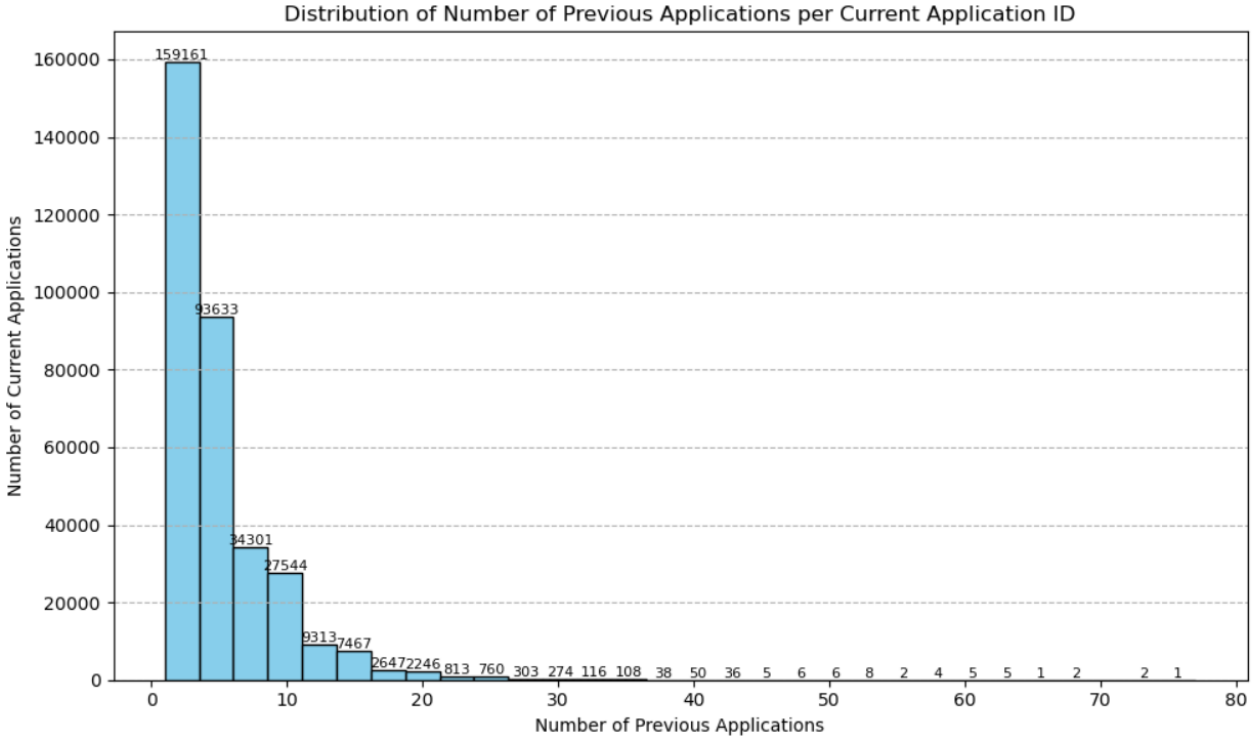
Insight :

- In the default category, the trend reveals that **smaller families still have the highest number of defaults,** with 12,625 defaulters among married individuals and 4,442 defaulters for singles. This indicates that even though smaller family units may generally be more financially stable, they can still encounter significant financial difficulties leading to defaults.
- In contrast, **larger families show relatively low default rates**, suggesting that having more family members can provide a buffer against financial hardships.

Image source: Icons8

Exploratory Data Analysis

Relationship between current application and previous application IDs



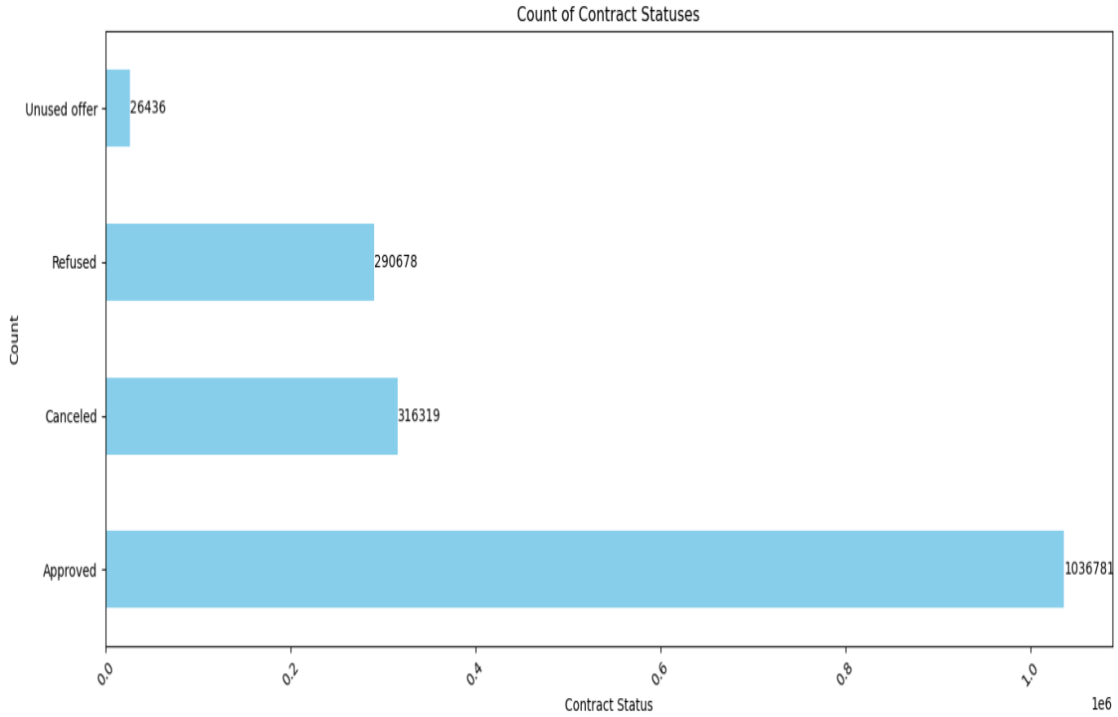
Insight :

- **Maximum number** of current applications have around **0 to 10 previous applications**.
- The **number keeps** on **decreasing with** the greater **number of previous applications**.
- only three customer had previous applications greater than 70

Image source: Icons8

Exploratory Data Analysis

Analyzing the Contract Status



Insight :

- The contract status data reveals a **strong approval rate**, with over 1 million contracts approved, indicating an effective acceptance process.
- However, the significant numbers of **cancellations** (316,319) and **refusals** (290,678)—**together** accounting for about **37% of total contracts**—highlight potential issues in negotiations or applications that warrant investigation.
- The relatively low count of unused offers (26,436) suggests effective follow-up, yet it also raises the possibility of missed opportunities for conversion.

Image source: Icons8

Exploratory Data Analysis

Analyzing the number of applications by Loan Type

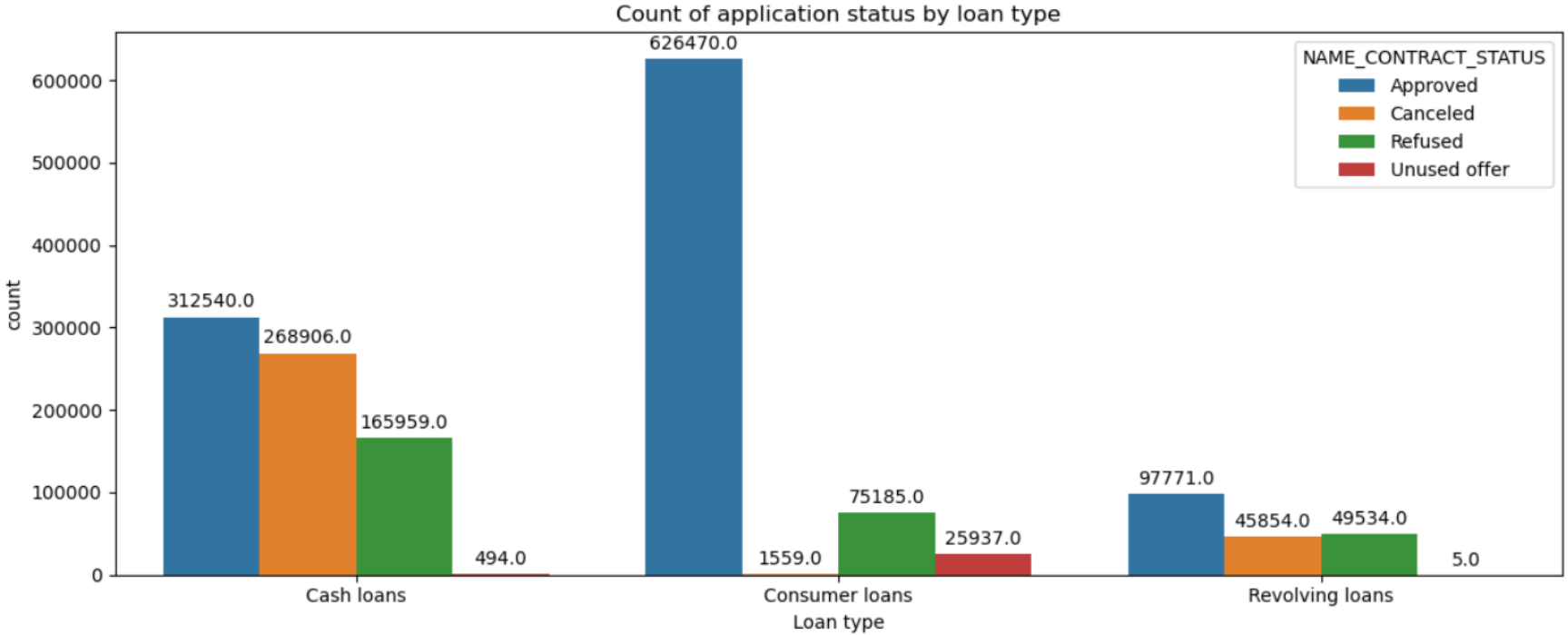


Image source: Icons8

Exploratory Data Analysis

Analyzing the number of applications by Loan Type

Insight :

- **Consumer loans** have the **highest approval count** at 626,470, indicating strong demand and acceptance.
- In contrast, **cash loans** show a **high cancellation rate** (268,906) and a notable refusal count (165,959), suggesting potential challenges in this category.
- **Revolving loans**, while having the lowest approval count (97,771), also exhibit a **significant number of refusals** (49,534) and **cancellations** (45,854).
- The low counts of unused offers across all loan types indicate effective follow-up, but the discrepancies in approval and refusal rates suggest targeted improvements may be needed, particularly for cash and revolving loans, to enhance overall conversion rates.

Exploratory Data Analysis

Analyzing the Purpose and the contract status

CONTRACT STATUS/NAME_CASH_LOAN_PURPOSE	Building a house or an annex	Business development	Buying a garage	Buying a holiday home / land	Buying a home	Buying a new car	Buying a used car
Approved	675	130	39	132	200	221	881
Canceled	98	19	8	19	39	50	98
Refused	1920	277	89	382	626	735	1896
Unused offer	0	0	0	0	0	6	13

NAME_CONTRACT_STATUS /NAME_CASH_LOAN_PURPOSE	Car repairs	Education	Everyday expenses	Money for a third person	Other	Payments on other loans	Purchase of electronic equipment
Approved	358	765	1236	12	6677	304	588
Canceled	17	21	13	0	314	70	8
Refused	422	782	1147	13	8519	1553	461
Unused offer	0	5	20	0	98	4	4

Image source: Icons8

Exploratory Data Analysis

Analyzing the Purpose and the contract status

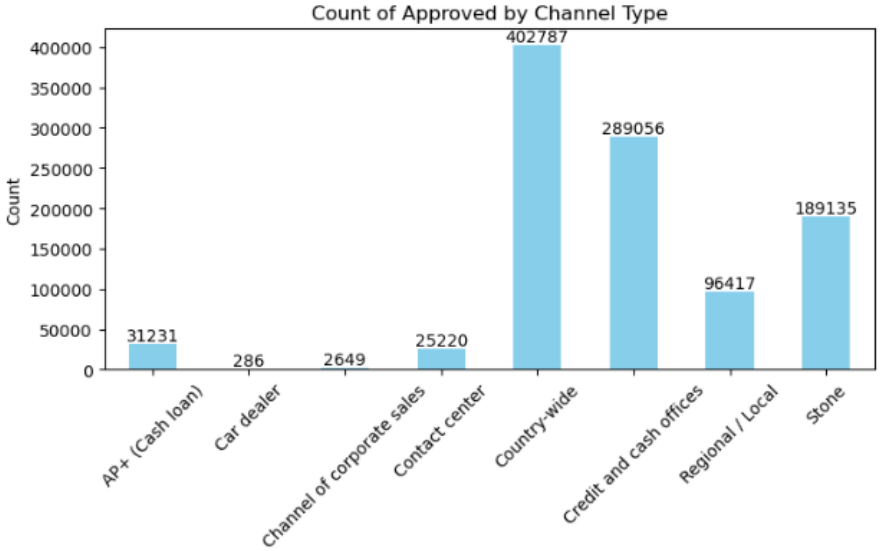
NAME_CONTRACT_STATUS /NAME_CASH_LOAN_PURPOSE	Refusal to name the goal	Repairs	Urgent needs	Wedding / gift / holiday	XAP	XNA
Approved	4	8677	3574	397	724241	285607
Canceled	0	621	148	23	47728	266952
Refused	11	14421	4690	542	124750	125070
Unused offer	0	46	0	0	25942	289

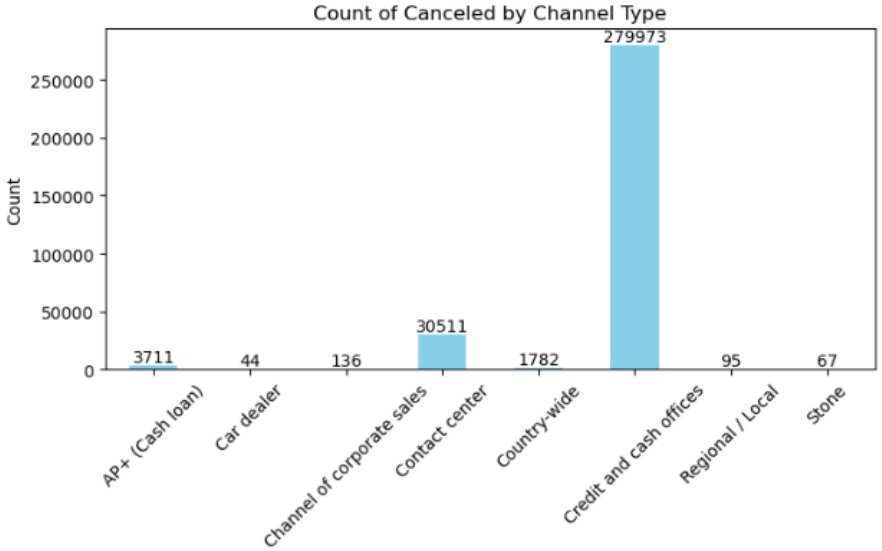
Insight :

- Purposes like XAP, purchase of electronics, every day expenses and education have maximum loan acceptance.
- Payment of other loans, refusal to name goal (can be suspicious) , buying new home or car have most refusals.
- 40% of XNA(Not available) purpose loans are cancelled, followed by buying a garage/home/car.
- % unused is too low to get any insight.

Exploratory Data Analysis

Analyzing Contract Status by Channel Type





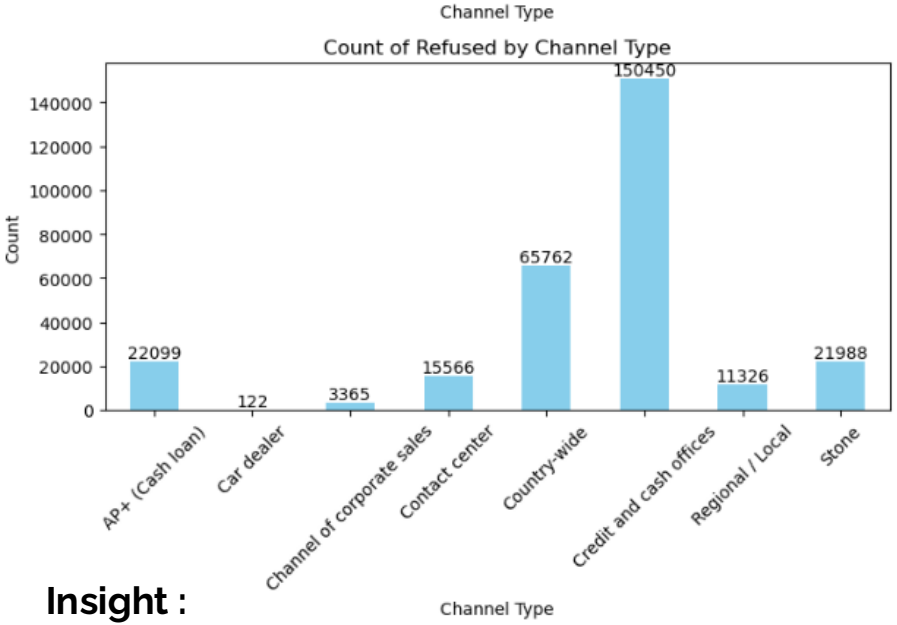
Insight :

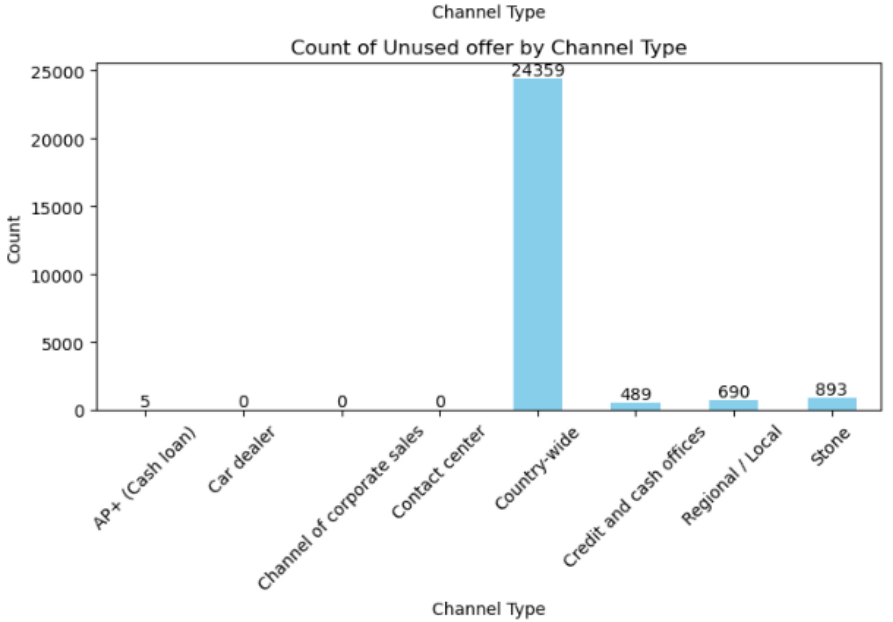
- The "Country-wide" channel has the highest approval count (402,787) and a relatively low cancellation rate (1,782), indicating strong performance in this channel.
- The "Contact center" channel has the highest cancellation count (30,511)

Image source: icons8

Exploratory Data Analysis

Analyzing Contract Status by Channel Type





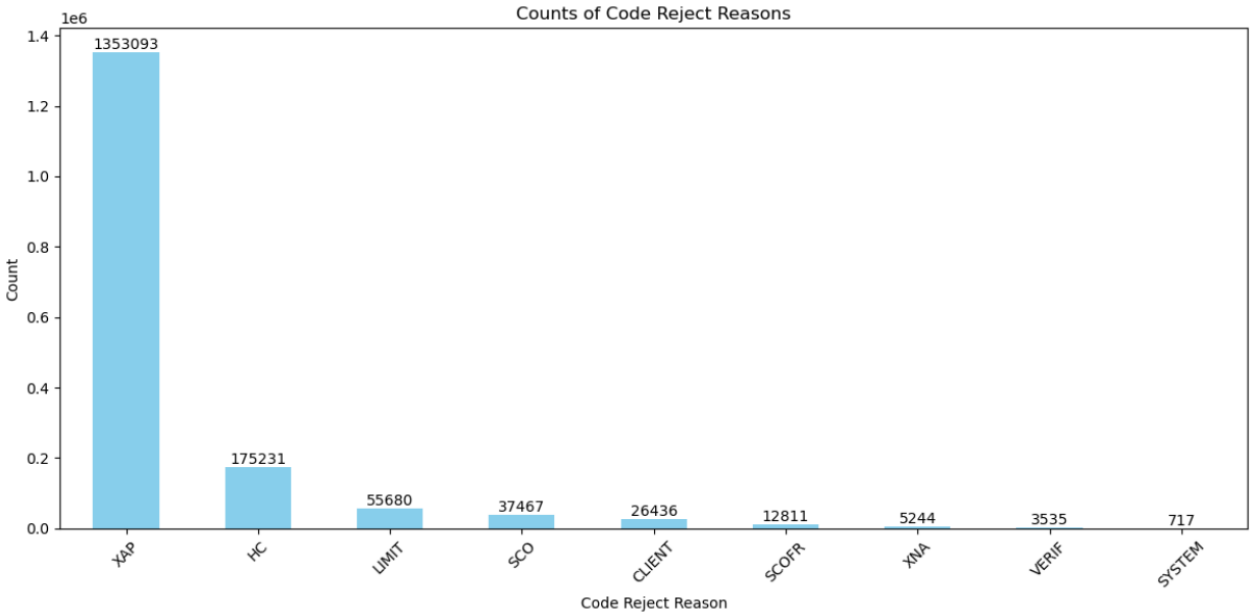
Insight :

- The "Contact center" channel has a significant number of refusals (15,566), suggesting potential issues in service or communication. The "Regional / Local" and "Stone" channels display moderate performance but still have noteworthy refusal counts.

Image source: icons8

Exploratory Data Analysis

Analyzing Contract Status by Channel Type



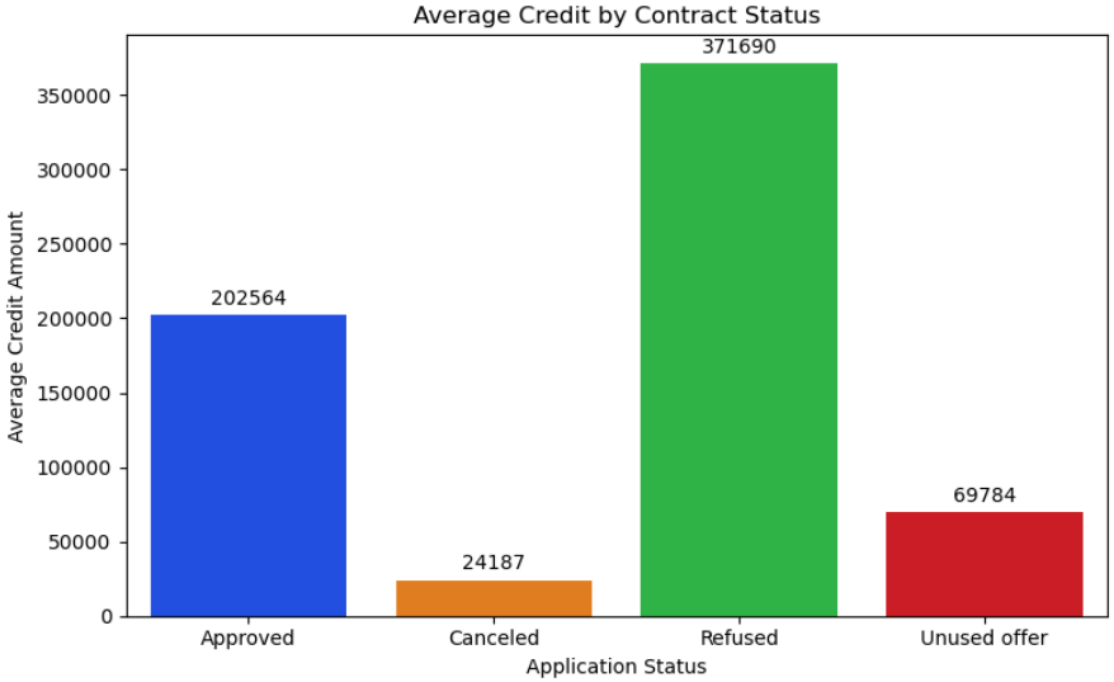
Insight :

- The rejection reason analysis reveals that **"XAP"** accounts for a staggering **80.34%** of total rejections (1,353,093), highlighting it as a critical issue that requires immediate attention to improve approval rates.
- The second most significant reason, **"HC,"** makes up **10.43%** (175,231), suggesting another area where systemic problems may exist.
- The smaller rejection categories, including **"CLIENT" (1.59%), "SCOFR" (0.77%), "XNA" (0.32%), "VERIF" (0.21%), and "SYSTEM" (0.04%), collectively account for a minor portion** of total rejections

Image source: Icons8

Exploratory Data Analysis

Analyzing Credit by Contract Status



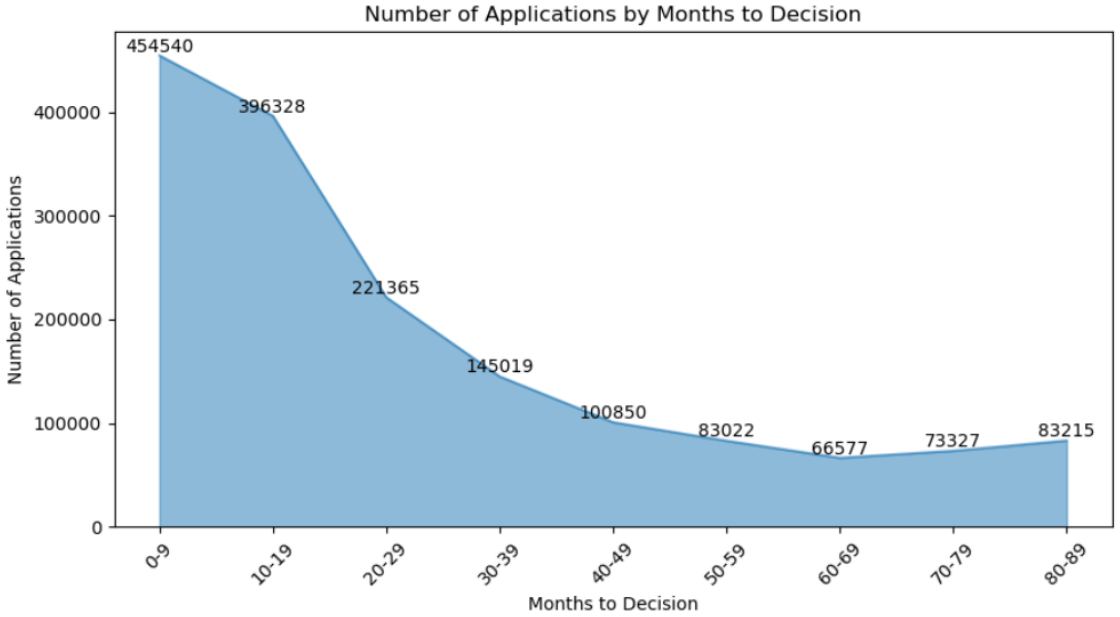
Insight :

- **Approved Contracts:** Average credit amount of 202,564, indicating robust lending for accepted applications.
- **Canceled Contracts:** Significantly lower average of 24,187, suggesting smaller loans may be abandoned.
- **Refused Contracts:** Highest average at 371,690, indicating larger loan requests often get denied.
- **Lending Criteria:** Strict criteria may lead to a mismatch between borrower requests and qualifications.
- **Unused Offers:** Average of 69,784 suggests moderate value, indicating opportunities for improved conversion strategies.

Image source: Icons8

Exploratory Data Analysis

Analyzing Days to Decide and number of applications



Insight :

- The graph tells us that most of the people had decided apply for a second application within the first 9 months of applying for the first time.
- From the 19th month onwards, there is a significant decline in the number of applications, indicating that with the passage of time, not more borrowers would like to apply for a second loan.

Exploratory Data Analysis

Analyzing Good Category and Contract Status

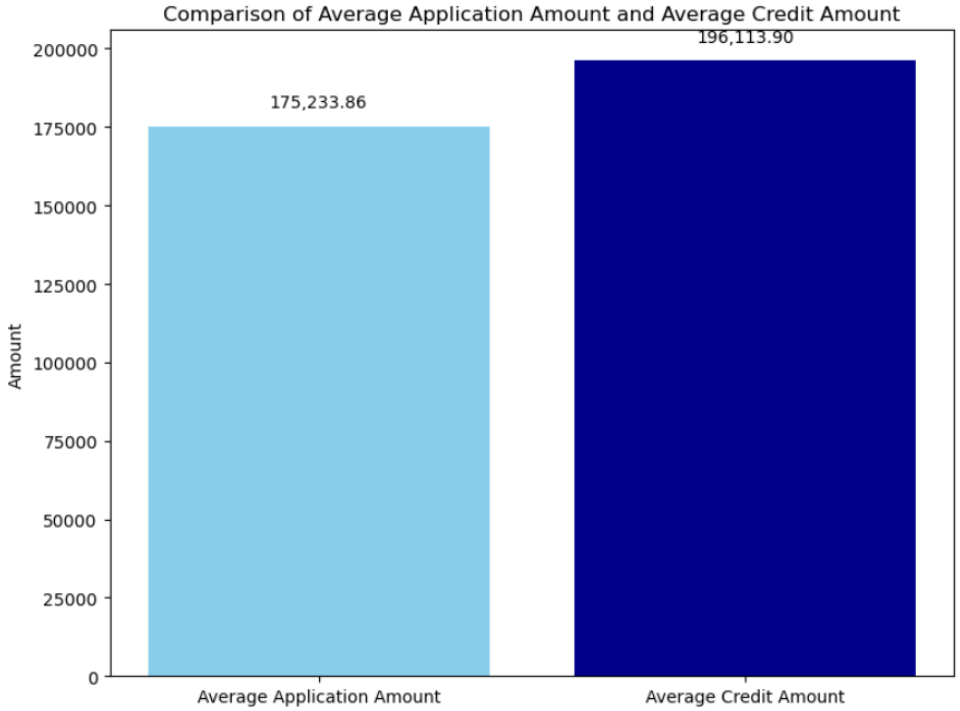
NAME_GOODS_CATEGORY/ NAME_CONTRACT_STATUS	Approved	Canceled	Refused	Unused offer
Additional Service	116	0	12	0
Animals	1	0	0	0
Audio/Video	89394	32	9080	935
Auto Accessories	6560	2	679	140
Clothing and Accessories	21460	0	2010	84
Computers	88050	35	13534	4150
Construction Materials	22471	7	2454	63
Consumer Electronics	111525	26	9100	925
Direct Sales	372	0	73	1
Education	91	0	16	0
Fitness	207	0	2	0
Furniture	49090	10	4342	214
Gardening	2469	0	189	10
Homewares	4540	0	466	17
House Construction	0	0	1	0
Insurance	52	0	10	2
Jewelry	5679	1	594	16
Medical Supplies	3539	1	301	2
Medicine	1448	0	102	0
Mobile	186174	88	20473	17973
Office Appliances	2082	0	240	11
Other	2432	0	122	0
Photo / Cinema Equipment	21379	8	2277	1357
Sport and Leisure	2718	0	250	13
Tourism	1462	1	191	5
Vehicles	2990	1	365	14
Weapon	70	0	7	0
XNA	410410	316107	223788	504

Insight :

- The analysis reveals that **"Mobile"** (186,174) and **"Computers"** (88,050) have the **highest approval counts**, indicating strong demand in technology-related products.
- Categories like **"Audio/Video"** (9,080 refusals) and **"Computers"** (13,534 refusals) **face significant challenges in meeting approval criteria.**
- **"XNA" shows a high cancellation rate** (316,107), suggesting potential issues that need addressing.

Exploratory Data Analysis

Analyzing the Average Application Amount and Average Credit Amount



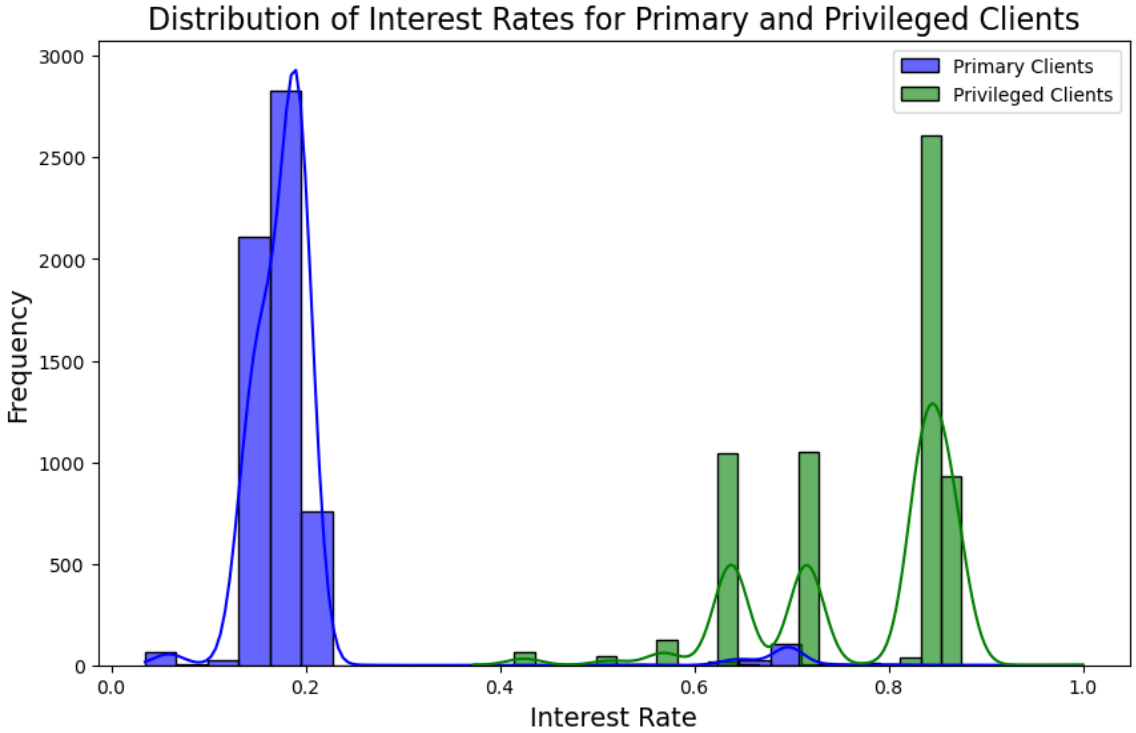
Insight :

- The average application amount is 175,233.86, while the average credit amount is higher at 196,113.90.
- This discrepancy suggests that **people are often asking for more money than they apply for**, which could show that **they feel confident about getting approved** or that **they need more funding than they originally requested.**

Image source: Icons8

Exploratory Data Analysis

Distribution of Interest Rates for Primary and Privileged Clients



Insight:

- Lower Interest for **Primary Clients (maximum 2%)**: Reflects more affordable loan options, likely due to lower loan amounts or better risk profiles.
- Higher Interest for **Privileged Clients (8%+):** Indicates larger loan amounts or specialized products with higher risks, which these clients accept for access to more exclusive terms.

Application Data

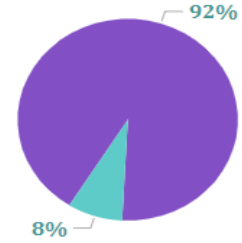
308K	25K	₹ 166K	₹ 169K
Total Application	Defaulted application	AVG Defaulter Income	AVG Repayer income

AGE: All

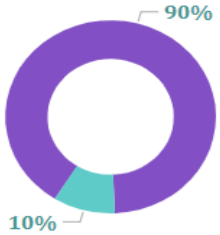
OCCUPATI...: All

EDUCATIO...: All

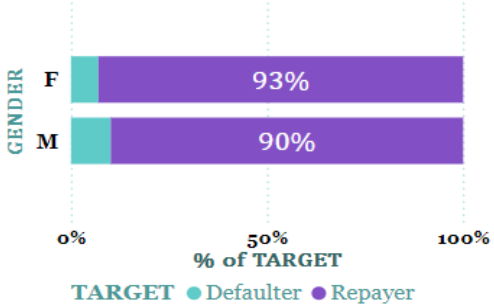
Defaulter Distribution



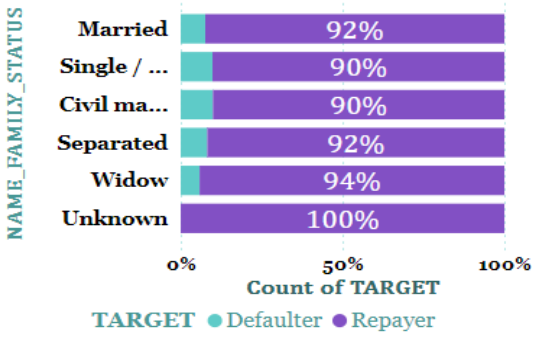
Defaulter Distrubution by Contract Type



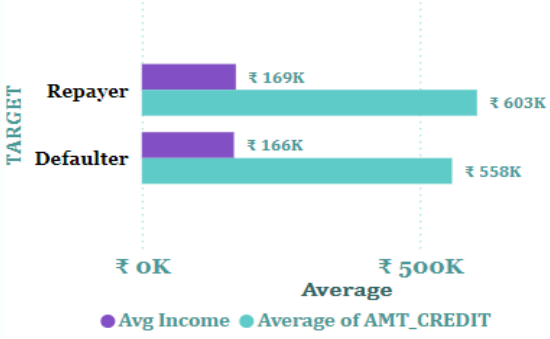
Defaulter Distribution by Gender



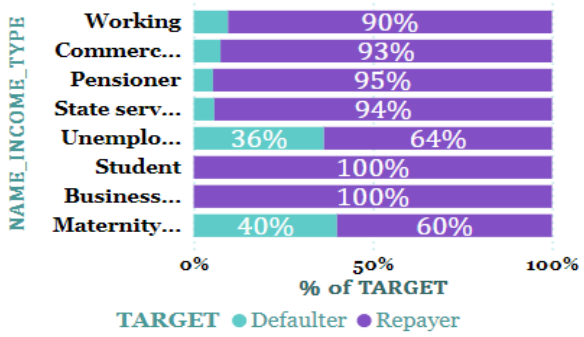
Defaulter Distribution by Family Status



Avg Income and Credit Amount By Defaulter



Defaulter Distribution by Income Type



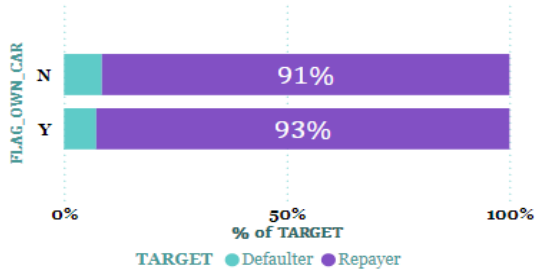
Application Data

₹ 184bn	₹ 8bn	₹ 14bn	4	3
Total Credit Amount	Total Annuity Amount	Defaulted Credit Amount	AVG DTI Rep...	AVG DTI Defa...

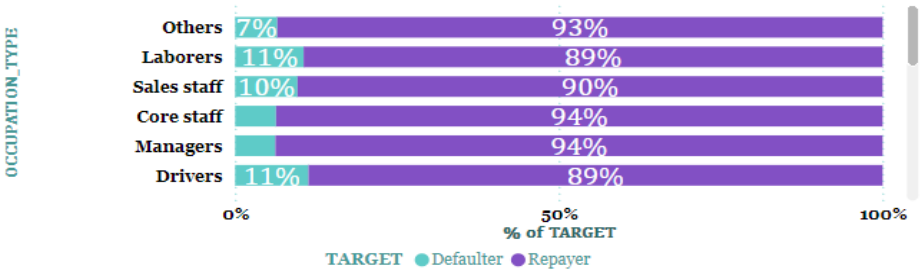
TARGET
All

AGE
All

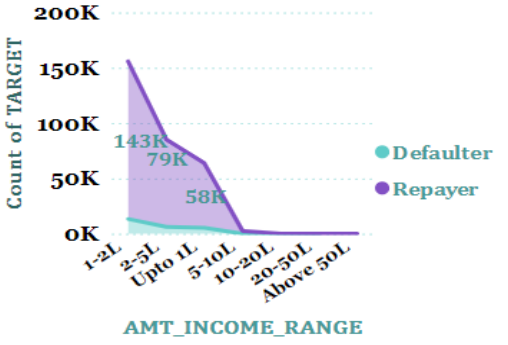
Defaulter Distribution by Car Status



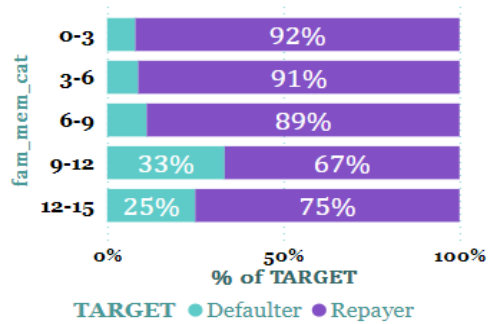
Defaulter Distribution by Occupation Type



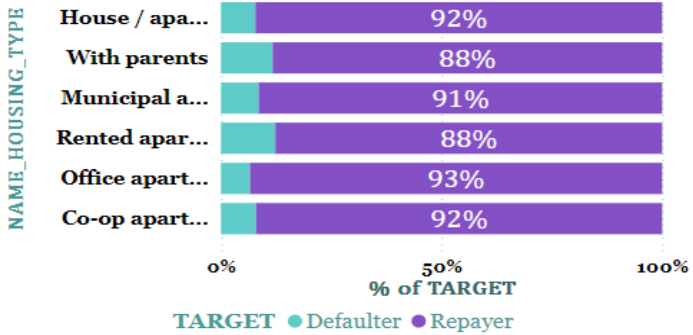
Defaulter Distribution by Income Amount Range



Defaulter Distribution by Family Members



Defaulter Distribution by Housing Type



Previous Data

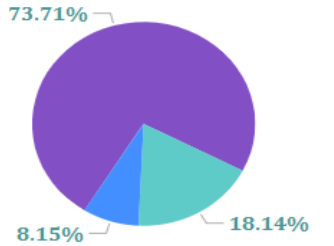
1,049K	₹ 183bn	₹ 204bn	882
Application Count	Total application Amount	Total Credit Amount	Avg Decision Time (Days)

LOAN_STATUS: All

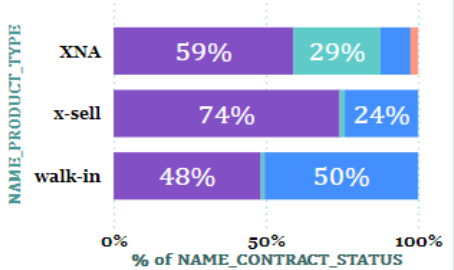
LOAN_PURPOSE: All

CHANNEL_TYPE: All

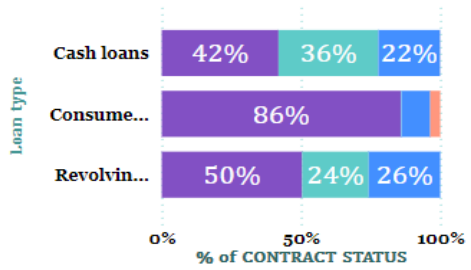
Client Distribution



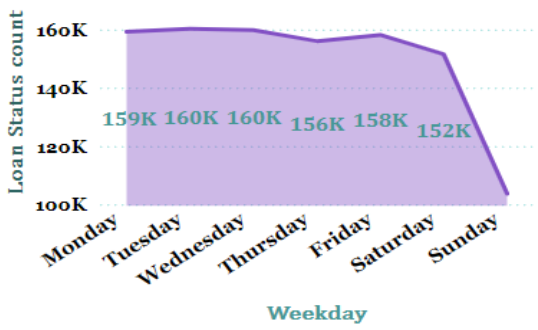
Product Type by Contract Status



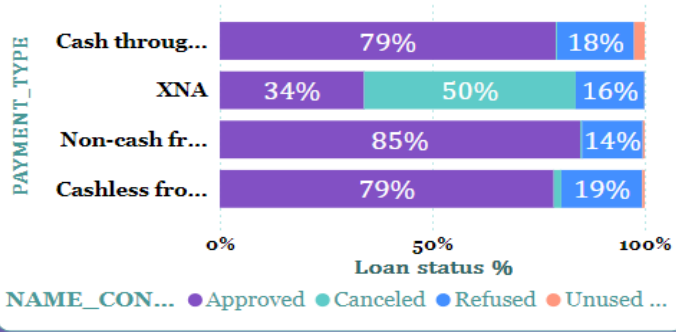
Contract Type by Contract Status



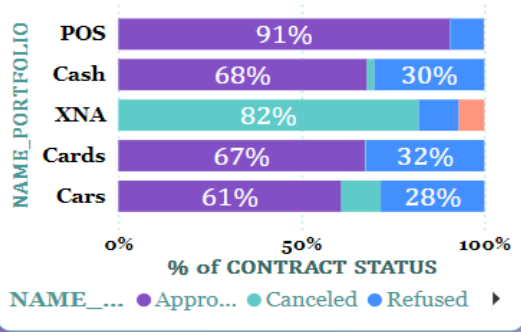
Count of Application on Weekday



Payment Type by Contract Status

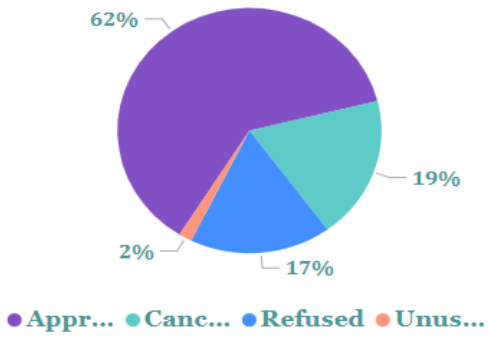


Portfolio by Contract Status

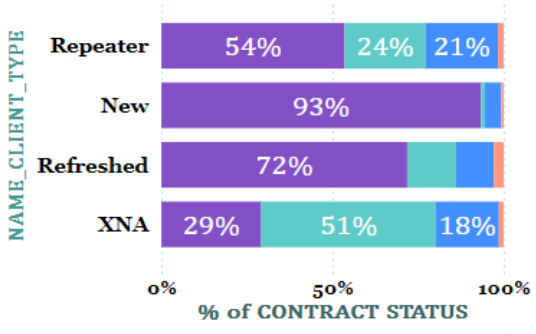


Previous Data

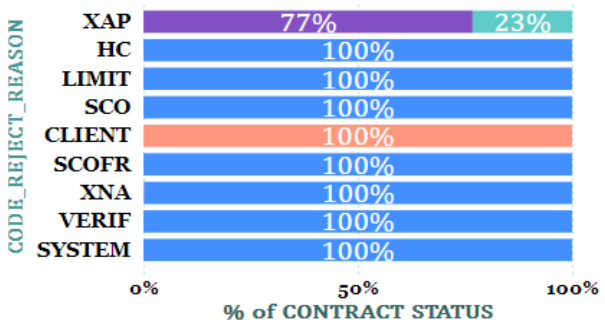
Contract Status Distribution



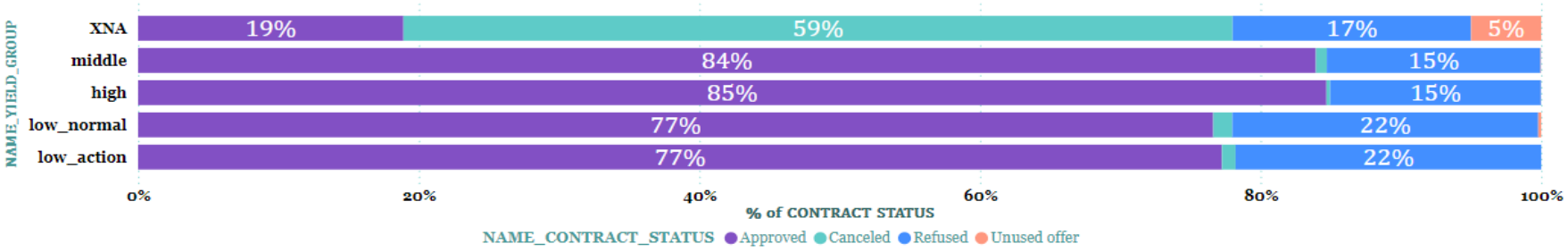
Client type by Contract Status



Reject Code by Contract Status

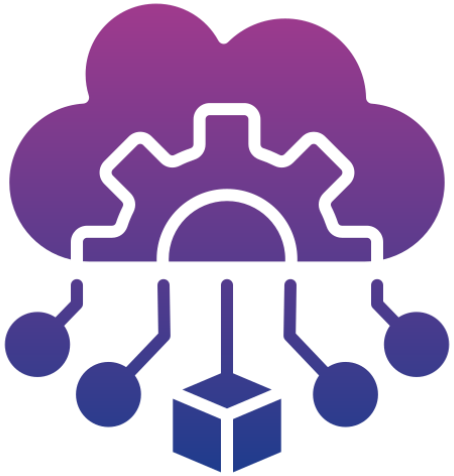


Interest Rate Group by Contract Status



STEPS THAT CAN BE TAKEN TO REDUCE DEFAUTERS AND INCREASE INCOME

- As a higher percentage of female applicants (7.51% or 14,170 out of 188,282) are non-defaulters compared to males (11.29% or 10,655 out of 94,404), banks can develop targeted marketing strategies specifically for female customers to enhance their loyalty.
- Banks should focus on targeted marketing strategies aimed at the "Working" demographic, which shows high application rates and lower default risk.
- The bank can require documentation of stable income from borrowers to verify their ability to afford the loan, thus helping to minimize the risk of defaults.
- By offering financial education and tailored support for vulnerable groups, individuals, can help reduce default rates and also maintain the customer loyality.
- Foster stronger relationships with customers by offering tailored products and services, encouraging them to seek help before defaulting.
- Enhance risk assessment models to better identify potential defaulters before they miss payments, allowing for early intervention.
- Bank can offer incentives for timely payments, such as interest rate reductions or discounts on future loans, to encourage responsible borrowing behavior.



Predictive Modeling for Risk Management in BFSI

An Overview of Predictive Models and their Performance

Agenda

- Objective of the Analysis
- Data Preprocessing and Feature Engineering
- Model Development
- Model Evaluation and Performance
- Insights and Conclusions

Objective

Purpose of the Analysis

- To develop predictive models that accurately identify defaulters and non-defaulters in the dataset.

Key Focus

- Leverage machine learning techniques to improve the accuracy and reliability of predictions.
- Understand the most significant features influencing the prediction of defaulters.

Goal

- Build a range of models to compare performance, identify strengths, and determine the most effective approach for predicting loan defaults.

Model Building Process

Data Preprocessing:

- The data was cleaned, missing values were handled, and outliers were treated to ensure a high-quality dataset.

Feature Engineering:

- Domain knowledge and correlation analysis were applied to select the most relevant features, enhancing model accuracy.

Dimensionality Reduction:

- PCA was used to reduce the feature space, improving computation speed while retaining significant variance in the data.

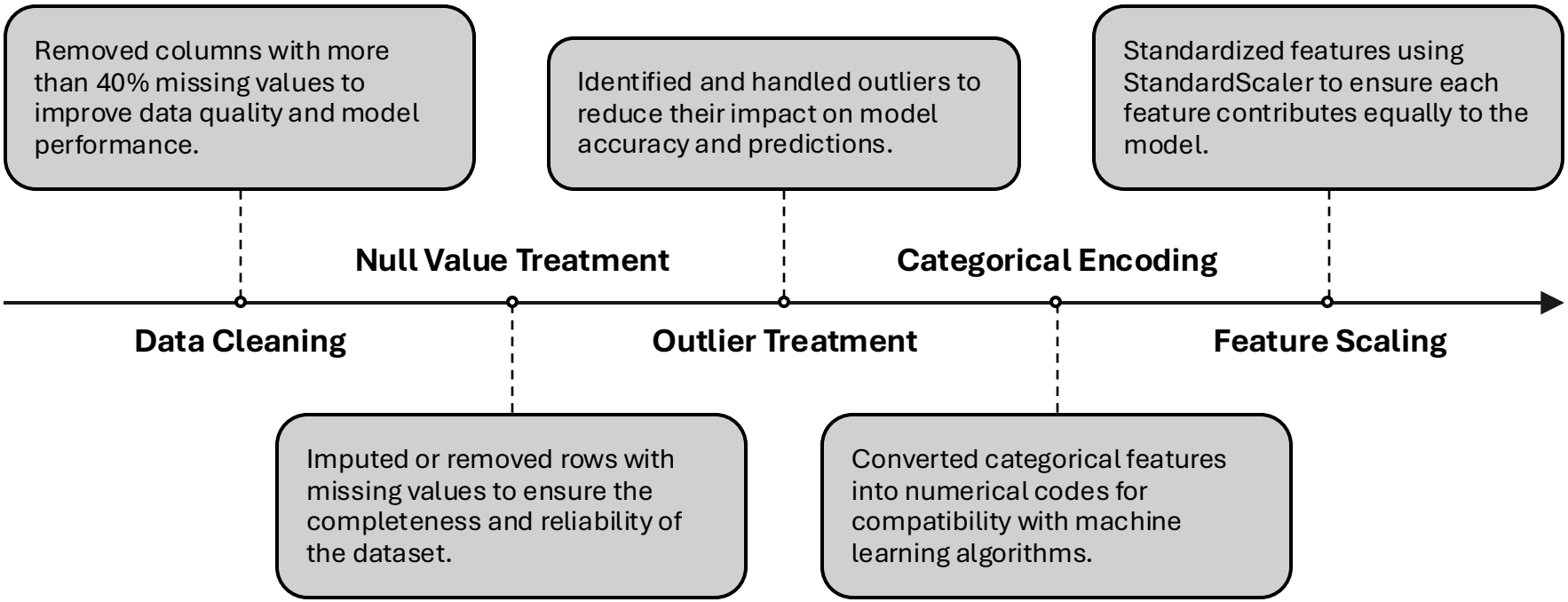
Handling Imbalanced Data:

- Techniques like down-sampling and the application of class weights were considered to balance the dataset for fair model training.

Algorithm Selection:

- Models including Random Forest and Logistic Regression were chosen for their robustness, interpretability, and ability to handle high-dimensional data effectively.

Data Preparation



Commonly Used Algorithms for Classification

Logistic Regression

- **Advantages:**
  - Simple and easy to implement.
  - Interpretable results with coefficients indicating feature impact.
  - Effective for binary classification problems.
- **Disadvantages**:
  - Assumes a linear relationship between features and the target variable.
  - Struggles with complex relationships and high-dimensional data.

Decision Trees

- **Advantages:**
  - Easy to interpret and visualize.
  - Handles both numerical and categorical data.
  - Requires little data preprocessing.
- **Disadvantages**:
  - Prone to overfitting, especially with deep trees.
  - Sensitive to small changes in data, leading to different splits.

Random Forest

- **Advantages:**
  - Reduces overfitting by averaging multiple trees.
  - Handles high-dimensional data well.
  - Provides feature importance metrics.
- **Disadvantages:**
  - Less interpretable compared to single decision trees.
  - Can be computationally intensive with large datasets.

Support Vector Machines (SVM)

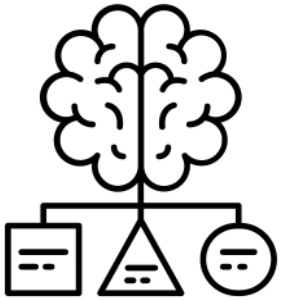
- **Advantages**:
  - Effective in high-dimensional spaces.
  - Works well for both linear and non-linear classification using kernel functions.
- **Disadvantages**:
  - Sensitive to the choice of kernel and regularization parameters.
  - Computationally expensive for large datasets.

K-Nearest Neighbors (KNN)

- **Advantages**:
  - Simple and intuitive, easy to implement.
  - Non-parametric, making no assumptions about data distribution.
- **Disadvantages:**
  - Computationally expensive during prediction, especially with large datasets.
  - Sensitive to irrelevant features and the scale of data.

Naive Bayes

- **Advantages:**
  - Fast and efficient for large datasets.
  - Performs well with categorical features and in text classification.
- **Disadvantages:**
  - Assumes independence among features, which may not hold in practice.
  - Less effective with small datasets.



Predictive Models Overview

Model 1: Random Forest Classifier (Initial Model)

- Trained using all features to establish a baseline performance.

Model 2: Random Forest Classifier +Domain Knowledge

- Reduced feature set chosen based on domain expertise for improved efficiency.

Model 3: Random Forest +Correlated features

- Focused on the most correlated features to target variable for optimized predictions.

Model 4: Random Forest with Grid Search Optimization

- Hyperparameters tuned using Grid Search to enhance model performance.

Model 5: Random Forest + PCA

- Dimensionality reduced via PCA to boost model speed and reduce complexity.

Model 6: Logistic Regression + PCA

- Implemented a simplified model with PCA for better interpretability and classification performance.

Model Evaluation Metrics

		Predicted values		
		True	False	
Actual	True	True Positive (TP)	False Negative (FN) Type 1 Error	$Recall = Sensitivity = \frac{TP}{TP+FN}$
	False	False Positive (FP) Type 1 Error	True Negative (TN)	$Specificity = \frac{TN}{TN+FP}$
		$Precision = \frac{TP}{TP+FP}$		$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ $F1 = \frac{2\,x\ Precision\ x\ Recall}{Precision + Recall}$

B. Classification Model Results

Accuracy:

- **Definition:** The percentage of total predictions that were correct.
- **Example:** If a model predicts 90 out of 100 applicants correctly (whether they will default or not), the accuracy is 90%.

Precision:

- **Definition:** The proportion of true positive predictions among all positive predictions (i.e., correctly identified defaulters).
- **Example:** If the model predicts 40 applicants as defaulters, but only 30 actually defaulted, precision is 75% (30/40).

Recall (Sensitivity):

- **Definition:** The proportion of true positive predictions among all actual positives (i.e., all actual defaulters).
- **Example:** If there are 50 actual defaulters and the model correctly identifies 30 of them, recall is 60% (30/50).

F1 Score:

- **Definition:** The harmonic mean of precision and recall, providing a balance between the two metrics.
- **Example:** For a precision of 75% and recall of 60%, the F1 Score is about 66.67%.

Support:

- **Definition:** The number of actual occurrences of the positive class (i.e., actual defaulters) in the dataset.
- **Example:** If there are 100 actual defaulters in the dataset, the support for this class is 100.

Model Evaluation Comparison Summary

Model	Features Used	Technique	Accuracy Score	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)
Model 1	All features	Random Forest	68.24%	0.68	0.69	0.74	0.62
Model 2	Domain Knowledge Features	Random Forest	61.34%	0.61	0.62	0.69	0.53
Model 3	Top Correlated Features	Random Forest	60.77%	0.61	0.61	0.69	0.52
Model 4	All Features	Random Forest (Grid Search)	68.24%	0.68	0.69	0.74	0.62
Model 5	PCA Components (27 components)	Random Forest + PCA	67.45%	0.67	0.68	0.73	0.61
Model 6	PCA Components (27 components)	Logistic Regression + PCA	68.51%	0.68	0.69	0.73	0.64

Best Model: Model 6

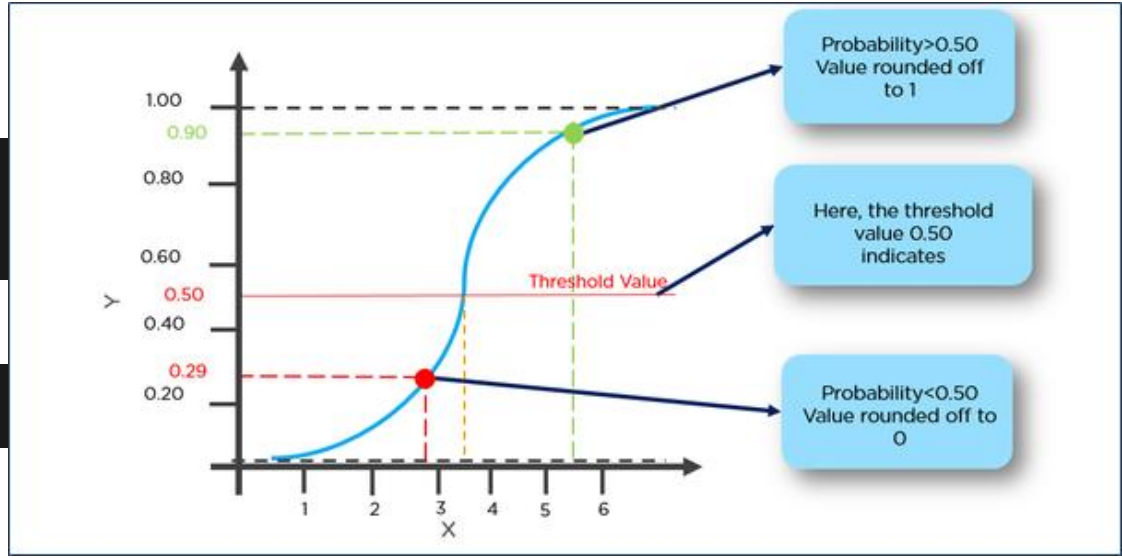
- **Higher Accuracy**: Model 6 achieved the highest accuracy (68.51%), indicating it provides the best generalization for this classification task.
- **Improved Recall for Class 1 (Defaulters)**: Model 6 has better recall for the defaulter class (64%), which is crucial in risk or default detection models, where it is essential to identify defaulters accurately.
- **Simplicity**: Logistic Regression with PCA offers a simpler interpretation due to the reduced number of dimensions, while still capturing 95% of the variance in the data.

Mathematical Equation of the Logistic Model:

Sigmoid Function:

$$P(Y=1|X) = \sigma(z) = \frac{1}{1+e^{-z}}$$

$$L(z) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$$



Mathematical Equation of the Logistic Model:

Logit(p) = (-0.100)*NAME_CONTRACT_TYPE + (0.207)*CODE_GENDER + (-0.133)*FLAG_OWN_CAR + (0.022)*FLAG_OWN_REALTY + (-0.007)*CNT_CHILDREN + (-0.057)*AMT_INCOME_TOTAL + (-0.037)*AMT_CREDIT + (0.073)*AMT_ANNUITY + (-0.008)*NAME_TYPE_SUITE + (0.034)*NAME_INCOME_TYPE + (0.187)*NAME_EDUCATION_TYPE + (0.026)*NAME_FAMILY_STATUS + (0.026)*NAME_HOUSING_TYPE + (0.022)*REGION_POPULATION_RELATIVE + (-0.016)*DAYS_BIRTH + (-0.184)*DAYS_EMPLOYED + (-0.016)*DAYS_REGISTRATION + (-0.037)*DAYS_ID_PUBLISH + (0.000)*FLAG_MOBIL + (0.000)*FLAG_WORK_PHONE + (0.007)*FLAG_CONT_MOBILE + (-0.026)*FLAG_PHONE + (0.000)*FLAG_EMAIL + (0.066)*REGION_RATING_CLIENT + (0.015)*WEEKDAY_APPR_PROCESS_START + (-0.022)*HOUR_APPR_PROCESS_START + (0.000)*REG_REGION_NOT_LIVE_REGION + (0.000)*REG_REGION_NOT_WORK_REGION + (0.000)*REG_CITY_NOT_LIVE_CITY + (0.018)*REG_CITY_NOT_WORK_CITY + (0.031)*ORGANIZATION_TYPE + (-0.426)*EXT_SOURCE_2 + (-0.610)*EXT_SOURCE_3 + (0.019)*OBS_30_CNT_SOCIAL_CIRCLE + (0.000)*DEF_30_CNT_SOCIAL_CIRCLE + (-0.063)*DAYS_LAST_PHONE_CHANGE + (0.000)*FLAG_DOCUMENT_2 + (0.054)*FLAG_DOCUMENT_3 + (0.000)*FLAG_DOCUMENT_4 + (0.000)*FLAG_DOCUMENT_5 + (0.000)*FLAG_DOCUMENT_6 + (0.000)*FLAG_DOCUMENT_7 + (0.000)*FLAG_DOCUMENT_8 + (0.000)*FLAG_DOCUMENT_9 + (0.000)*FLAG_DOCUMENT_10 + (0.000)*FLAG_DOCUMENT_11 + (0.000)*FLAG_DOCUMENT_12 + (0.000)*FLAG_DOCUMENT_13 + (0.000)*FLAG_DOCUMENT_14 + (0.000)*FLAG_DOCUMENT_15 + (0.000)*FLAG_DOCUMENT_16 + (0.000)*FLAG_DOCUMENT_17 + (0.000)*FLAG_DOCUMENT_18 + (0.000)*FLAG_DOCUMENT_19 + (0.000)*FLAG_DOCUMENT_20 + (0.000)*FLAG_DOCUMENT_21 + (0.000)*AMT_REQ_CREDIT_BUREAU_HOUR + (0.000)*AMT_REQ_CREDIT_BUREAU_DAY + (0.000)*AMT_REQ_CREDIT_BUREAU_WEEK + (0.000)*AMT_REQ_CREDIT_BUREAU_MON + (0.000)*AMT_REQ_CREDIT_BUREAU_QRT + (0.028)*AMT_REQ_CREDIT_BUREAU_YEAR + (-0.096)

Top Positive Influences:

Higher values in these features increase the predicted likelihood of default.

Feature	Coefficient
CODE_GENDER	0.207187
NAME_EDUCATION_TYPE	0.186688
AMT_ANNUITY	0.072619
REGION_RATING_CLIENT	0.066265
FLAG_DOCUMENT_3	0.053576
NAME_INCOME_TYPE	0.034388
ORGANIZATION_TYPE	0.031004
AMT_REQ_CREDIT_BUREAU_YEAR	0.028368
NAME_FAMILY_STATUS	0.026147
NAME_HOUSING_TYPE	0.026032
REGION_POPULATION_RELATIVE	0.022004
FLAG_OWN_REALTY	0.021773
OBS_30_CNT_SOCIAL_CIRCLE	0.018871
REG_CITY_NOT_WORK_CITY	0.017515
WEEKDAY_APPR_PROCESS_START	0.015325
FLAG_CONT_MOBILE	0.007133

CODE_GENDER (0.207):

- Specific gender codes increase the likelihood of default

NAME_EDUCATION_TYPE (0.187):

- Higher education levels are linked to higher default risk, possibly due to related debt or career types.

AMT_ANNUITY (0.073):

- Higher loan payments suggest a greater risk of default as debt burden increases.

REGION_RATING_CLIENT (0.066):

- Higher regional ratings imply areas where clients are more prone to default, perhaps due to economic factors.

FLAG_DOCUMENT_3 (0.054):

- Specific document flags may indicate higher risk profiles.

NAME_INCOME_TYPE (0.034):

- Certain income types, like self-employment, show higher default risk due to less predictable income.

ORGANIZATION_TYPE (0.031):

- Employment in particular sectors may relate to higher default risk, likely from job instability.

AMT_REQ_CREDIT_BUREAU_YEAR (0.028):

- More annual credit inquiries suggest potential financial stress, increasing default probability.

NAME_FAMILY_STATUS (0.026):

- Certain marital statuses may correlate with higher risk, reflecting diverse financial behaviors.

NAME_HOUSING_TYPE (0.026):

- Specific housing types may link to higher default rates due to differences in asset stability.

Top Negative Influences:

In these cases, higher values generally indicate financial stability, resulting in a lower likelihood of default.

Feature	Coefficient
EXT_SOURCE_3	-0.609754
EXT_SOURCE_2	-0.426294
DAYS_EMPLOYED	-0.184311
FLAG_OWN_CAR	-0.133061
NAME_CONTRACT_TYPE	-0.100418
DAYS_LAST_PHONE_CHANGE	-0.062597
AMT_INCOME_TOTAL	-0.056781
AMT_CREDIT	-0.037234
DAYS_ID_PUBLISH	-0.03657
FLAG_PHONE	-0.026375
HOUR_APPR_PROCESS_START	-0.02161
DAYS_BIRTH	-0.015774
DAYS_REGISTRATION	-0.015696
NAME_TYPE_SUITE	-0.008454
CNT_CHILDREN	-0.006896

EXT_SOURCE_3 (-0.610):

- Higher scores from external sources are linked to lower default risk, suggesting good financial standing.

EXT_SOURCE_2 (-0.426):

- Higher external credit scores indicate stability, reducing default probability.

DAYS_EMPLOYED (-0.184):

- Longer employment duration is associated with decreased default risk, as stable employment often signals reliable income.

FLAG_OWN_CAR (-0.133):

- Car ownership may reflect financial stability, reducing the chance of default.

NAME_CONTRACT_TYPE (-0.100):

- Certain contract types are less risky, possibly reflecting more secure financial agreements.

DAYS_LAST_PHONE_CHANGE (-0.063):

- Recent phone changes may indicate instability; less frequent changes are associated with lower default rates.

AMT_INCOME_TOTAL (-0.057):

- Higher income correlates with a lower likelihood of default, showing greater ability to manage financial obligations.

AMT_CREDIT (-0.037):

- Higher total credit amounts may correlate with reduced risk, possibly due to effective credit management.

DAYS_ID_PUBLISH (-0.037):

- The stability in ID-related documents suggests a lower risk of default.

FLAG_PHONE (-0.026):

- Consistent phone records might indicate financial stability and responsibility.

Thank you!