SCALETRIX.AI

Risk Analysis – Banking, Financial Services and Insurance

AICPA SOC

HIPAA COMPLIANT

GDPR COMPLIANT

Unleash Growth with Digital Transformation & Analytics

• NEW DELHI • BANGALORE • UNITED STATES

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

2 Tables :

Application Table	
Columns	122
Entries / rows	3,07,511
Primary Key	**SK_ID_CURR**
Contains information of all the current loan applications, including loan id, contract type, credit amount, gender of the applicant, income level, education level, family status, regional population, contact information, real estate ownership information, different documents submitted etc, along with the TARGET (indicator if the applicant can be a potential defaulter)	

Previous Application Table	
Columns	37
Entries / rows	16,70,214
Primary Key	**SK_ID_PREV**
Contains information of all the previous home loan applications, including loan id, contract type, credit amount, status of previous loan, Relative to current application when was the decision about previous application made, Payment method that client chose to pay for the previous application, portfolio type, yield type etc	

Observations made during Data Auditing:

01 Application Data has many columns with more than 40% null values

02 Previous Application Data also has significant number of columns more than 40% null values but less than the Application Data

Data Cleaning and Pre-processing

01 The name of tables as application_data, previous_application_data to keep the consistency in the data while clubbing the analysis

02 Will maintain the null value simulation by keeping the dataset intact with the null value

03 Provide the proper definition of the categorical values to get the better view over the analysis

Image source: icons8

Exploratory Data Analysis

Understanding the demographics

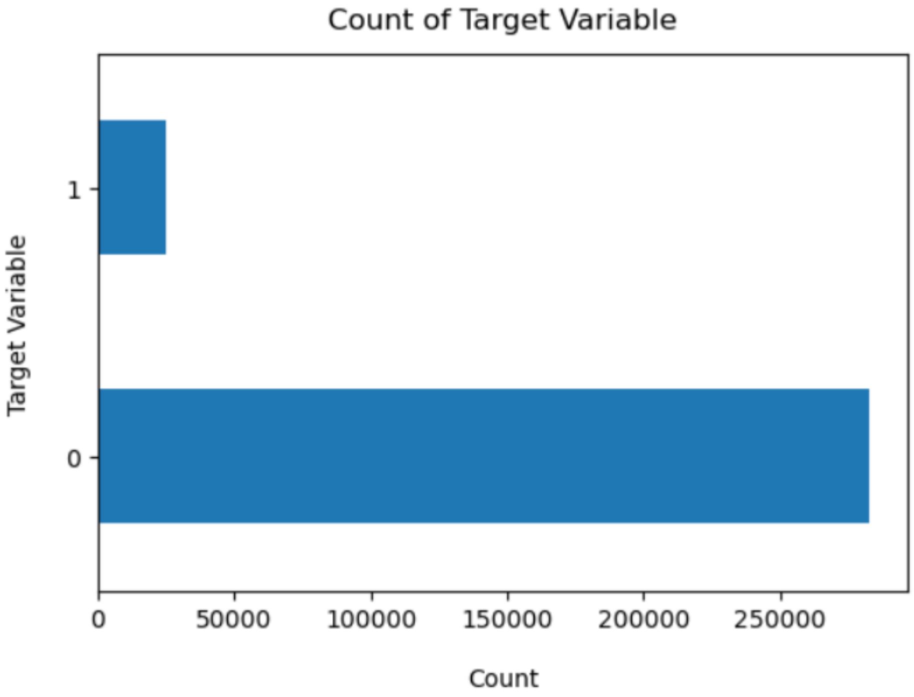
Male Vs Female
Ratio : 1.93

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

Lets explore the TARGET field

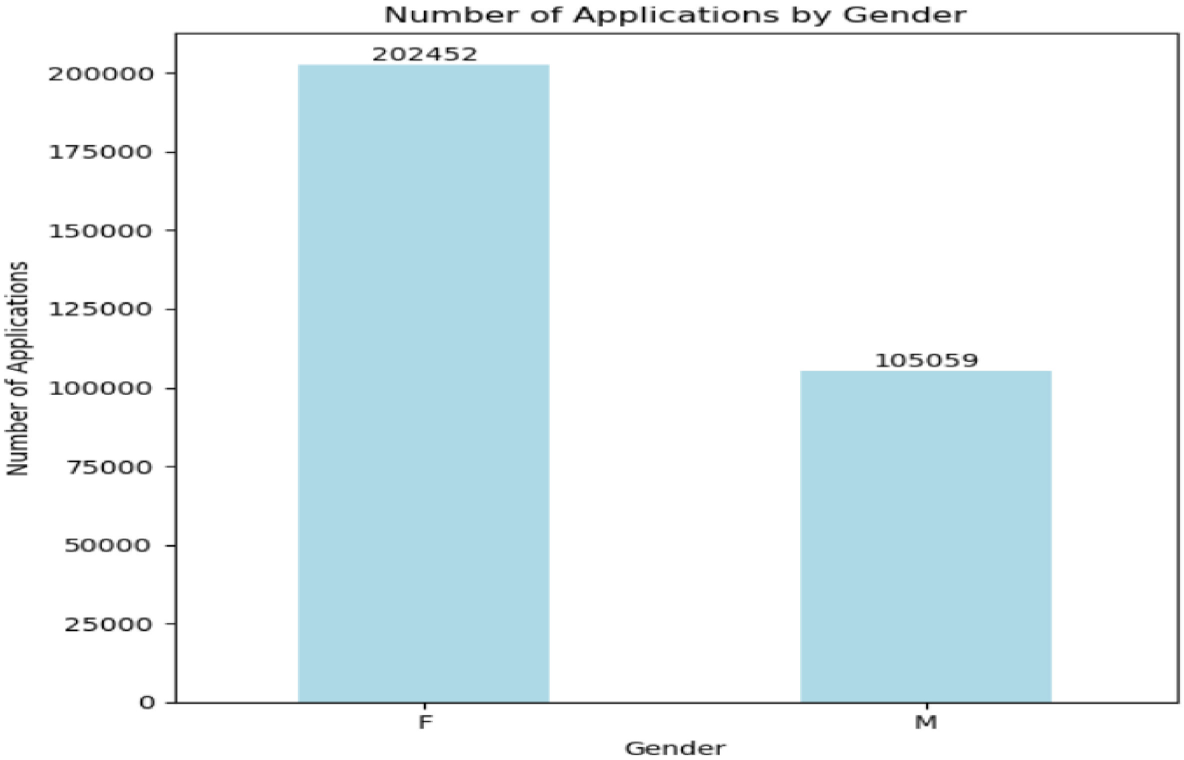


Insight :

- The majority of the individuals in the dataset are non-defaulters, making up approximately **91.3%** (282,686 out of 307,511) of the total sample
- Defaulters are only about **8.7%** 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



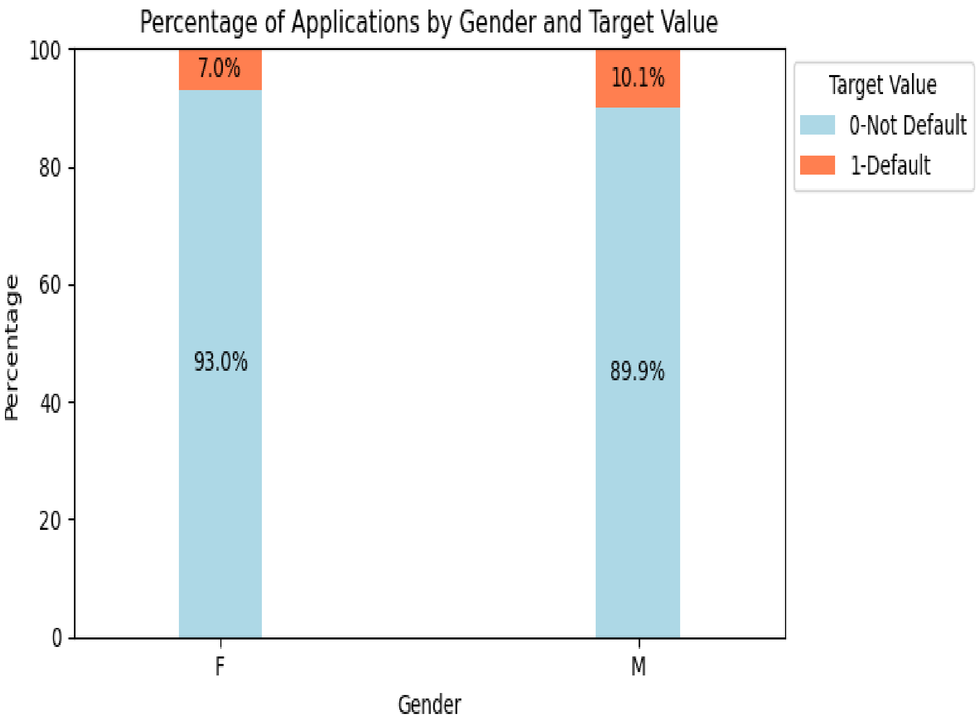
Insight :

- **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.

Image source: icons8

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).

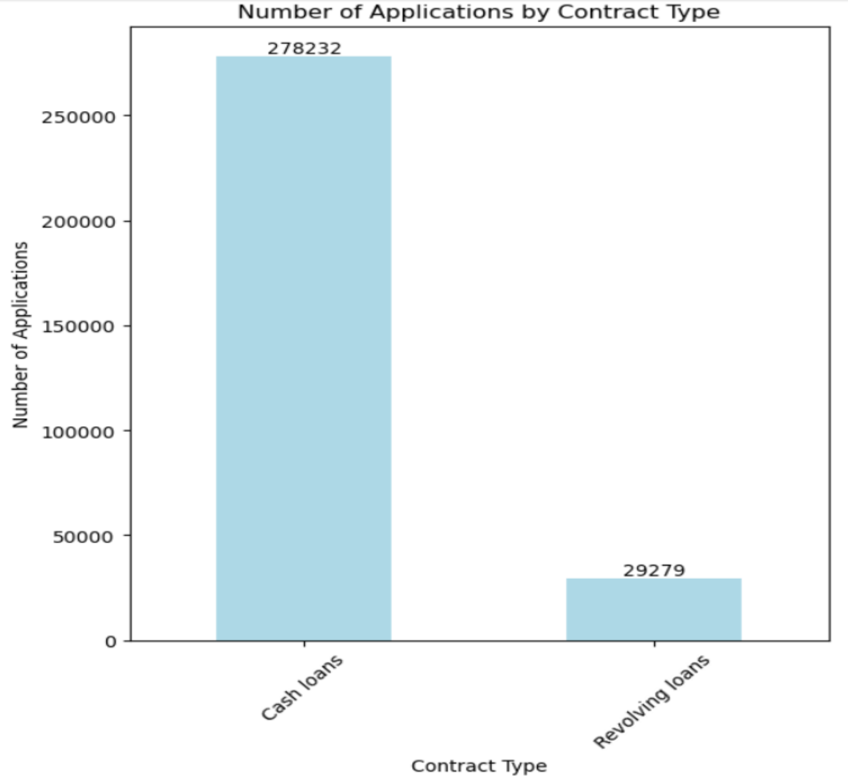
Default Rates:

- 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.

Exploratory Data Analysis

➢ Analyzing the contract type

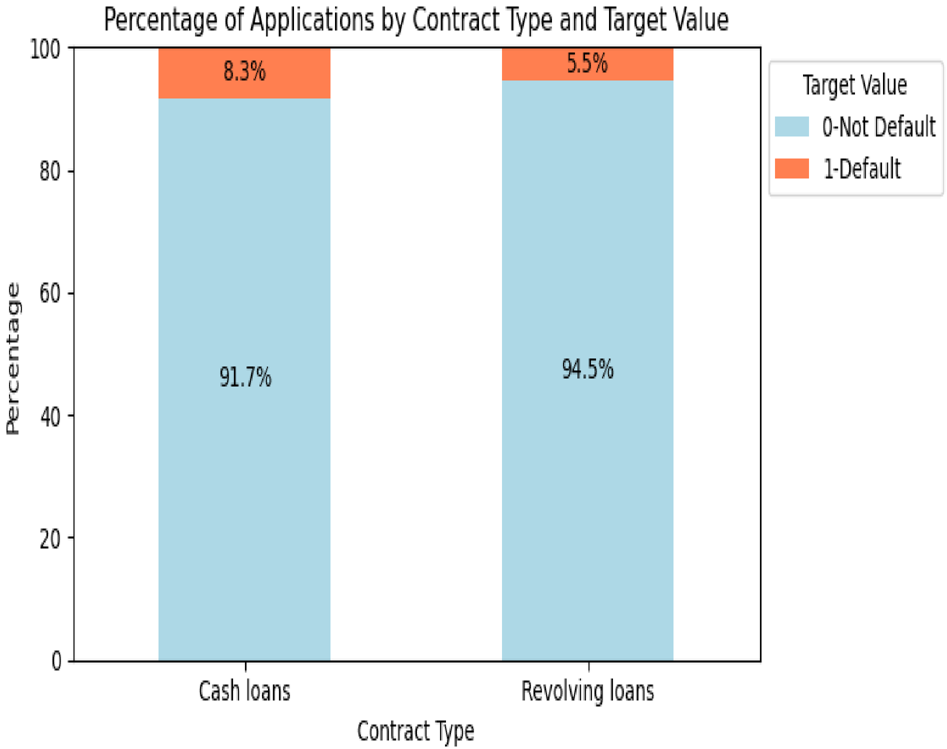


Insight :

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

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).

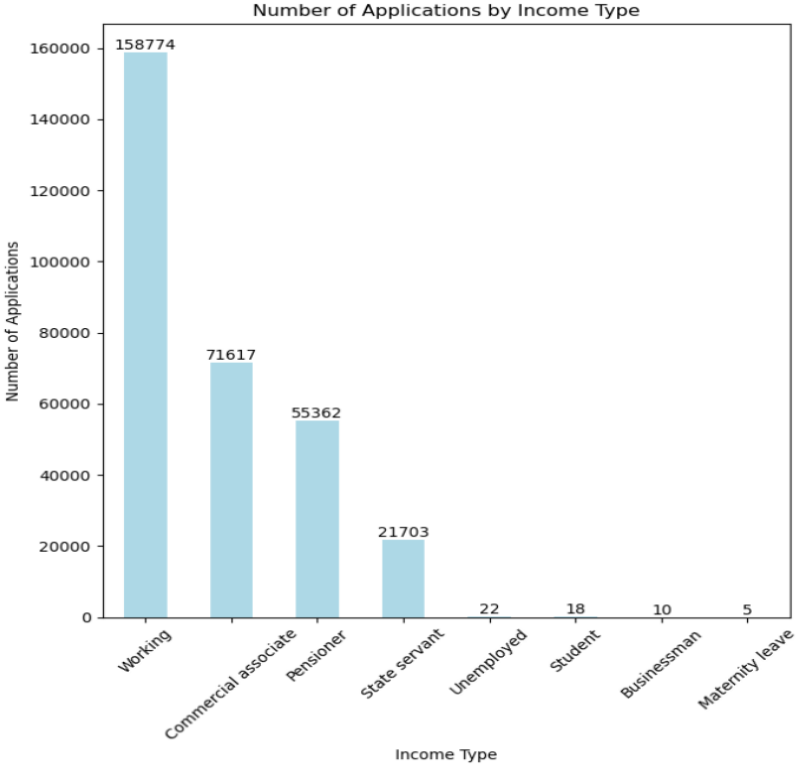
Non-Default Rates by Contract Type:

•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

➢ Lets look at the income type

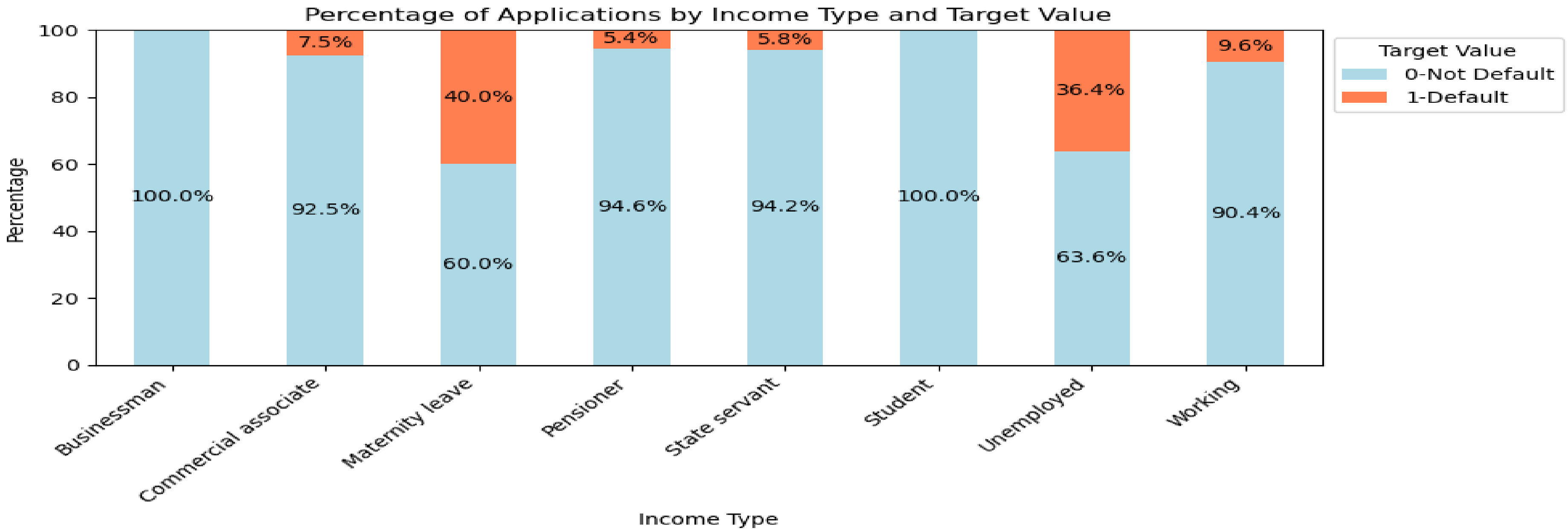


Insight :

- 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.

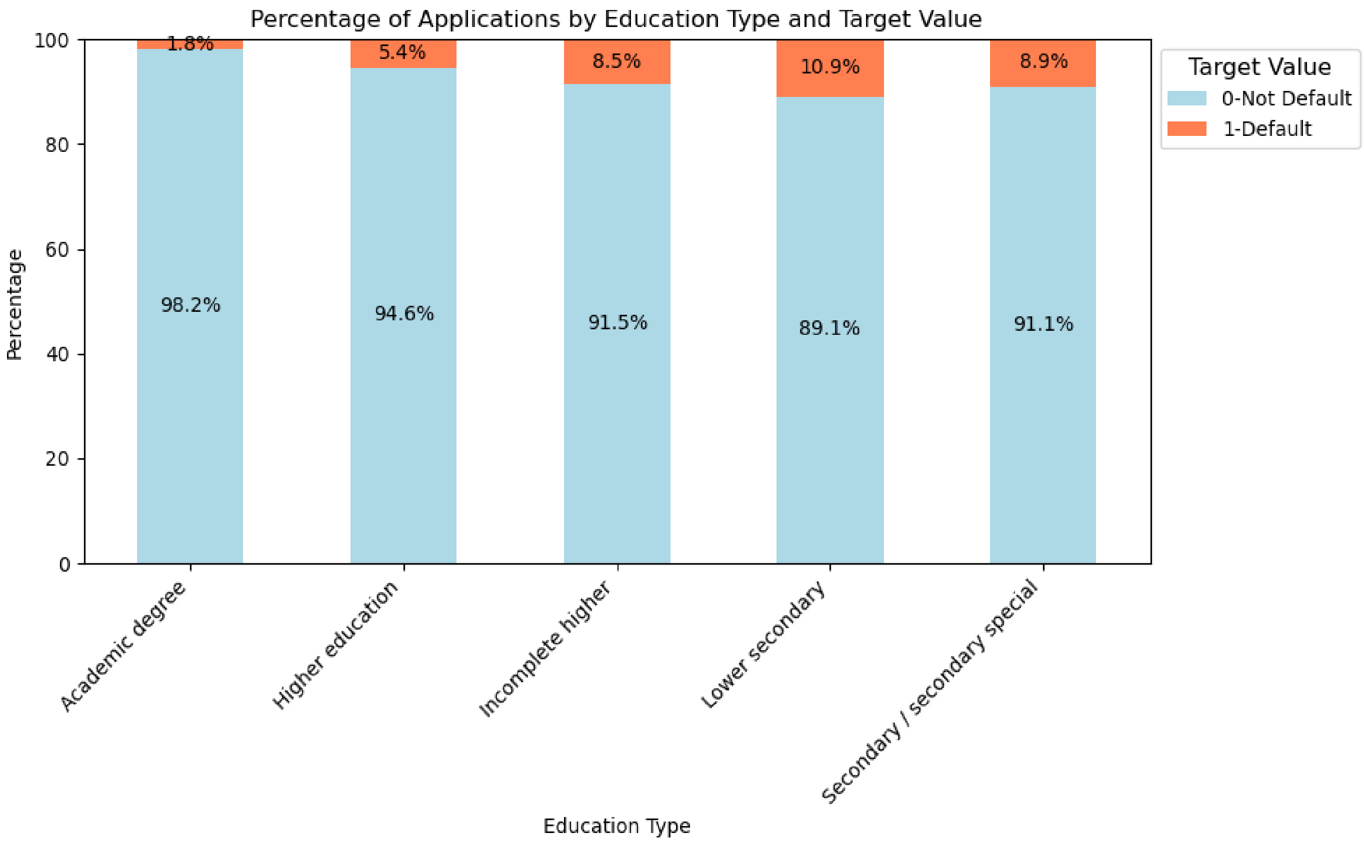
Exploratory Data Analysis

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.

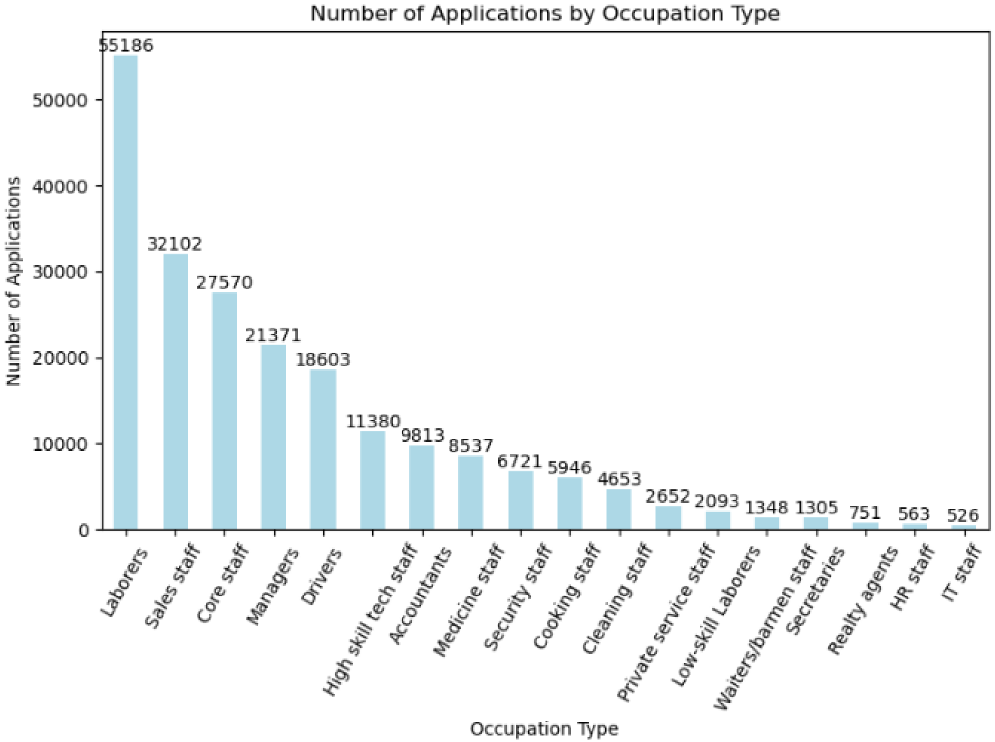
Exploratory Data Analysis ➢ Analyzing the Education Type



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

Exploratory Data Analysis

Analyzing the Occupation Type



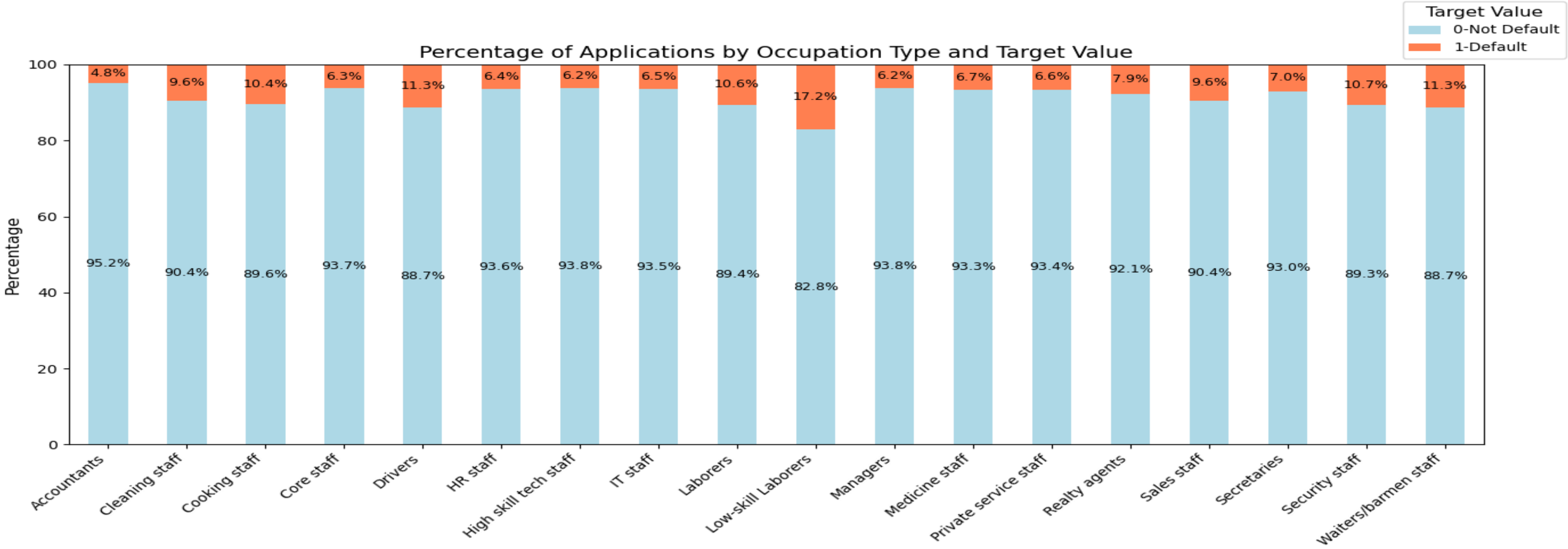
Insight :

- The dataset reveals a diverse range of loan applications across various occupation types, with **Laborers** being the **most prominent** group, representing about **22.4%** of total applications. Other significant segments include Sales Staff and Core Staff, indicating strong interest in loans among lower and middle-skilled workers. While Managers also show notable application numbers, high-skill roles like IT Staff and HR Staff are less represented, possibly reflecting their more stable financial situations.

SCALETRIX.AI

Image source: icons8

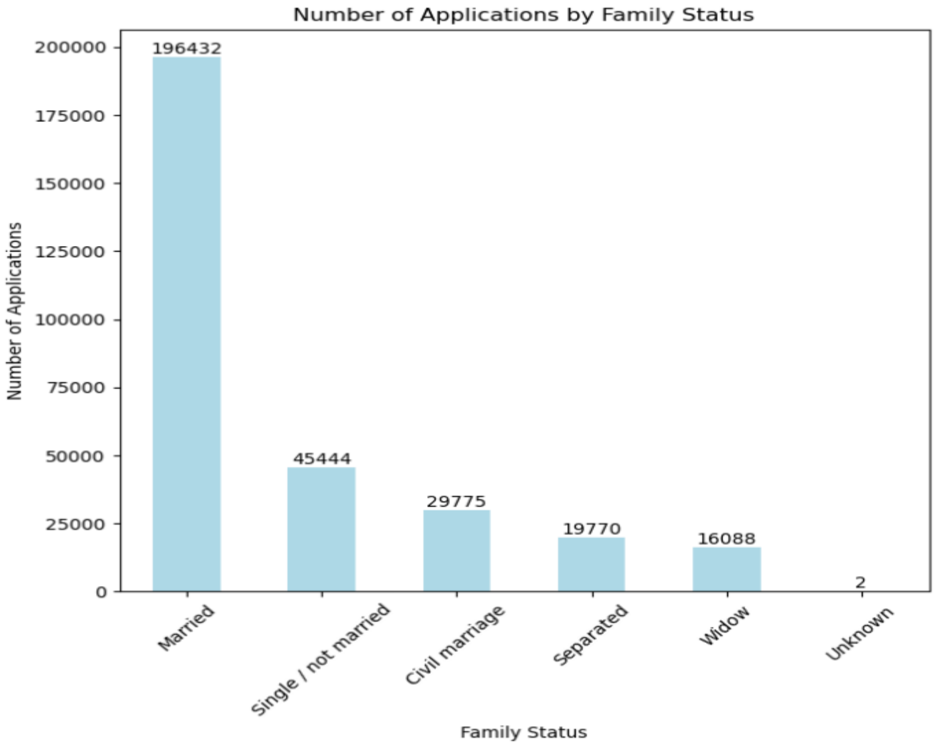
Exploratory Data Analysis ➢ Analyzing the Occupation Type



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.

Exploratory Data Analysis

➢ Analyzing the Family Status

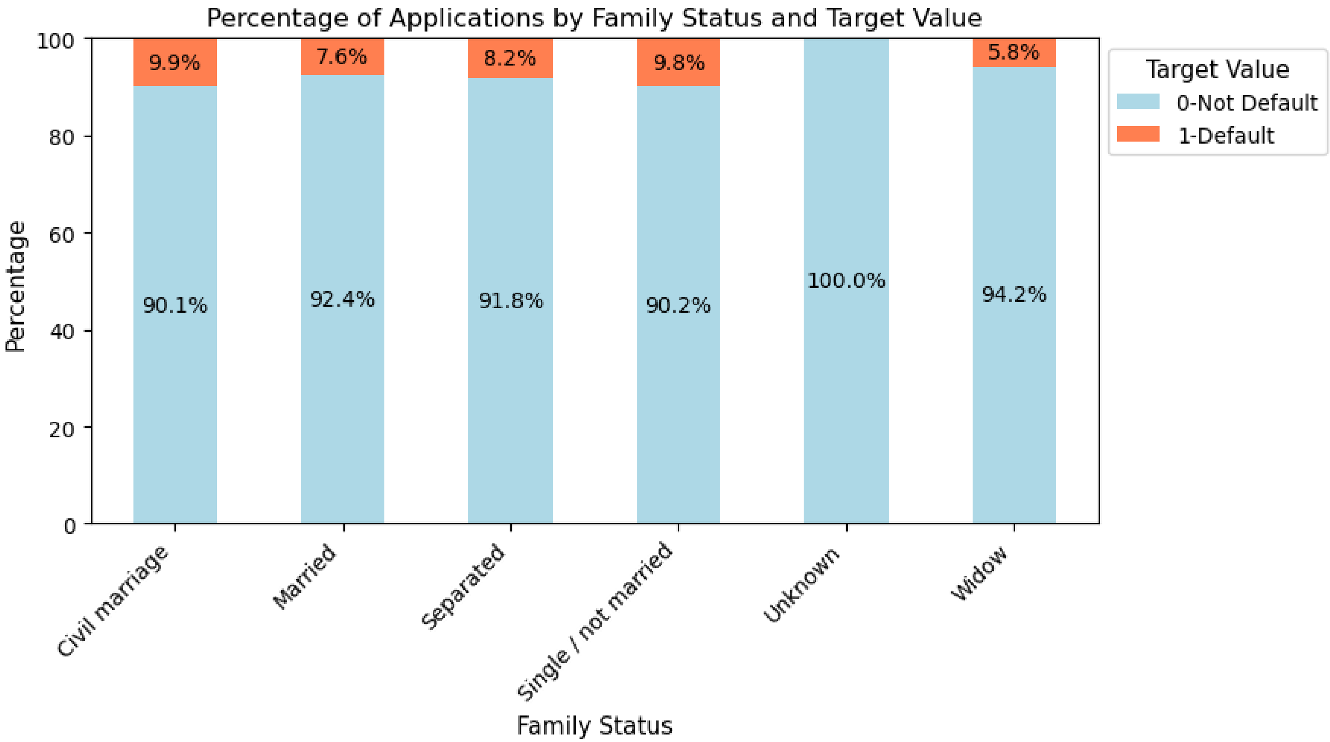


Insight :

- **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

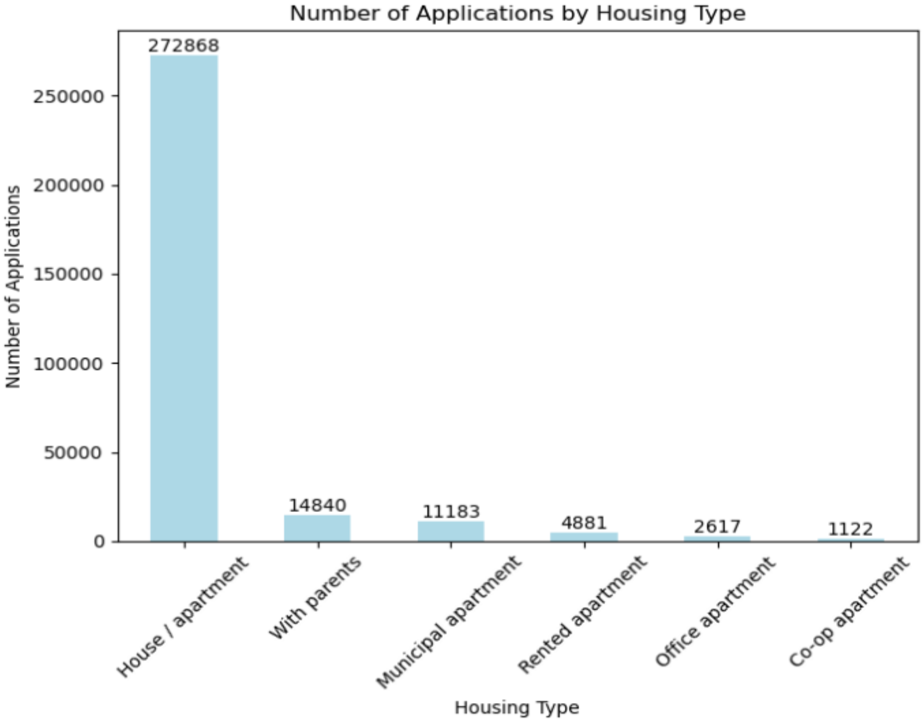


Insight:

- 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%.

Exploratory Data Analysis

Analyzing the Housing Type

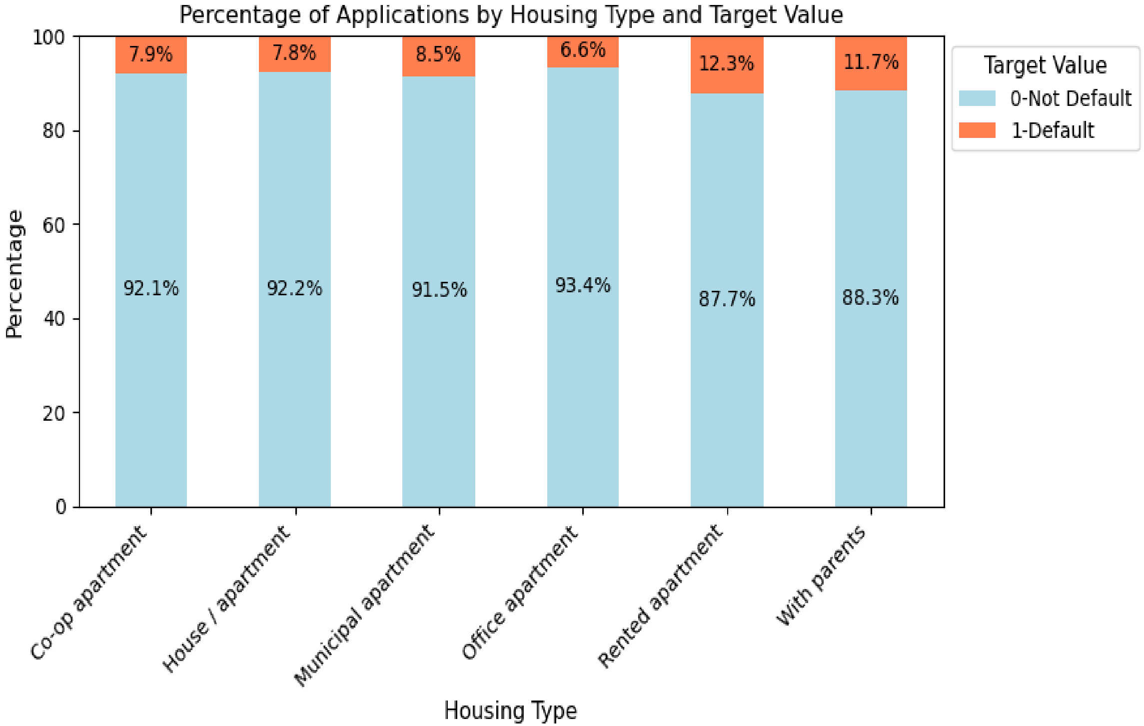


Insight :

- The data on loan applications shows that the **majority, 272,868 applications** (about **70%**), are from individuals in stable housing situations, highlighting a strong link between stable housing and loan-seeking behavior.
- **Rented apartments** account for 4,881 applications (**1.3%**), indicating that renting is less common among loan seekers. **Office apartments** have 2,617 applications (**0.7%**), showing they are a rare choice, while **co-op apartments**, with just 1,122 applications (**0.3%**), attract a specific demographic.

Exploratory Data Analysis

Analyzing the Housing Type

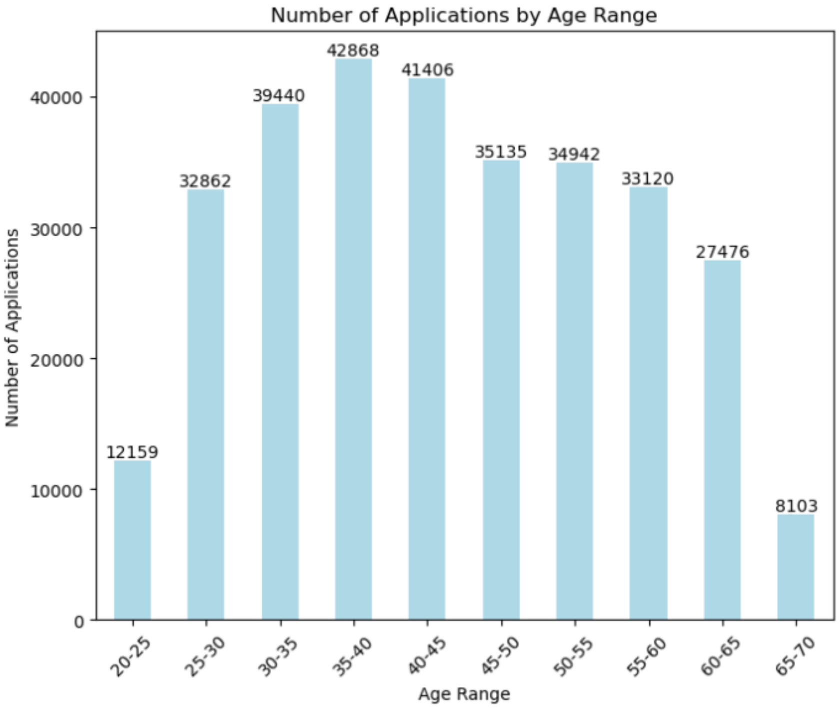


Insight:

- **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).

Exploratory Data Analysis

Analyzing the Age Range



Insight :

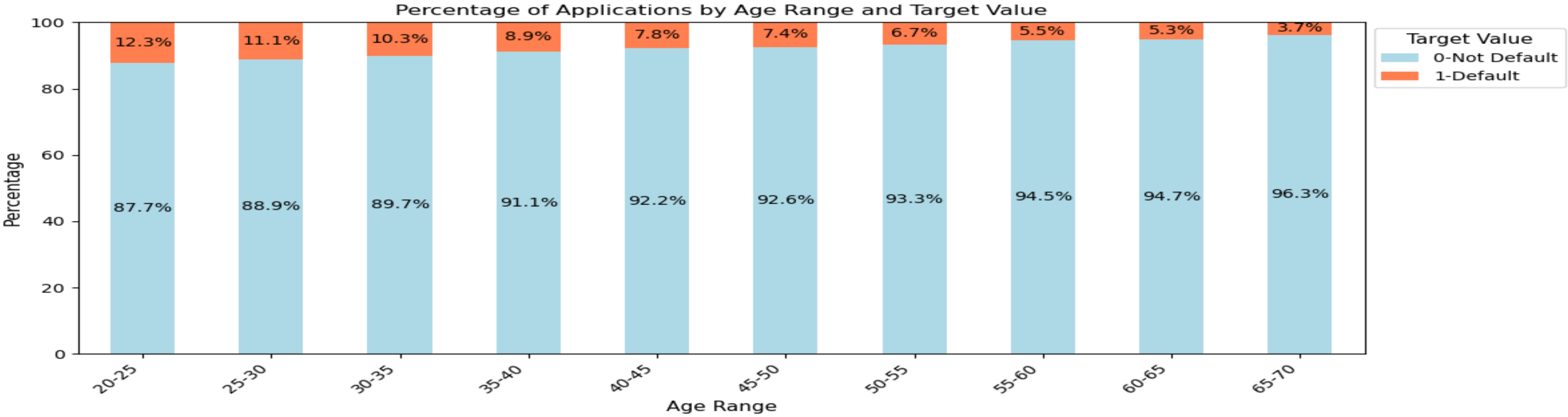
- The **35-40 age group leads** with 42,868 applications, about **16.2%** of the total, The **30-35 age group follows** closely with 39,440 applications (**15.0%**), indicating financial activity related to career and family investments.
- The **25-30 age group** has 32,862 applications, making up around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

SCALETRIX.AI

Image source: icons8

Exploratory Data Analysis

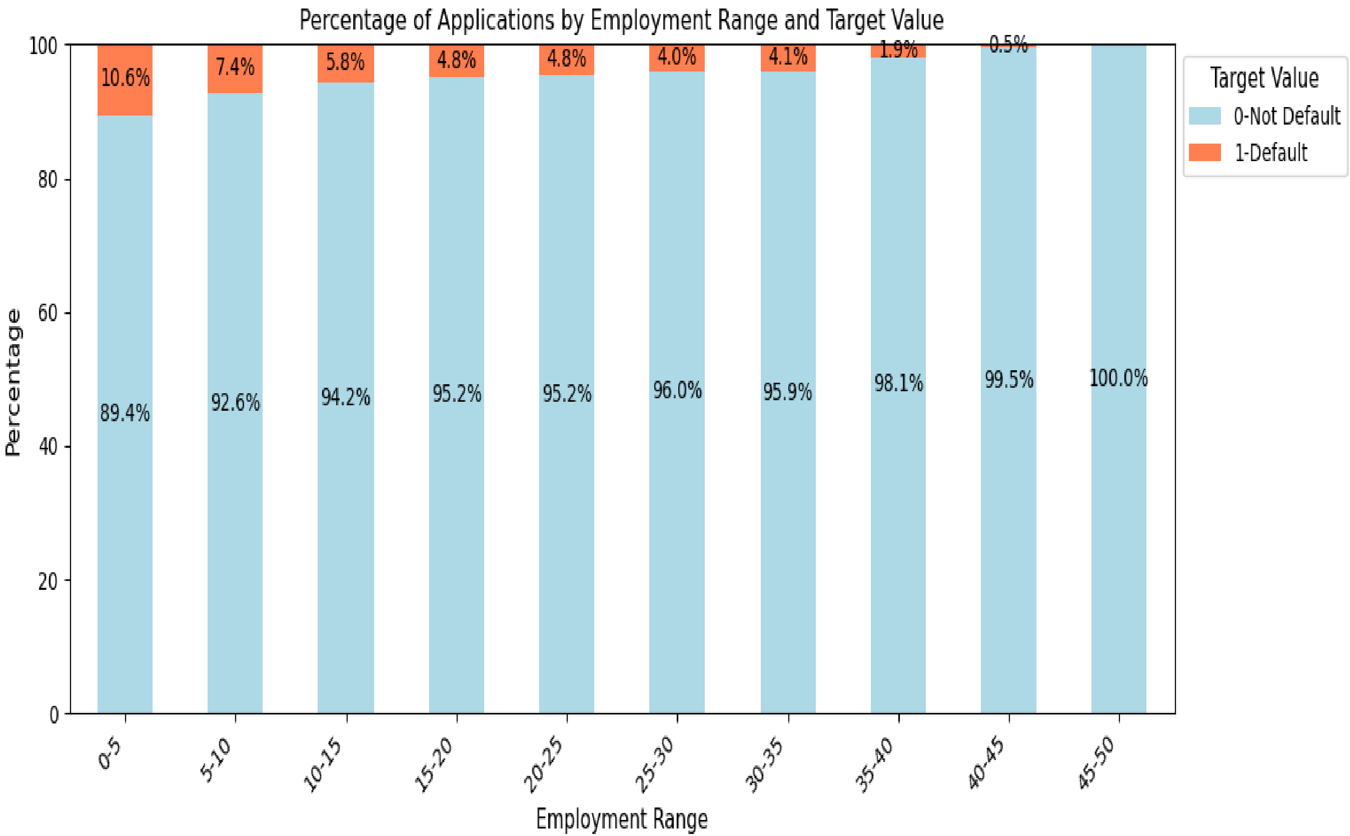
Analyzing the Age Range



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

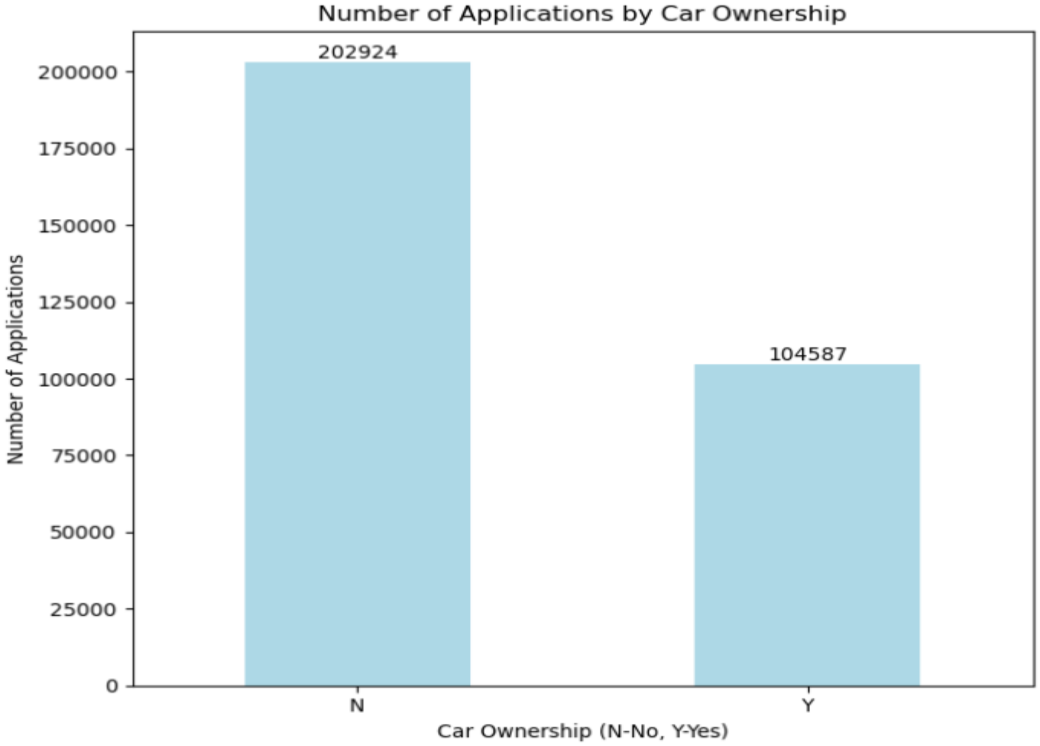


Insight:

- 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.

Exploratory Data Analysis

Analyzing the Car Ownership

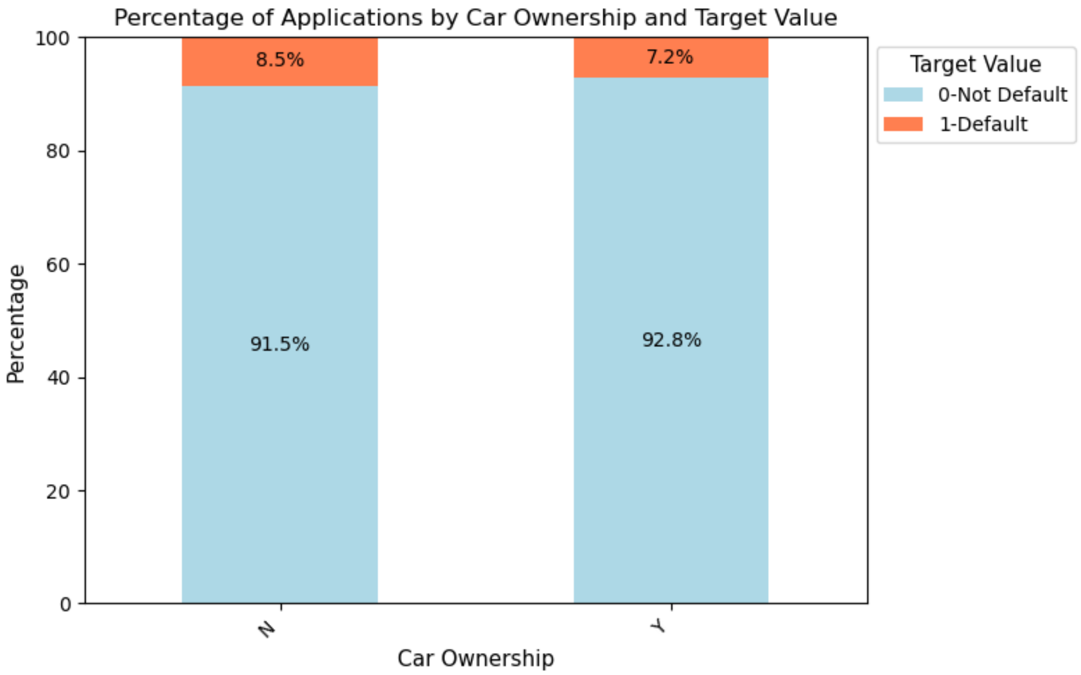


Insight :

- 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.

Exploratory Data Analysis

Analyzing the Car Ownership

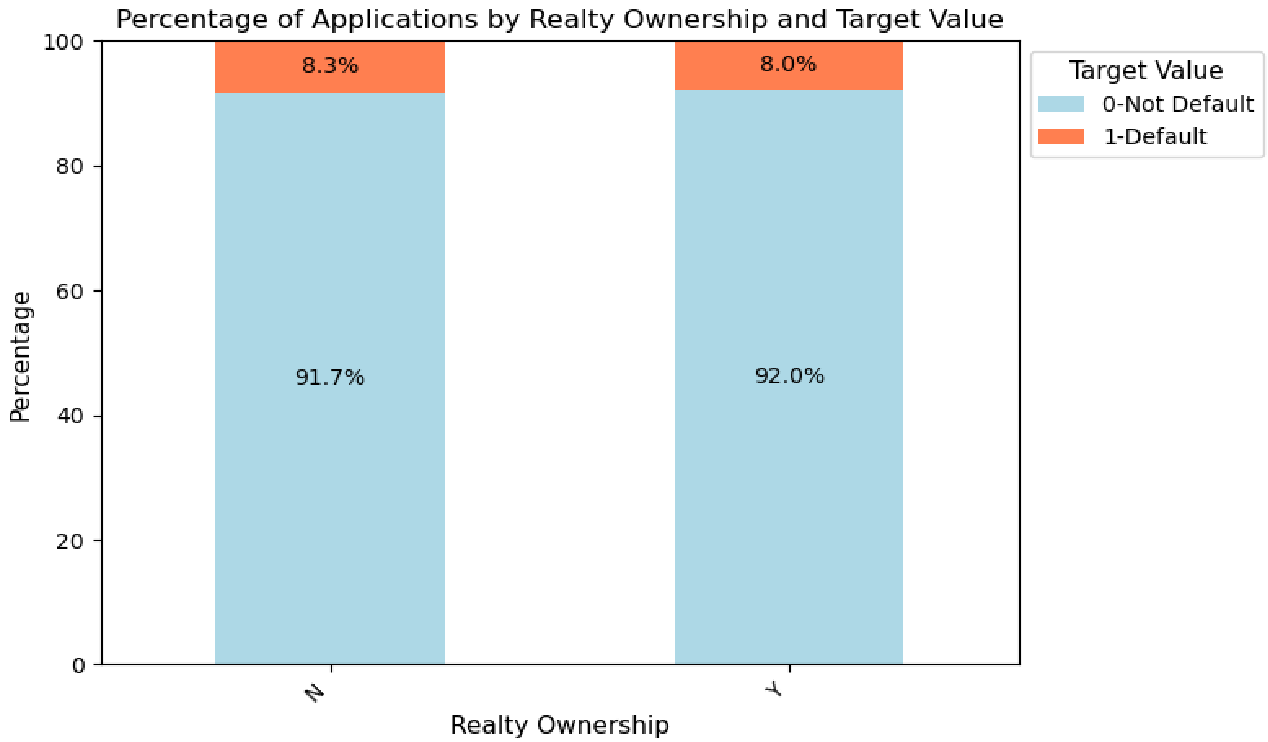


Insight :

- 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.

Exploratory Data Analysis

Analyzing the Reality Ownership

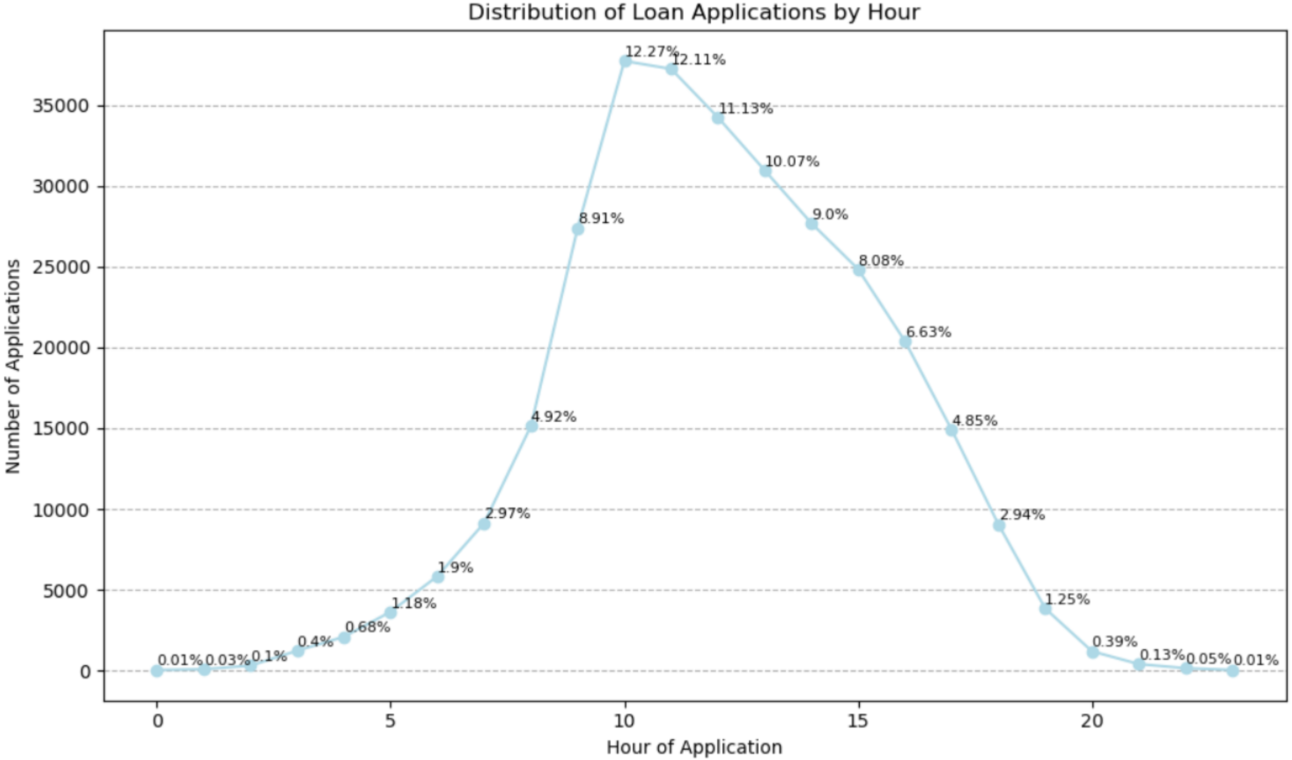


Insight :

- 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.

Exploratory Data Analysis

➢ Analyzing the loan applications by the hour of the day

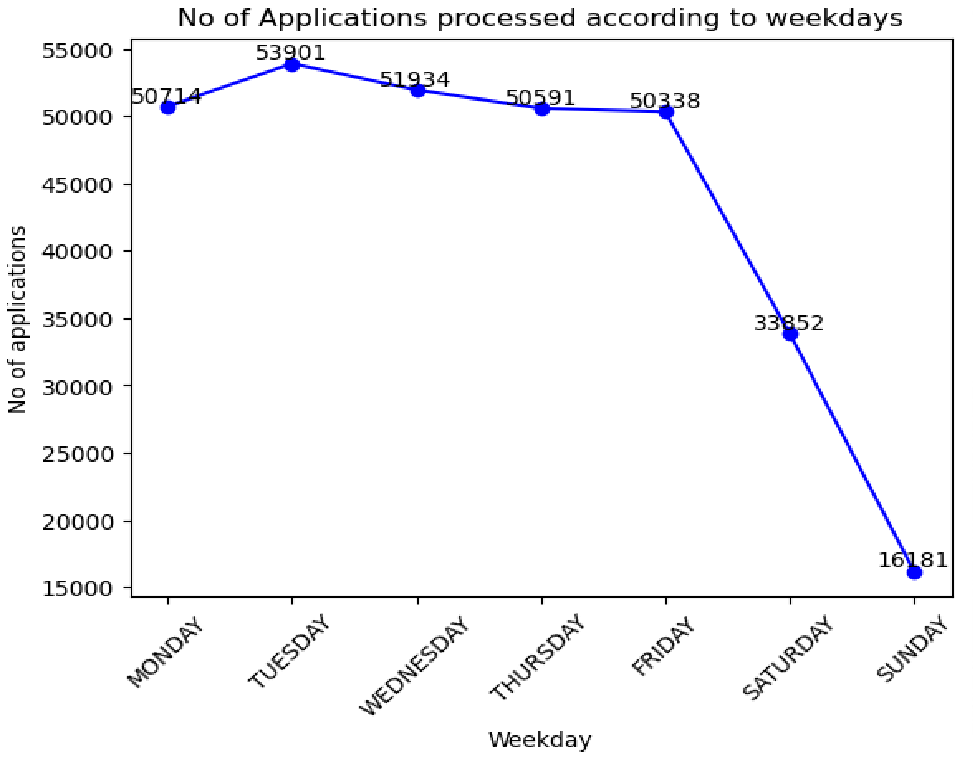


Insight :

- 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%).

Exploratory Data Analysis

➢ Analyzing the loan applications by the hour of the day

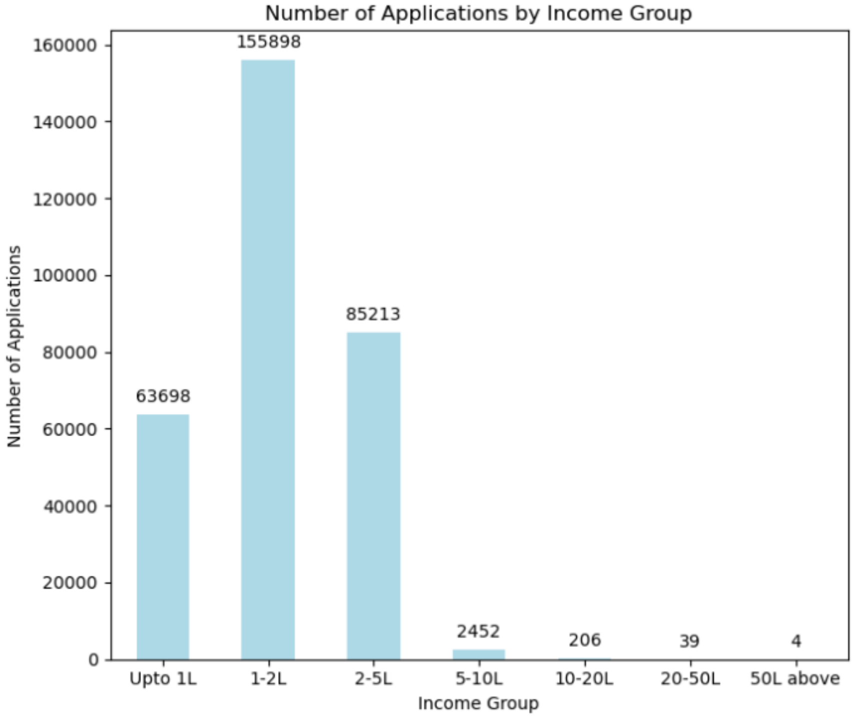


Insight :

- **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.

Exploratory Data Analysis

Analyzing the Income group

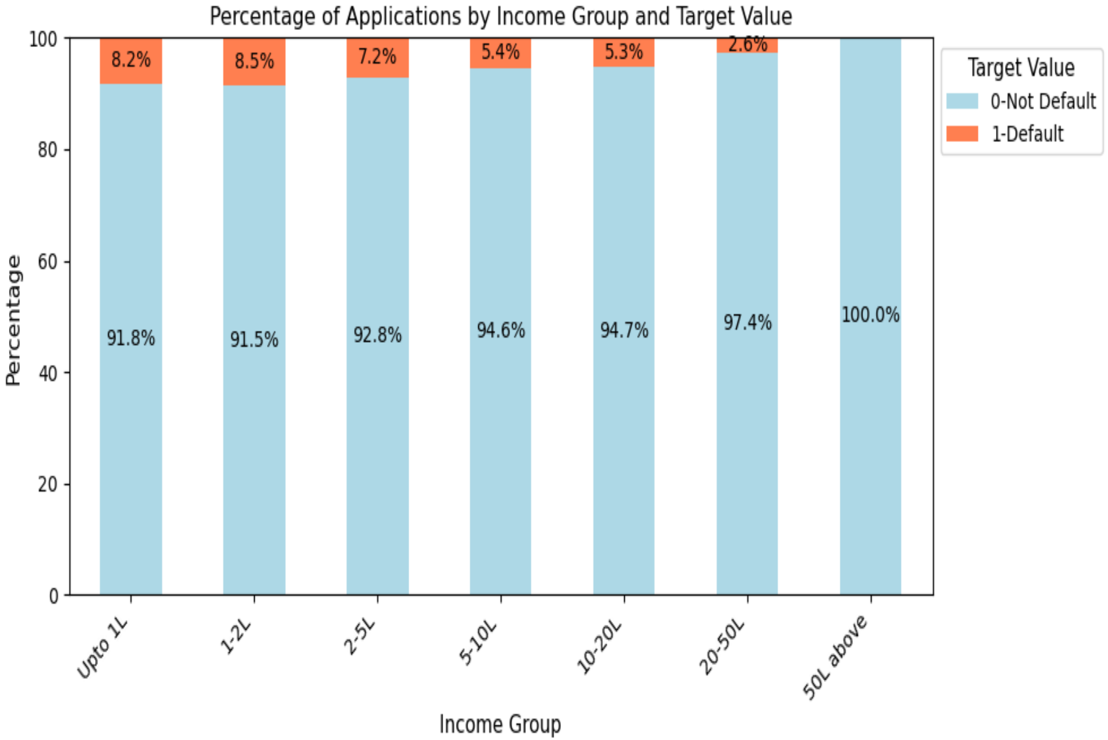


Insight :

- The application distribution reveals significant trends across various income brackets. The **"Upto 1 lakh"** group has 63,698 applications, making up about **17.3%** of the total, reflecting engagement from lower-income individuals. The **"1-2 lakh"** group leads with 155,898 applications, representing **42.5%**, indicating strong activity for major purchases or investments.
- However, engagement drops sharply in **h**igher income brackets, with only **0.7%** of applications (2,452) from the **"5-10 lakh"** group, and **even fewer** in the **"10-20 lakh"** (206 applications), **"20-50 lakh"** (39 applications), and **"50 lakh and above"** (4 applications) categories.
- This data suggests that most loan applicants are concentrated in the lower to middle-income ranges, with significantly fewer in the higher income brackets.

Exploratory Data Analysis

Analyzing the Income group

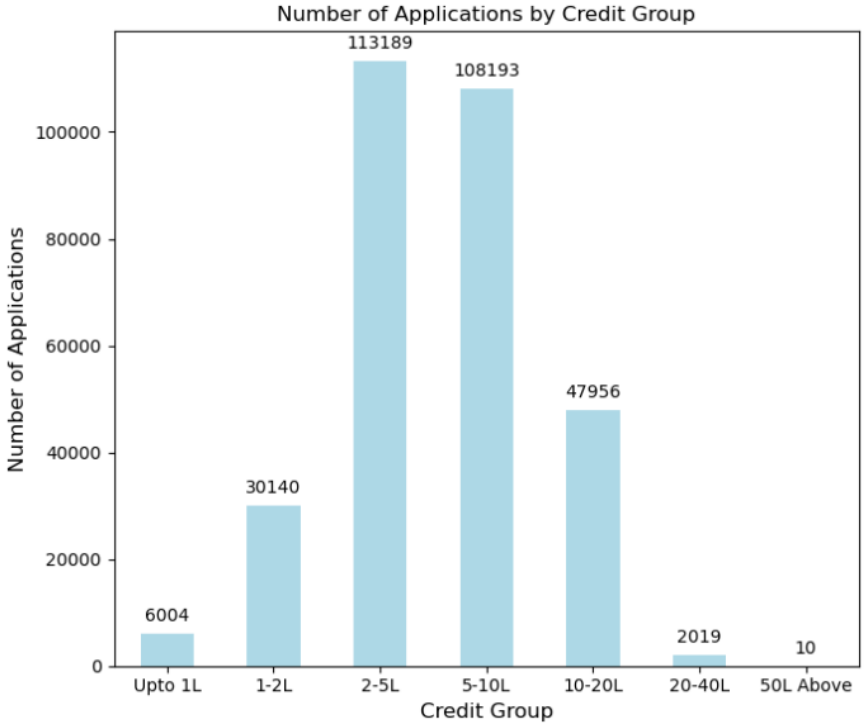


Insight:

- 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.

Exploratory Data Analysis

Analyzing the Credit group

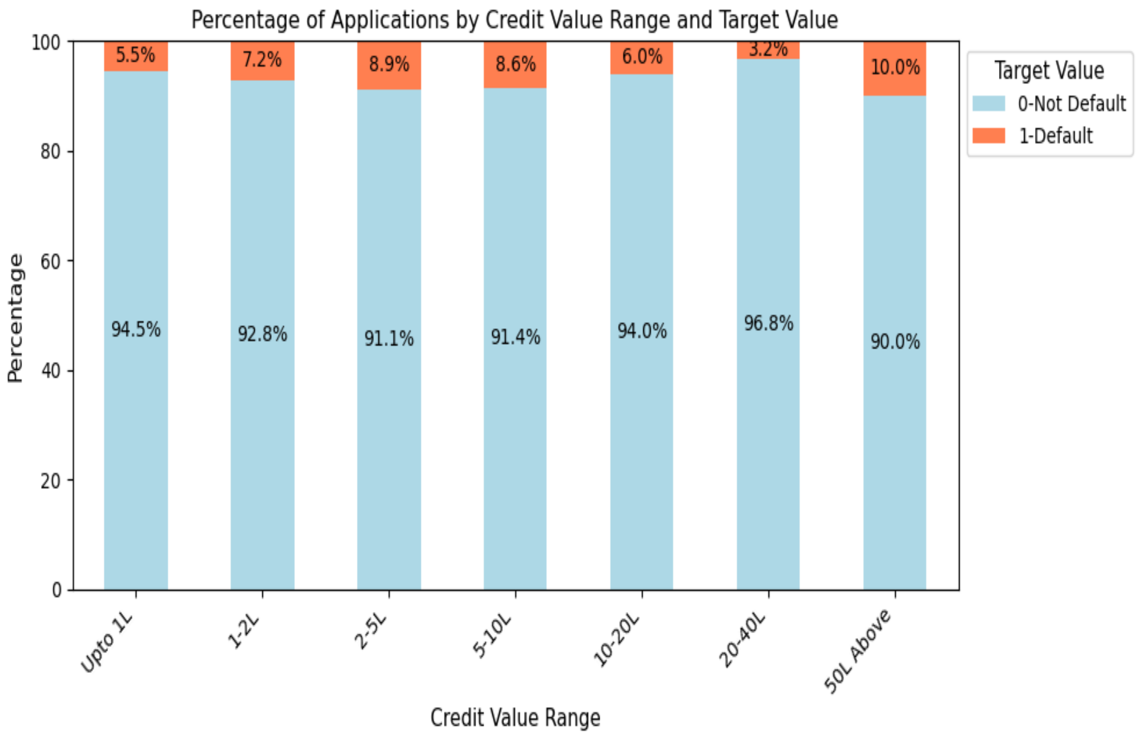


Insight :

- 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

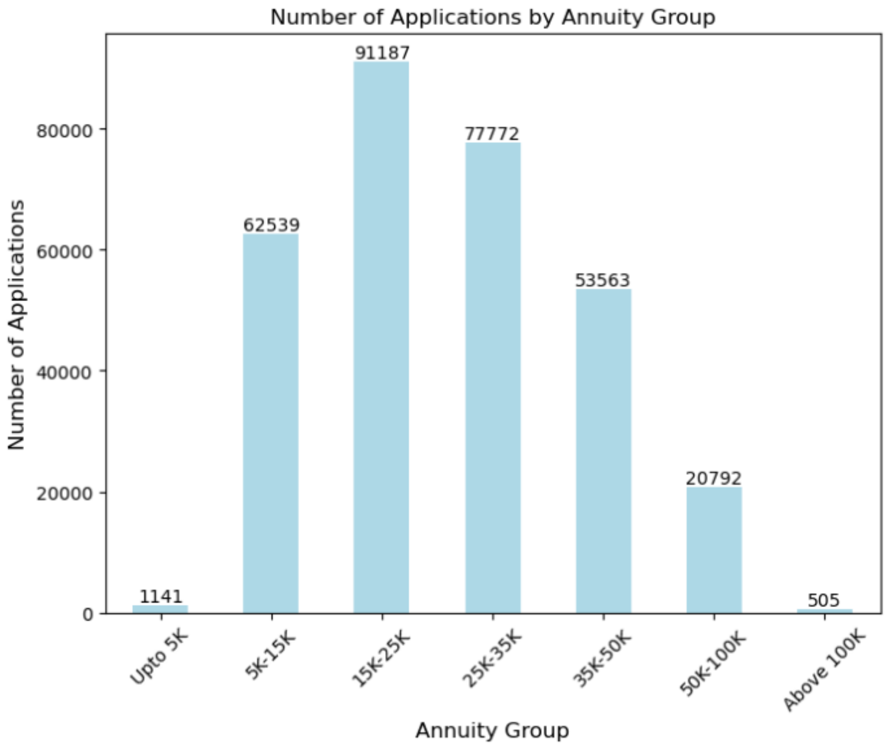


Insight :

- The **2-5 lakh** credit group has the highest loan applications, indicating strong demand for major purchases.
- 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.

Exploratory Data Analysis

Analyzing the Annuity group



Insight :

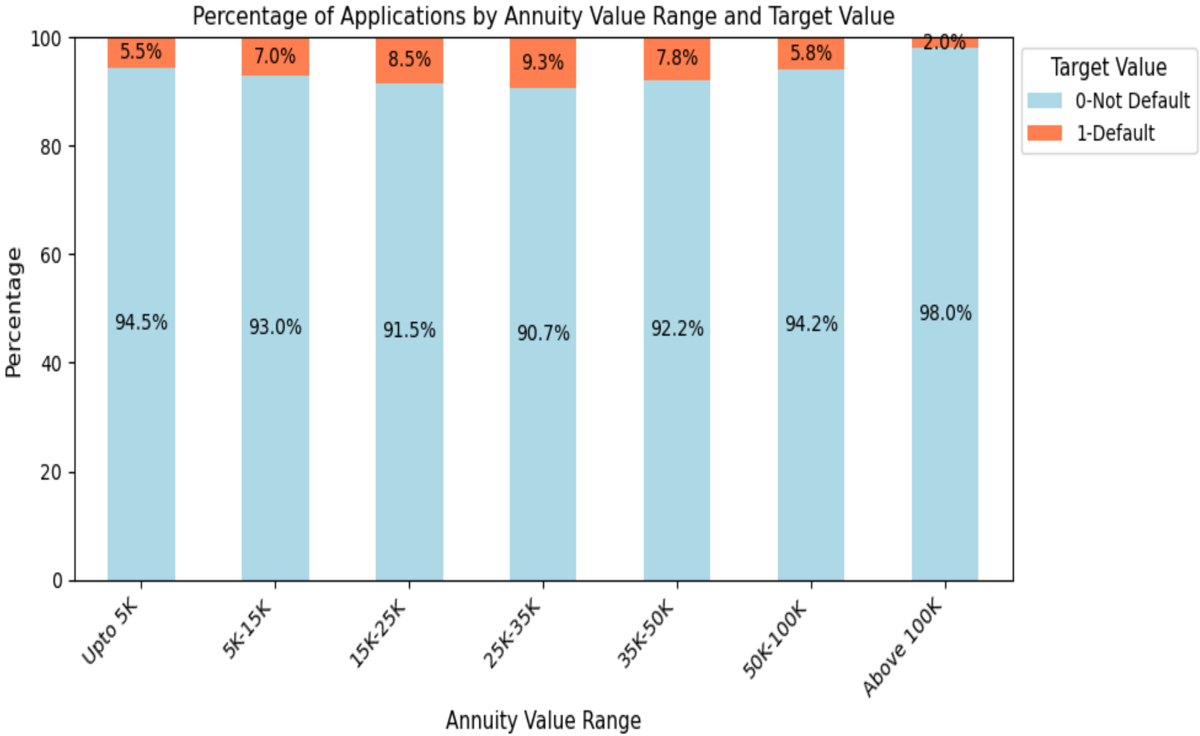
- The **15K-25K annuity group is the most popular**, with 91,187 applications (approximately **30.7%** of all applications), indicating strong demand for loans in this monthly payment range. **The 5K-15K group also attracts significant interest**, with 62,539 applications (about **21.3%**).
- Borrowers show a willingness to commit to higher monthly payments, as evidenced by the **25K-35K** group, which has 77,772 applications (approximately **26.5%**). However, interest drops sharply in the higher payment ranges, with the Above 100K category receiving only 505 applications (around 0.2%), suggesting that the prospect of high monthly payments deters many potential borrowers.

SCALETRIX.AI

Image source: icons8

Exploratory Data Analysis

Analyzing the Annuity group

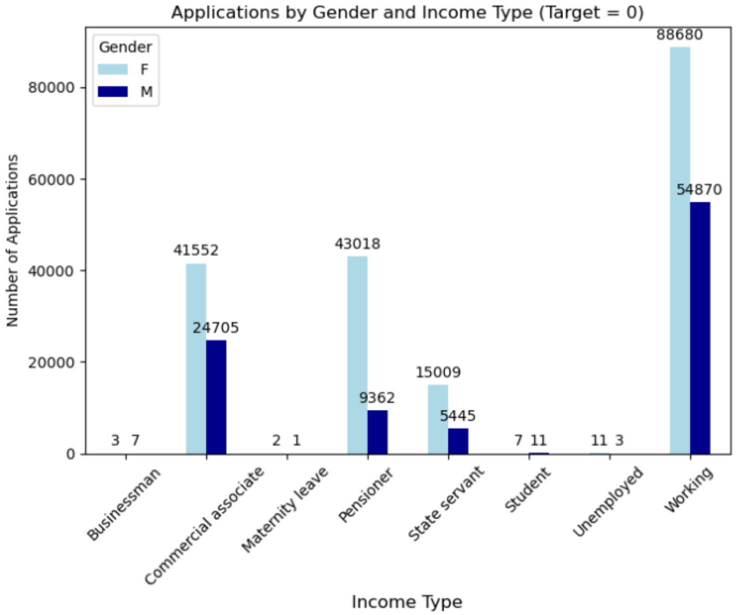


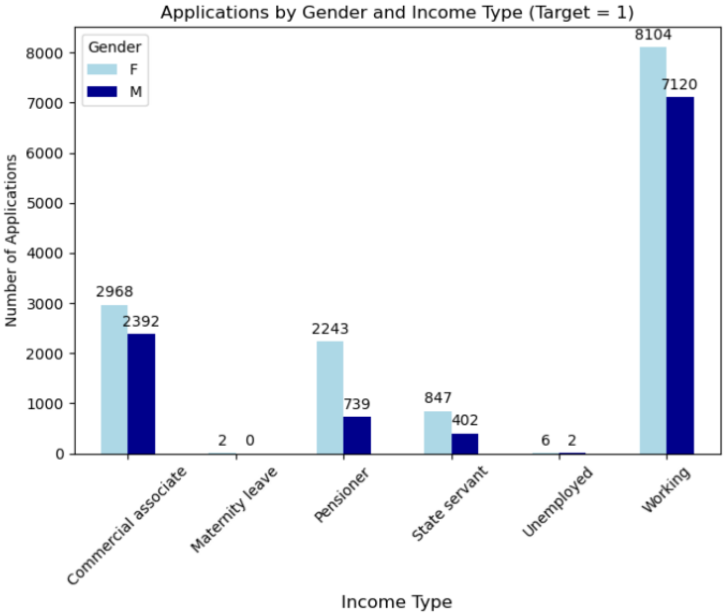
Insight :

- The **15K-25K** annuity group leads with **91,187** applications (30.7%), followed by **5K-15K** at **62,539 (21.3%)**.
- 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% (Upto 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.

Exploratory Data Analysis

➢ **Analyzing the Income type by gender**



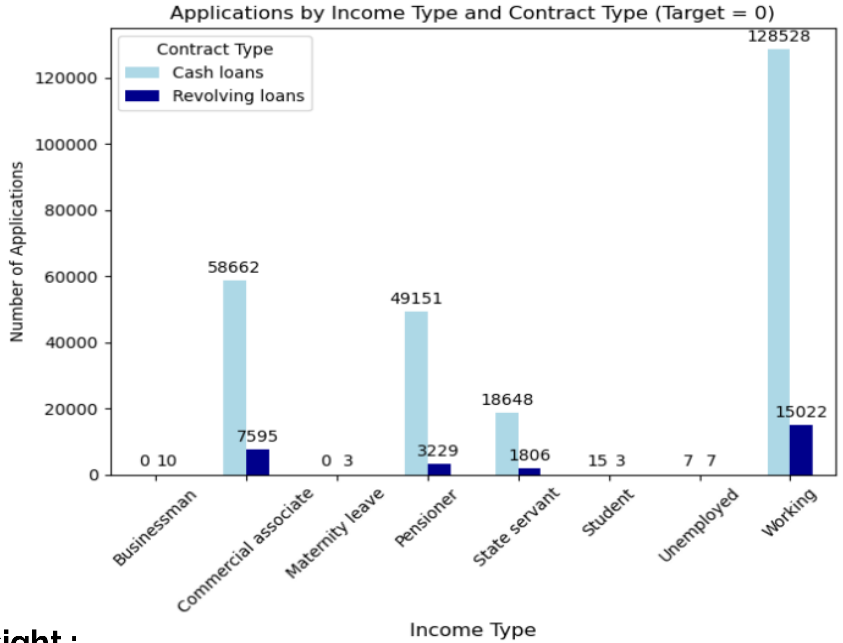


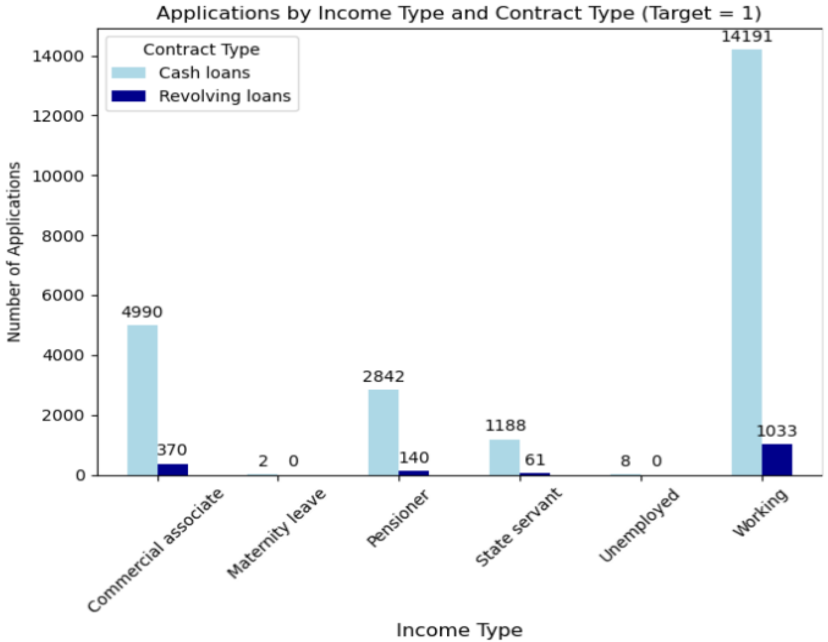
Insight :

- **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.

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.

SCALETRIX.AI

Image source: icons8

Exploratory Data Analysis

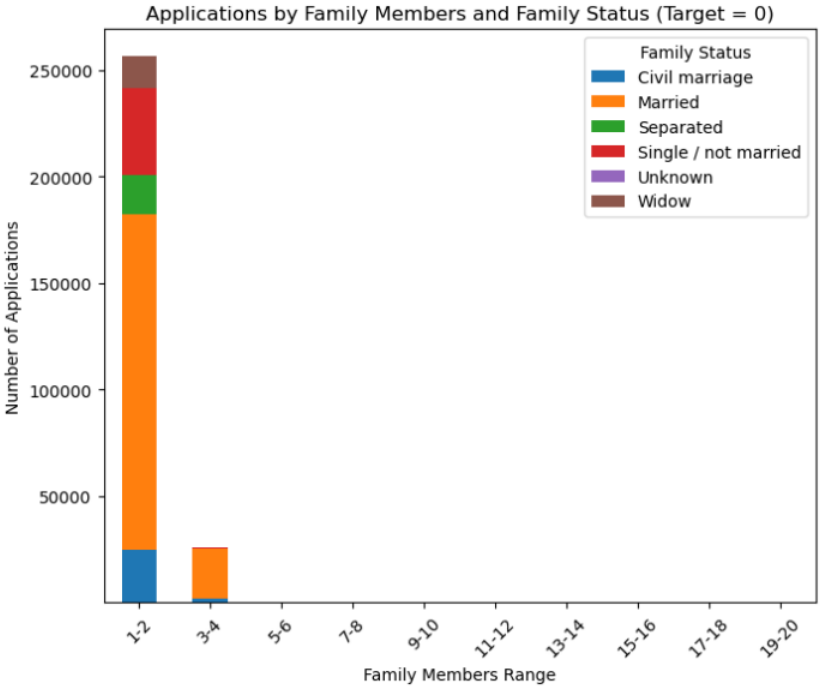
Average Income type by gender according to Target

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.

Exploratory Data Analysis

Analyzing family member and family status

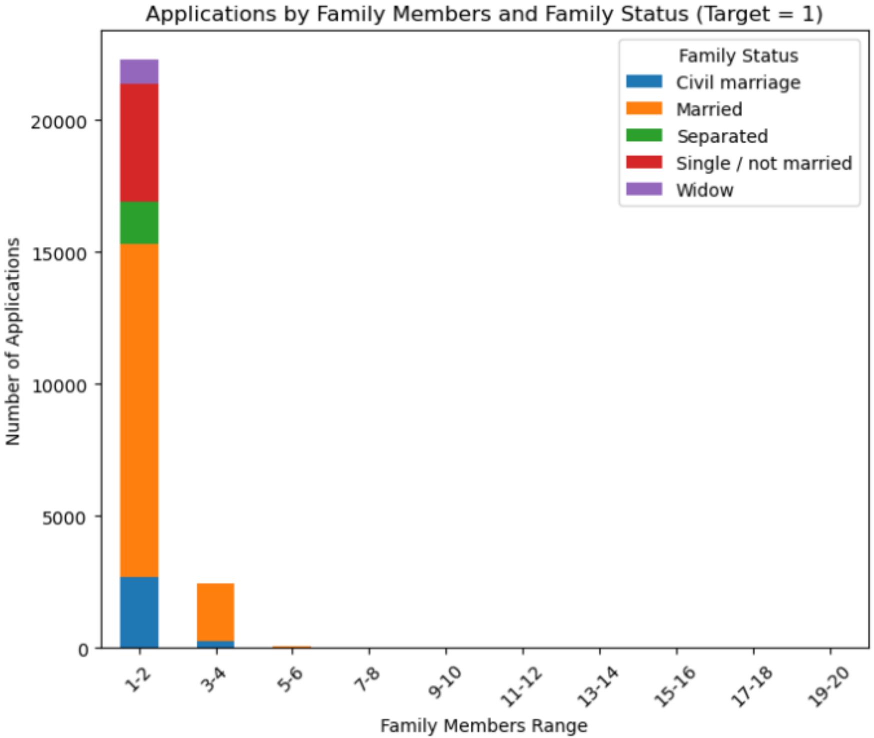


Insight :

- The **majority of non-defaulters** come from families with **1-2 members**, with 157,646 applications from married individuals and 40,904 applications from single or unmarried applicants. This suggests that smaller family units may find it easier to manage finances, leading to higher rates of loan repayment.
- In contrast, applications from families with 3-4 members remain significant, with 23,517 applications from married couples in this category. However, there is a sharp decline in applications for families with 5 or more members, indicating that larger families may encounter more financial challenges.

Exploratory Data Analysis

➢ Analyzing family member and family status

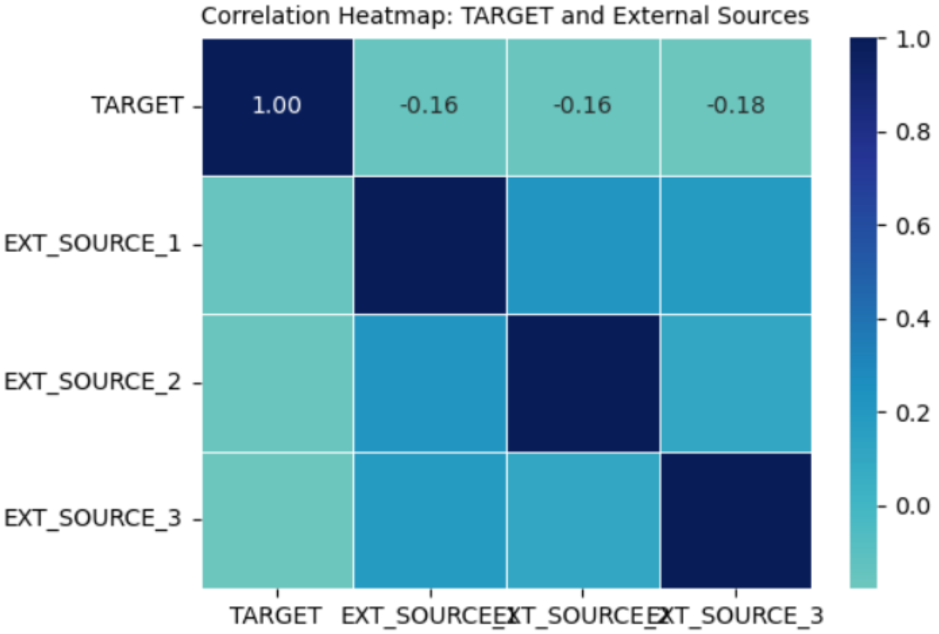


Insight :

- 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.

Exploratory Data Analysis

➢ Analyzing the relationship of Target and External Sources

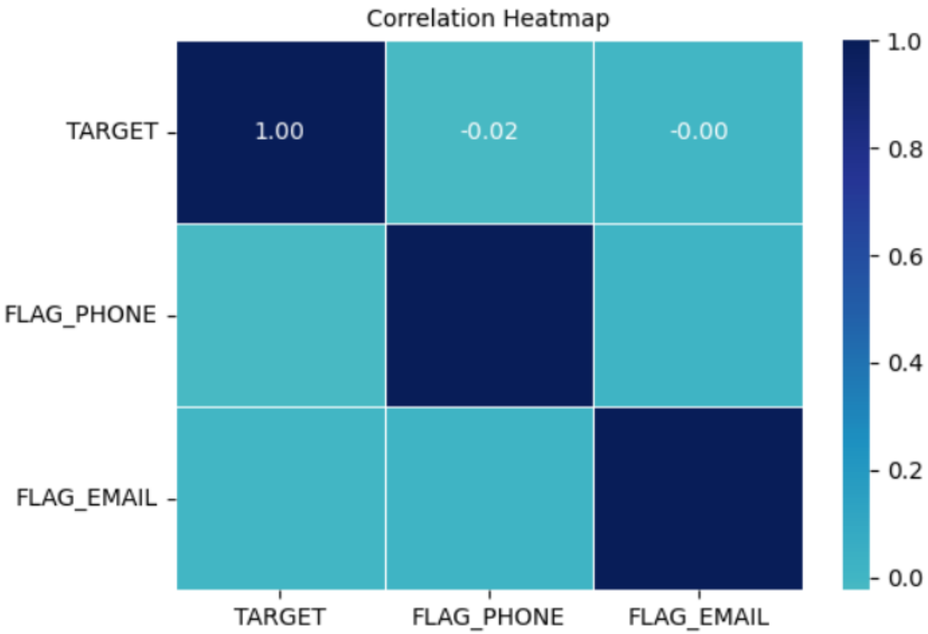


Insight :

- Negative Correlation with **TARGET: All external sources (EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3) show a negative correlation** with the target variable, suggesting that higher values in these external sources may be associated with a lower likelihood of the target event occurring (e.g., defaults).

Exploratory Data Analysis

➢ Correlation of Target, Flag_Phone and Flag_Email

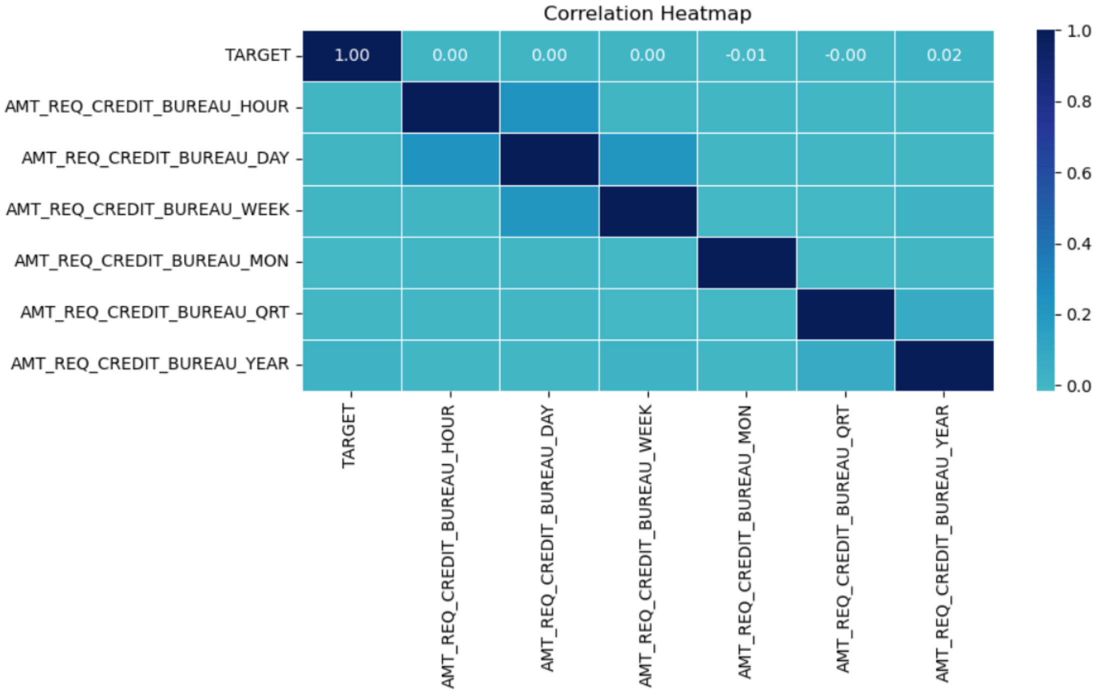


Insight :

- **Both FLAG_PHONE and FLAG_EMAIL exhibit very weak negative correlations with the target variable, at -0.0238 and -0.0018,** respectively. This suggests that having a phone or email contact does not significantly influence the likelihood of the target event occurring.

Exploratory Data Analysis

➤ Correlation of Target and Enquiries to Credit Bureau

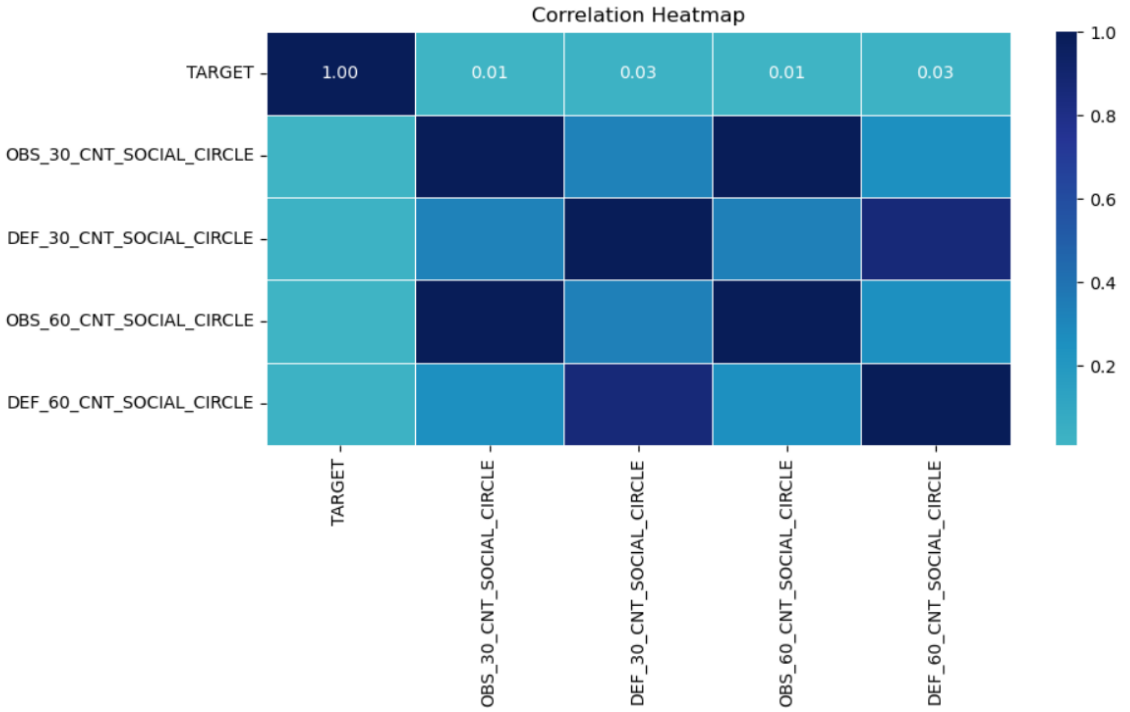


Insight :

- The **correlation values between** the **TARGET variable** (indicating default status) and the **Credit Bureau request** metrics are all **very low**, suggesting that the frequency of credit bureau inquiries does not strongly influence whether an applicant is a defaulter or non-defaulter.

Exploratory Data Analysis

➢ Correlation of Target and Client's Social Surroundings



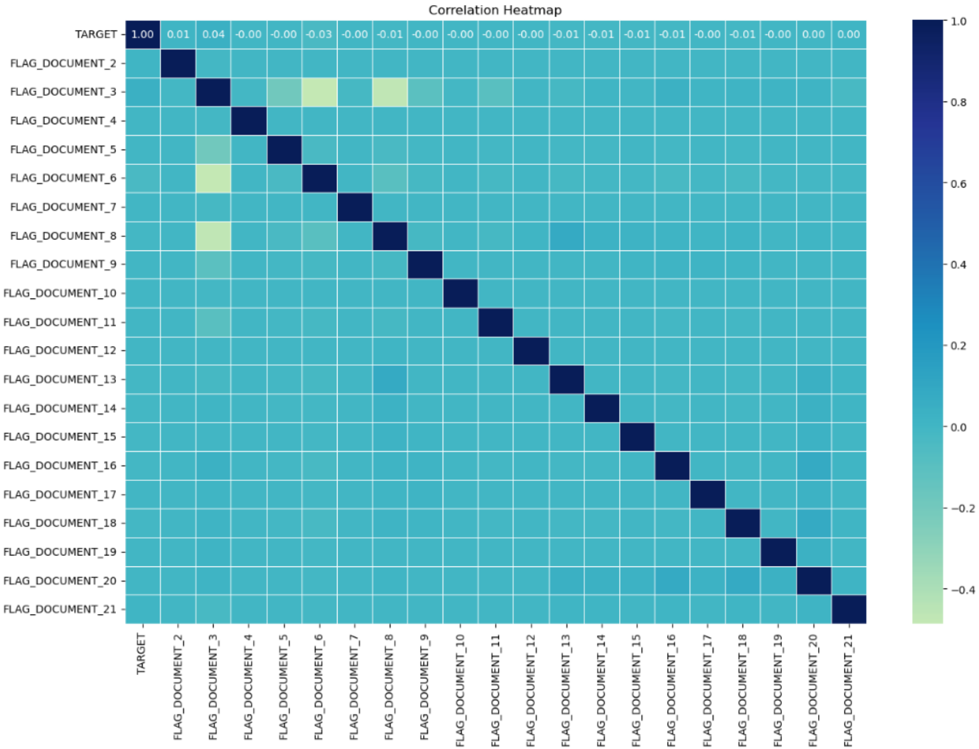
Insight :

- The correlation of **TARGET with OBS_30_CNT_SOCIAL_CIRCLE (0.009131) and OBS_60_CNT_SOCIAL_CIRCLE (0.009022) is very low**, indicating that the number of observed social connections does not significantly influence the likelihood of default.
- The correlation with **DEF_30_CNT_SOCIAL_CIRCLE (0.032248) and DEF_60_CNT_SOCIAL_CIRCLE (0.031276) suggests a minor relationship**, indicating that a higher number of defaulters within social circles may slightly correlate with an increased likelihood of default. However, this relationship is still weak.

Image source: icons8

Exploratory Data Analysis

Correlation of Target and the documents submitted

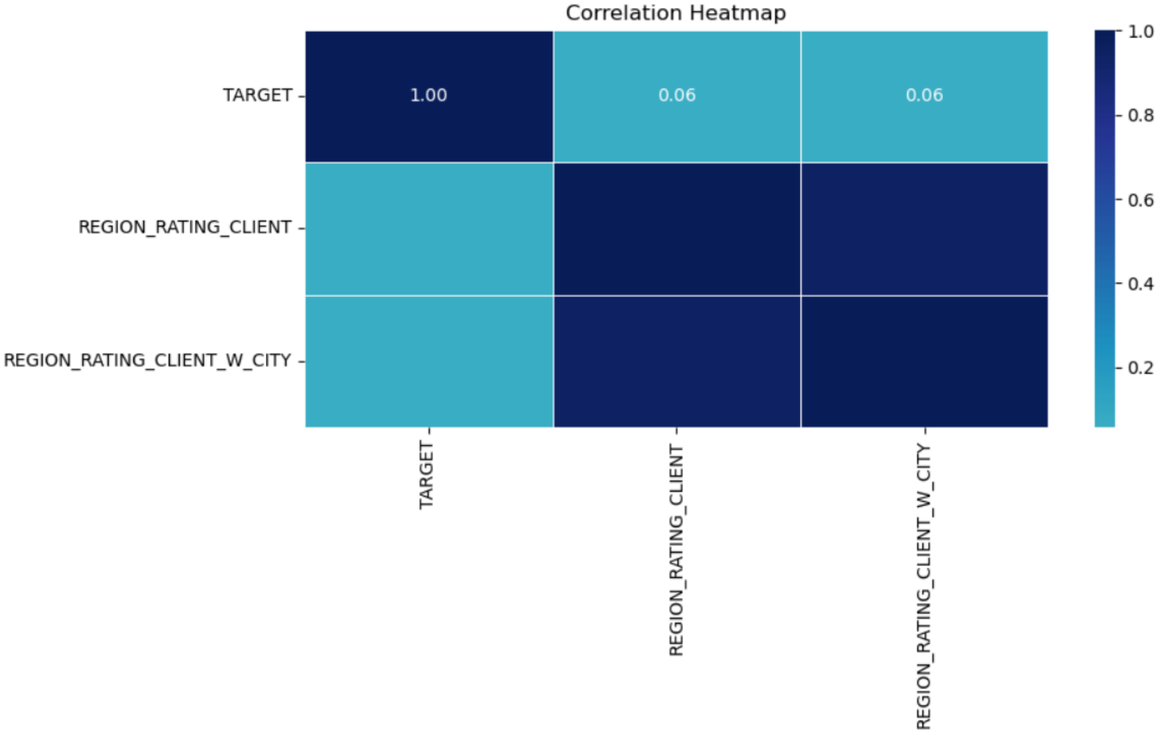


Insight :

- Several flags, such as **FLAG_DOCUMENT_3** and **FLAG_DOCUMENT_6**, exhibit **slight negative correlations**, suggesting that higher values may relate to a lower likelihood of default.
- Many flags, including **FLAG_DOCUMENT_4** and **FLAG_DOCUMENT_5**, have **near-zero correlation values**, indicating they likely do not significantly influence defaulting behavior.
- While **FLAG_DOCUMENT_18** shows a **slightly positive correlation** of 0.081589 with **TARGET**, and **FLAG_DOCUMENT_20 is close to zero**, both suggest minimal predictive relevance.

Exploratory Data Analysis

Region Rating and City Rating with respect to Target

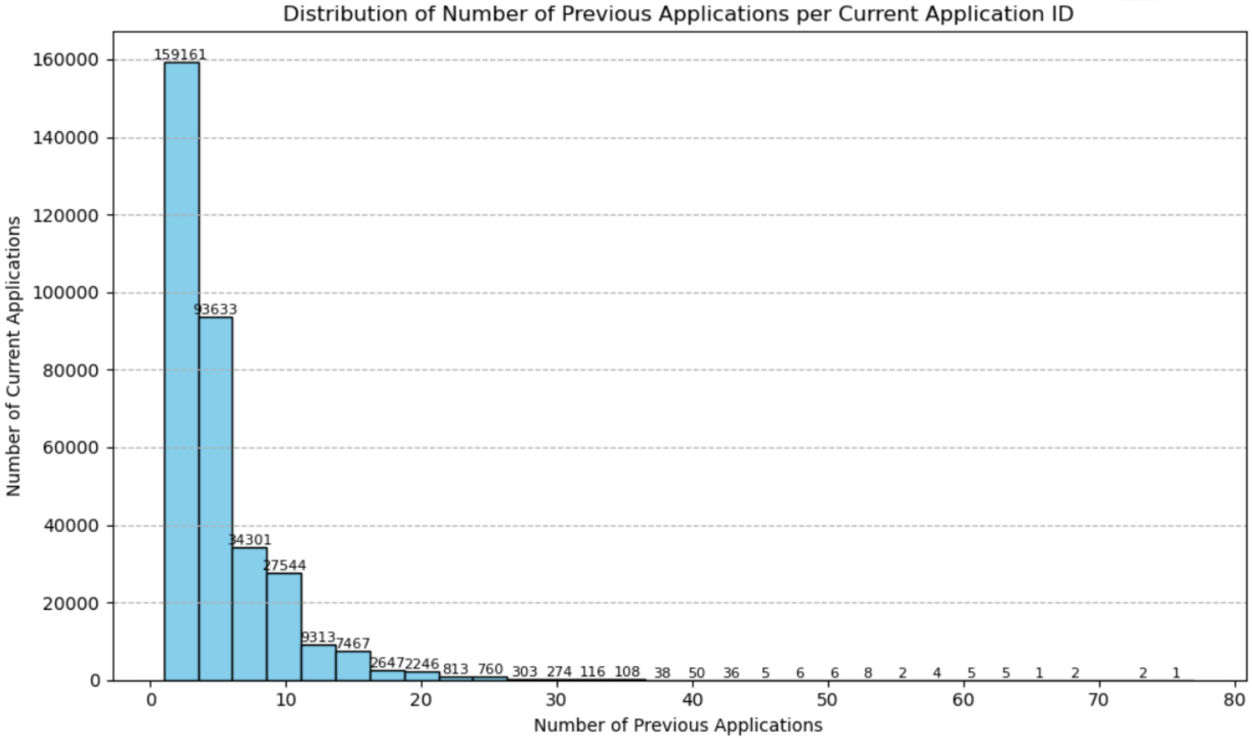


Insight :

- The weak correlations suggest that while regional and city ratings have some connection to default behavior, they are not strong predictors.

Exploratory Data Analysis

➢ Relationship between current application and previous application IDs

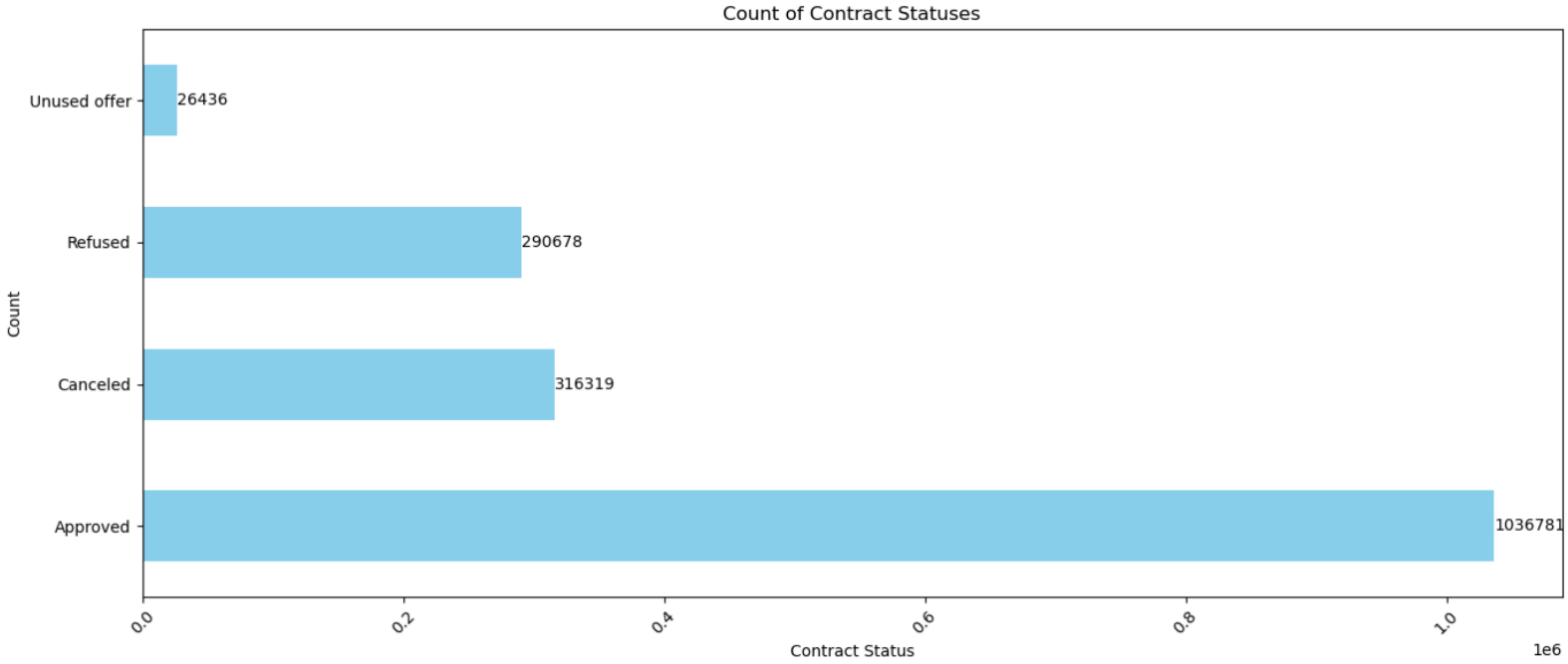


Insight :

- **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

Exploratory Data Analysis

➢ Analyzing the Contract Status



Exploratory Data Analysis

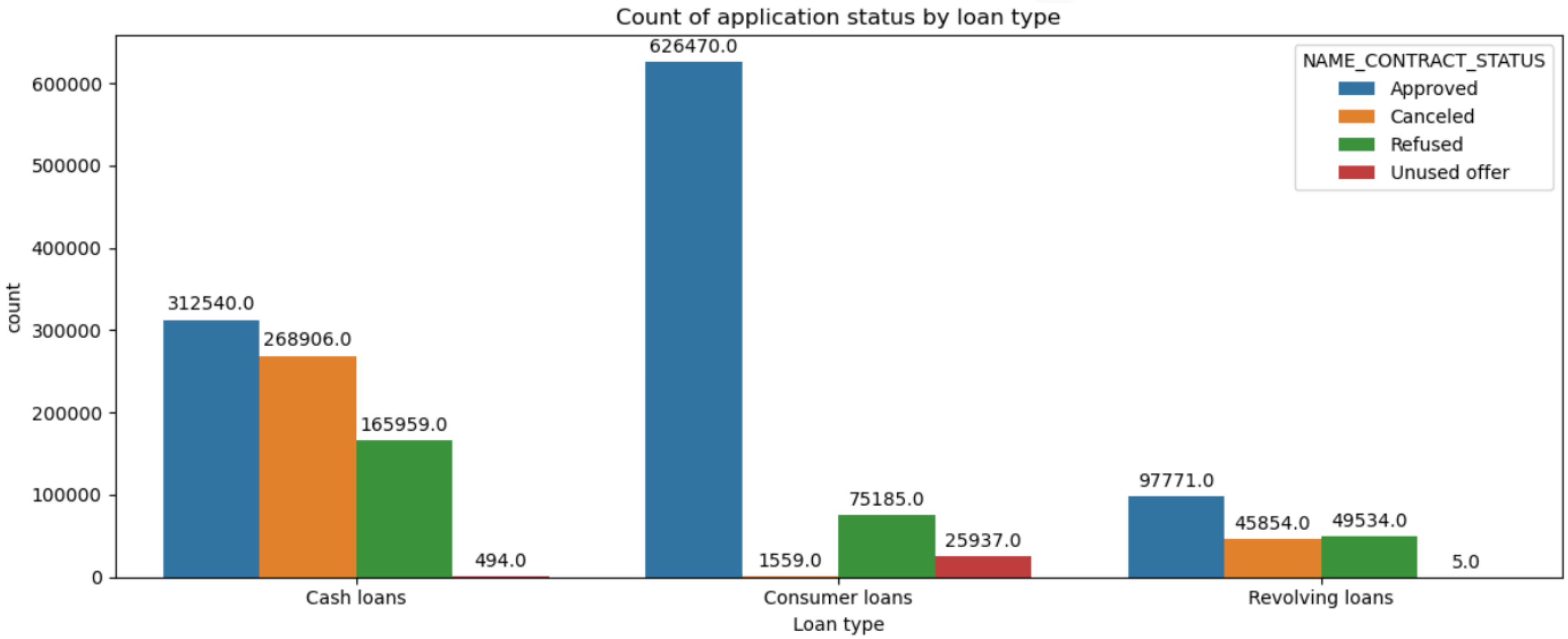
Analyzing the Contract Status

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.

Exploratory Data Analysis

- Analyzing the number of applications by Loan Type



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

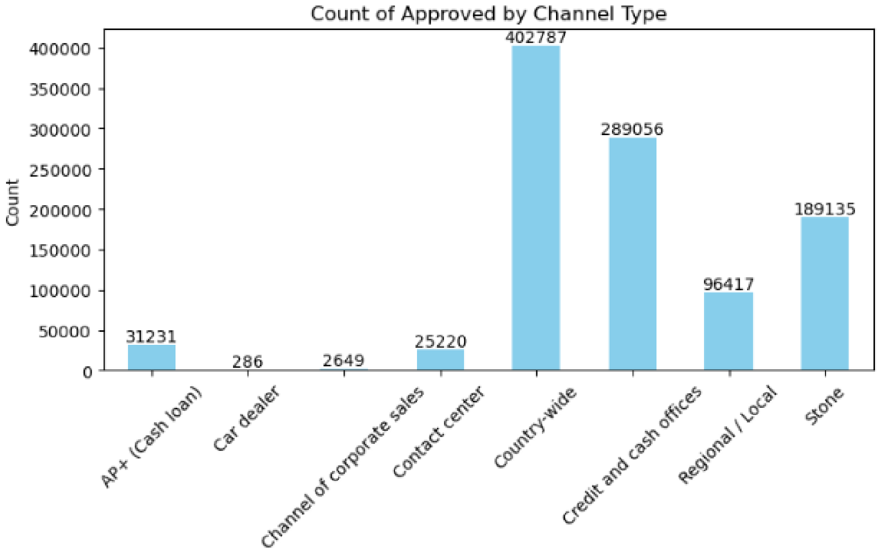
NAME_CONTRACT_STATUS /NAME_CASH_LOAN_PURPOSE	Refusal to name the goal	Repairs	Urgent needs	Wedding / gift / holiday	XAP	XNA
Approved	4	8677	3574	397	724241	285607
Canceled	0	621	148	23	47728	266952
Refused	11	14421	4690	542	124750	125070
Unused offer	0	46	0	0	25942	289

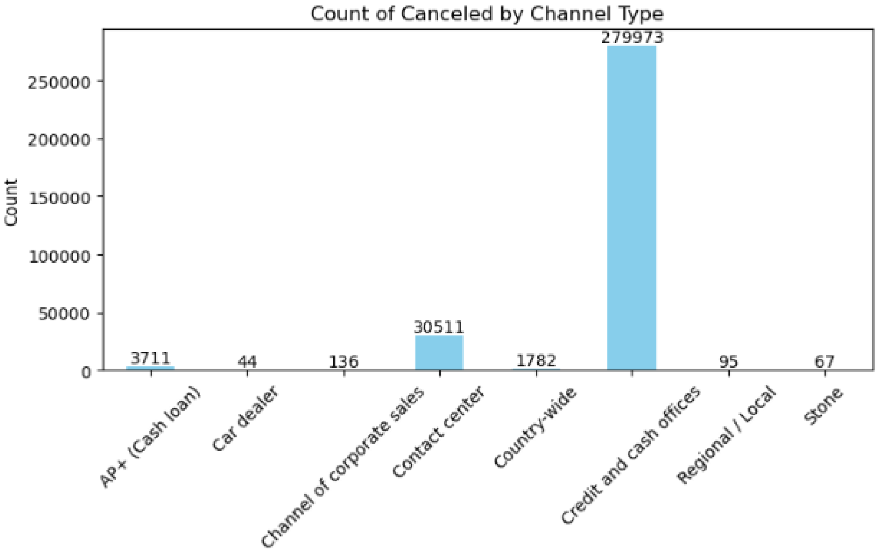
Insight :

- 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



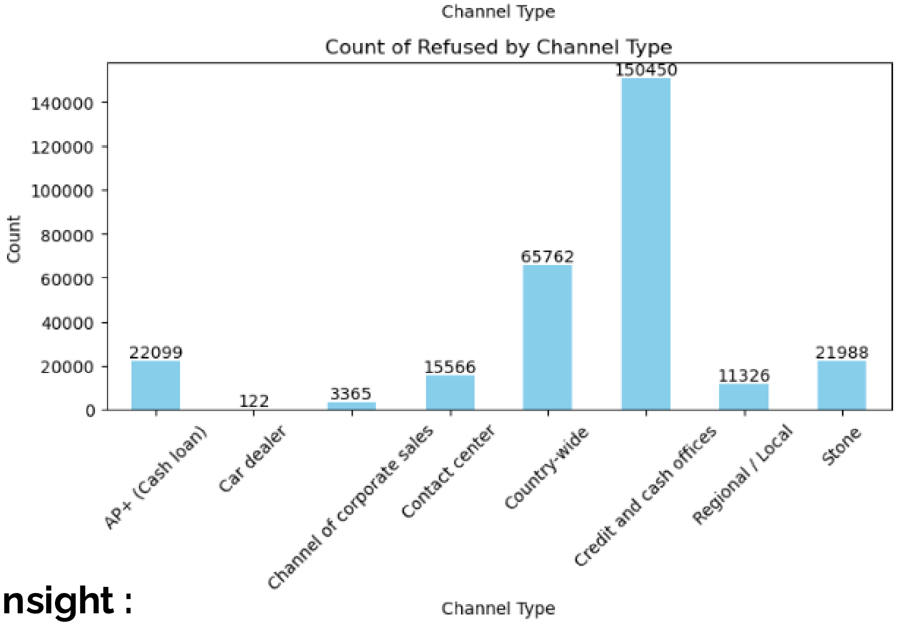


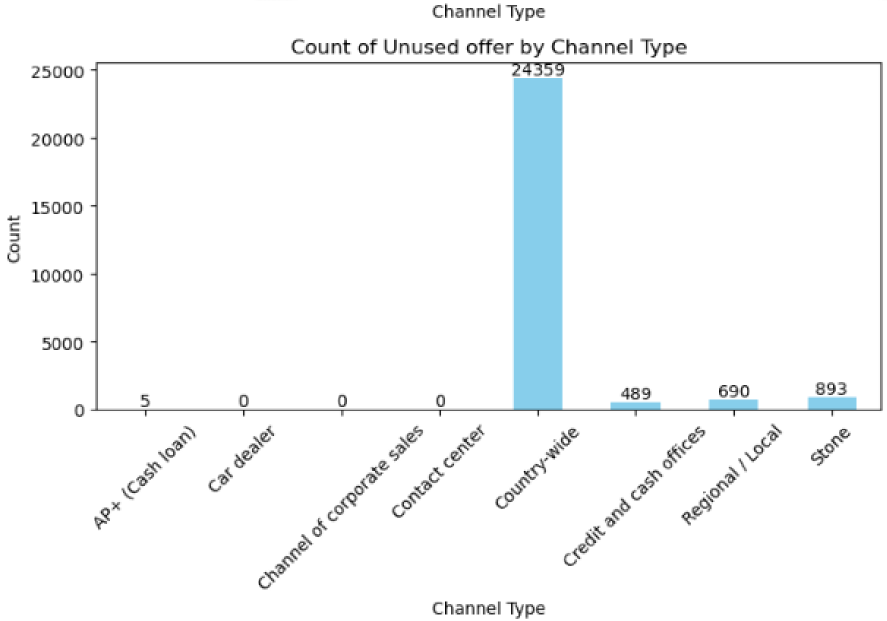
Insight :

- 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)

Exploratory Data Analysis

➢ Analyzing Contract Status by Channel Type



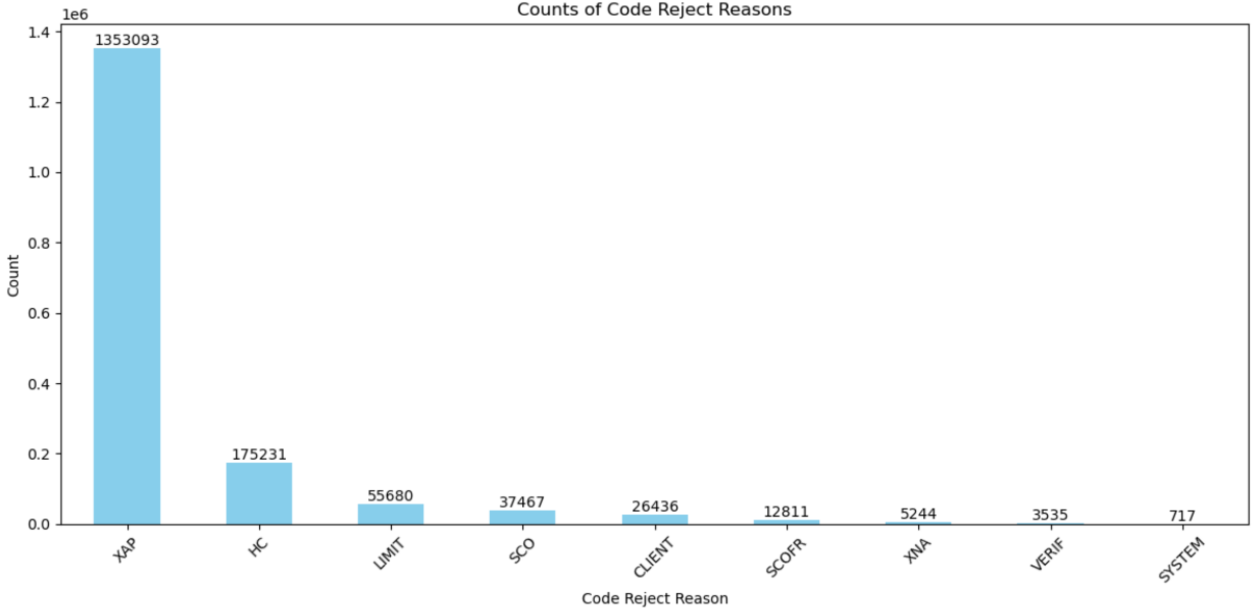


Insight :

- 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.

Exploratory Data Analysis

Analyzing Contract Status by Channel Type

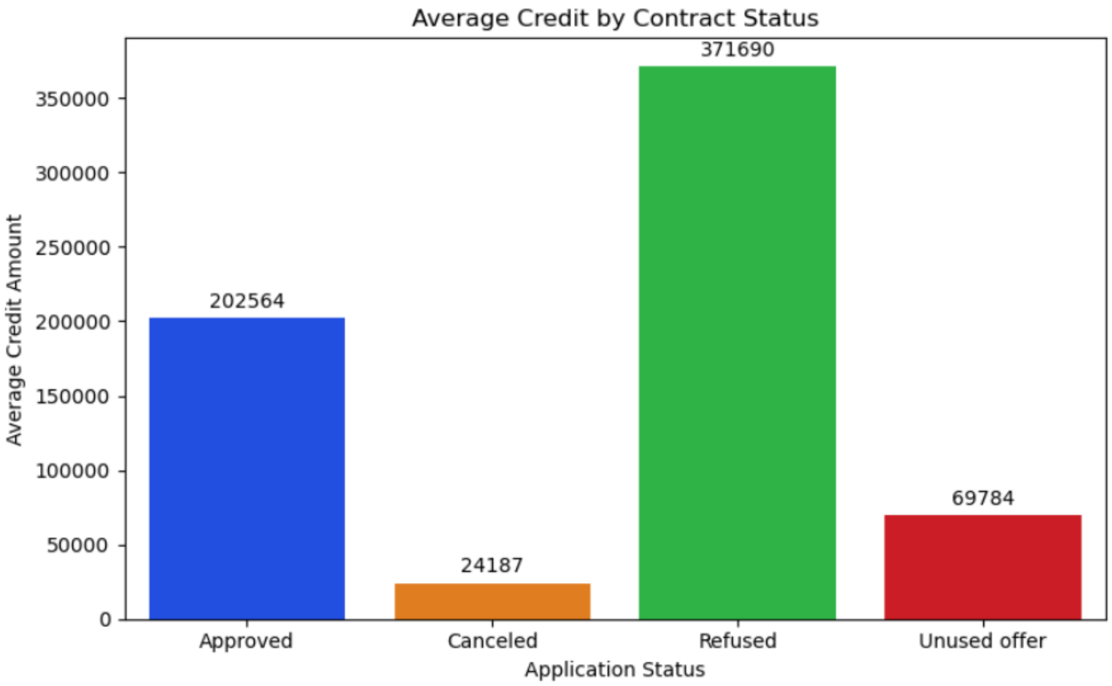


Insight :

- 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

Exploratory Data Analysis

Analyzing Credit by Contract Status

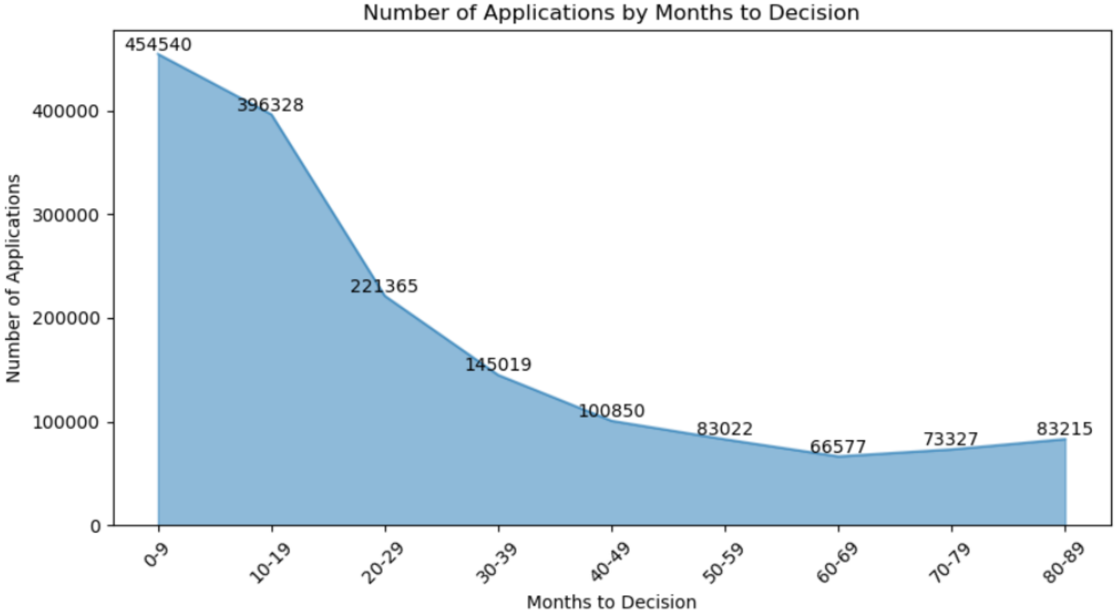


Insight :

- **Approved contracts** have an **average credit amount** of 202,564, indicating robust lending practices for accepted applications.
- In contrast, **canceled contracts** show a significantly **lower average** of 24,187, which may suggest that these amounts are less committed or that smaller loans are being abandoned.
- **Refused contracts** have the **highest average credit amount** at 371,690, which means that larger loan requests are often denied. This could suggest that the lending criteria are strict or that there is a mismatch between what borrowers are asking for and what they qualify for.
- The average for **unused offers** at 69,784 suggests a **moderate value**, indicating opportunities for conversion if follow-up strategies are improved.

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

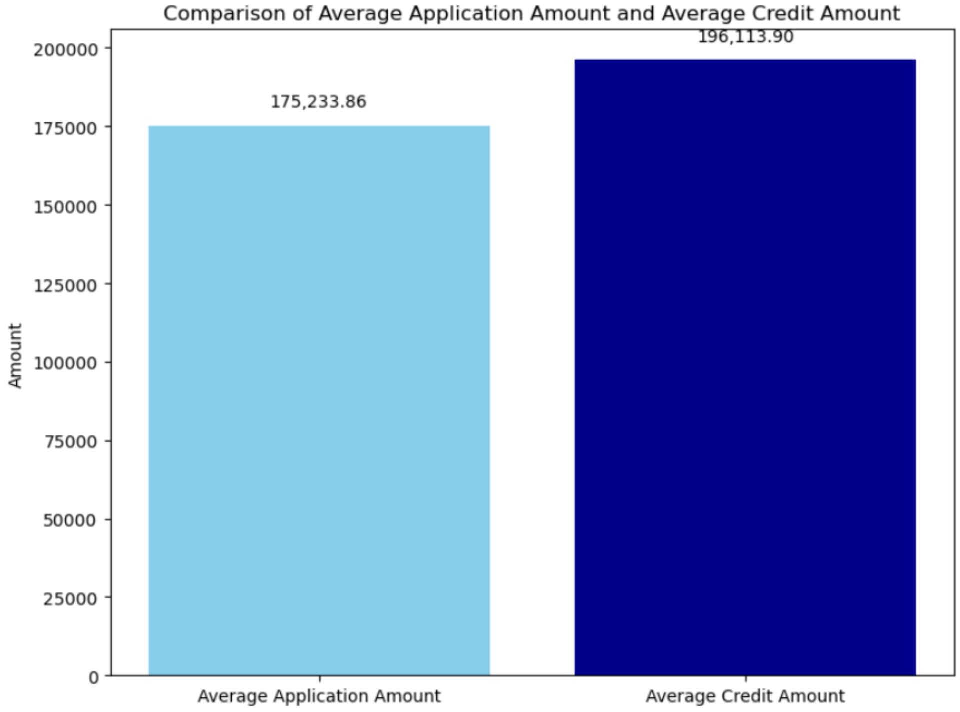
NAME_GOODS_CATEGORY/ NAME_CONTRACT_STATUS	Approved	Canceled	Refused	Unused offer
Additional Service	116	0	12	0
Animals	1	0	0	0
Audio/Video	89394	32	9080	935
Auto Accessories	6560	2	679	140
Clothing and Accessories	21460	0	2010	84
Computers	88050	35	13534	4150
Construction Materials	22471	7	2454	63
Consumer 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

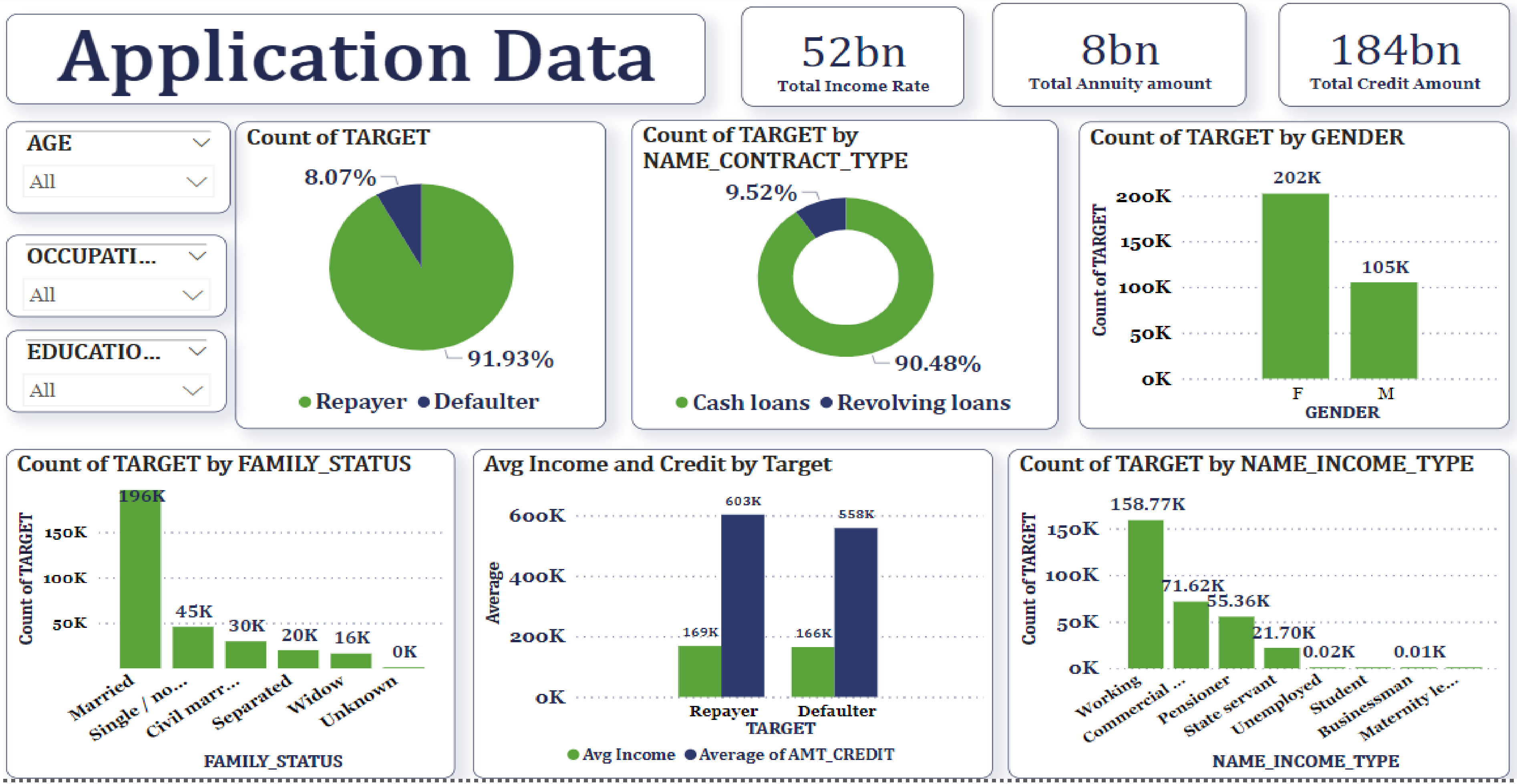


Insight :

- 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.**

SCALETRIX.AI

Image source: icons8



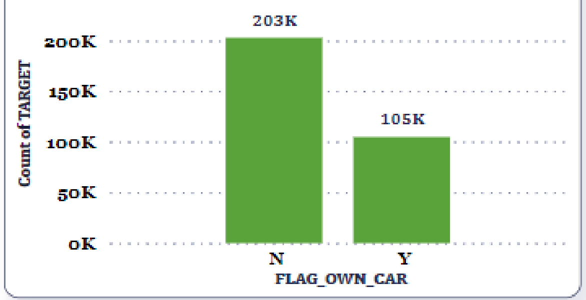
SCALETRIX.AI

Application Data

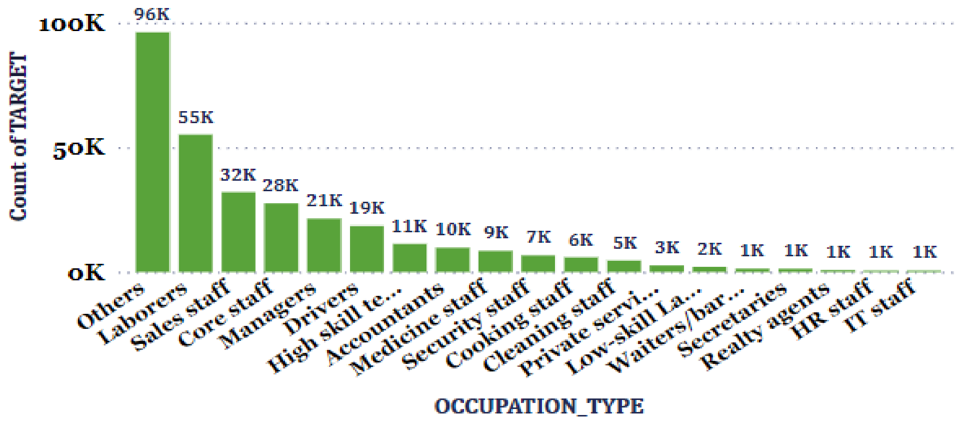
TARGET
All

AGE
All

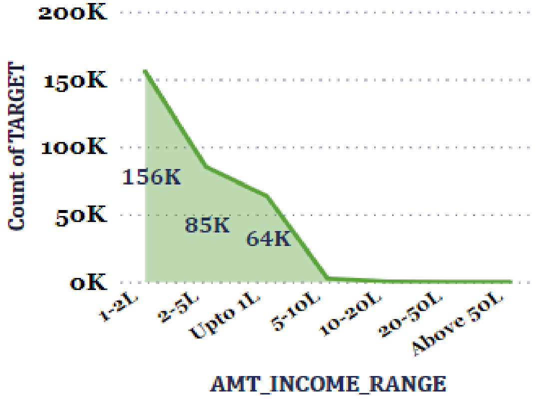
Count of TARGET by FLAG_OWN_CAR



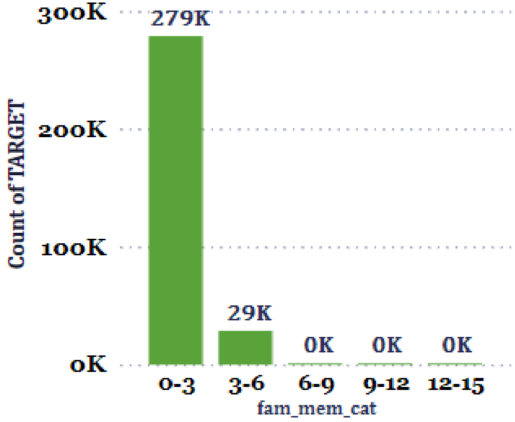
Count of TARGET by OCCUPATION_TYPE



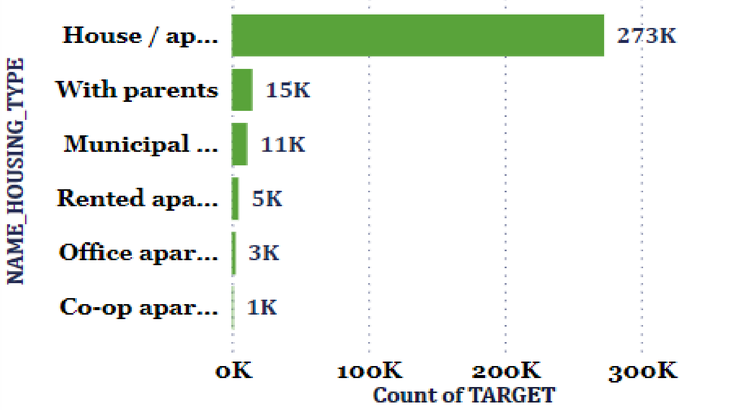
Count of TARGET by AMT_INCOME_RANGE



Target as per Family Members



Count of TARGET by NAME_HOUSING_TYPE



Previous Data

1,049K
Application Count

183bn
Total application price

204bn
Total Credit amount

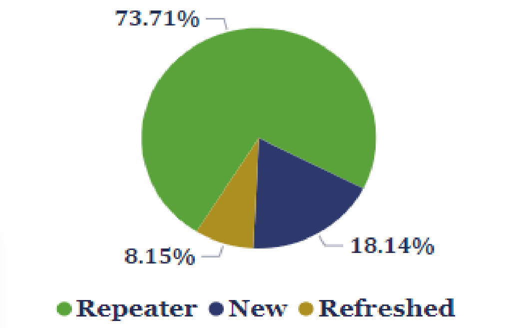
882
Avg DAYS_DECISION

LOAN_STATUS
All

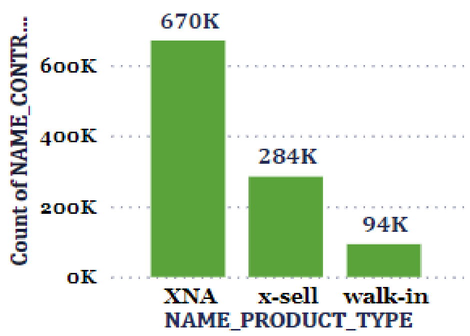
LOAN_PURPOSE
All

CHANNEL_TYPE
All

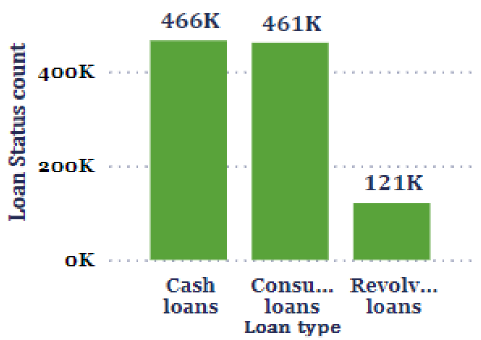
Client Type Percentage(%)



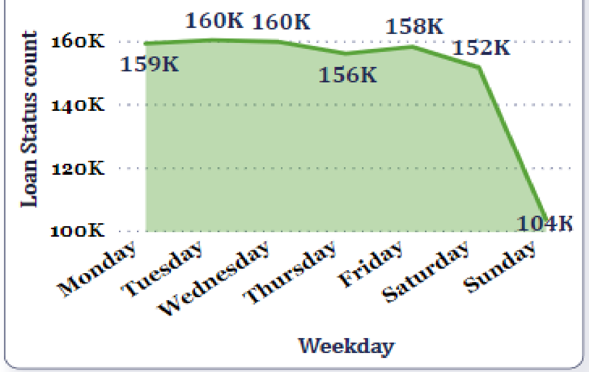
Contract Status by Product Type



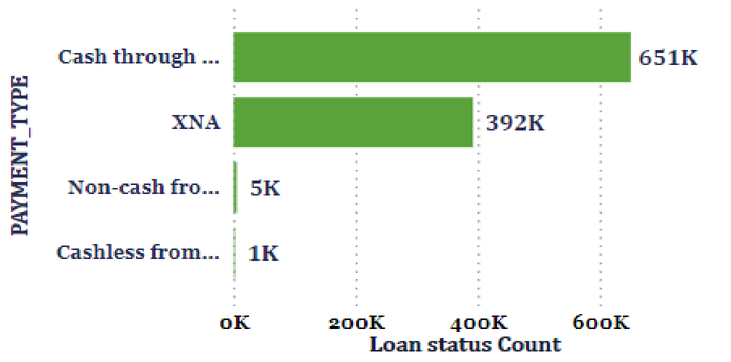
Loan Status count by Loan type



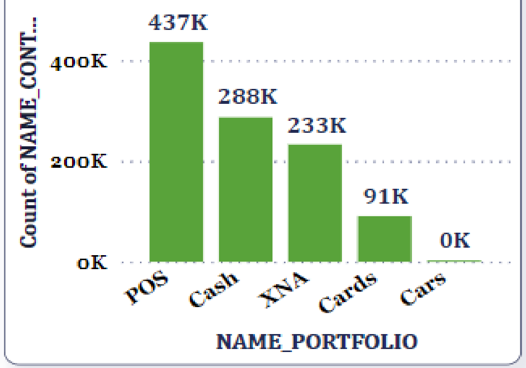
Weekday Loan Approval Status



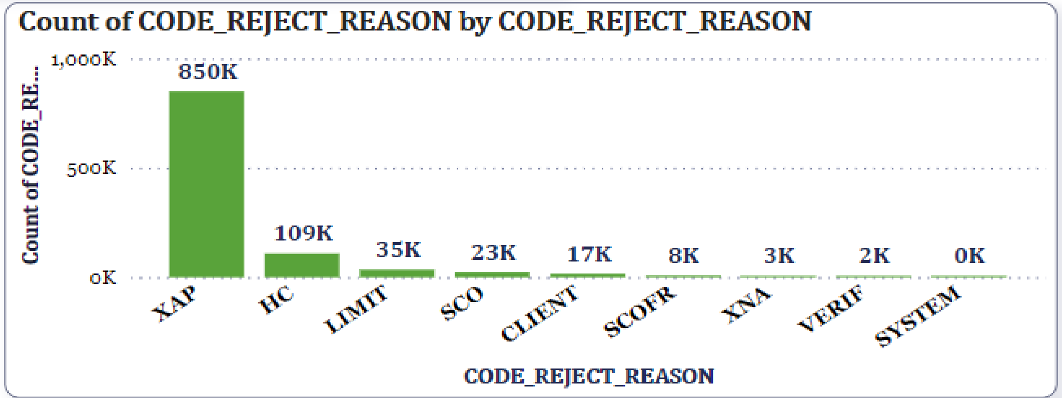
Loan status by Payment type

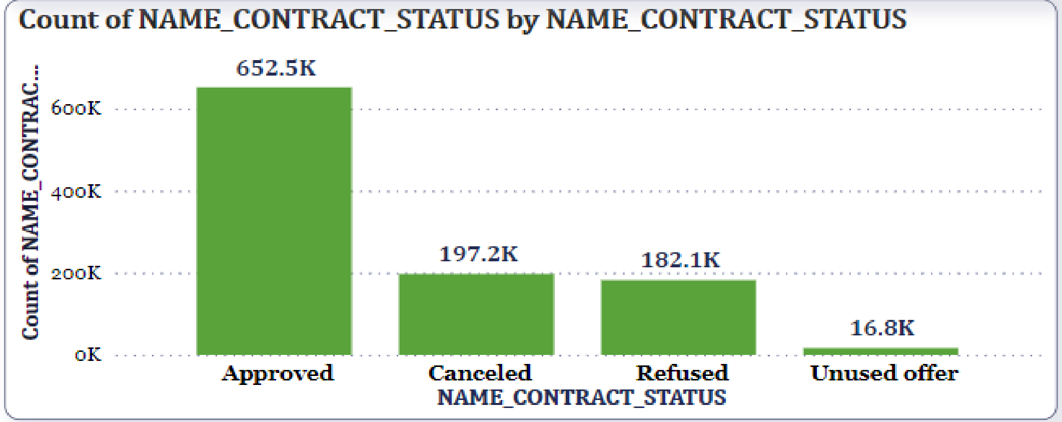


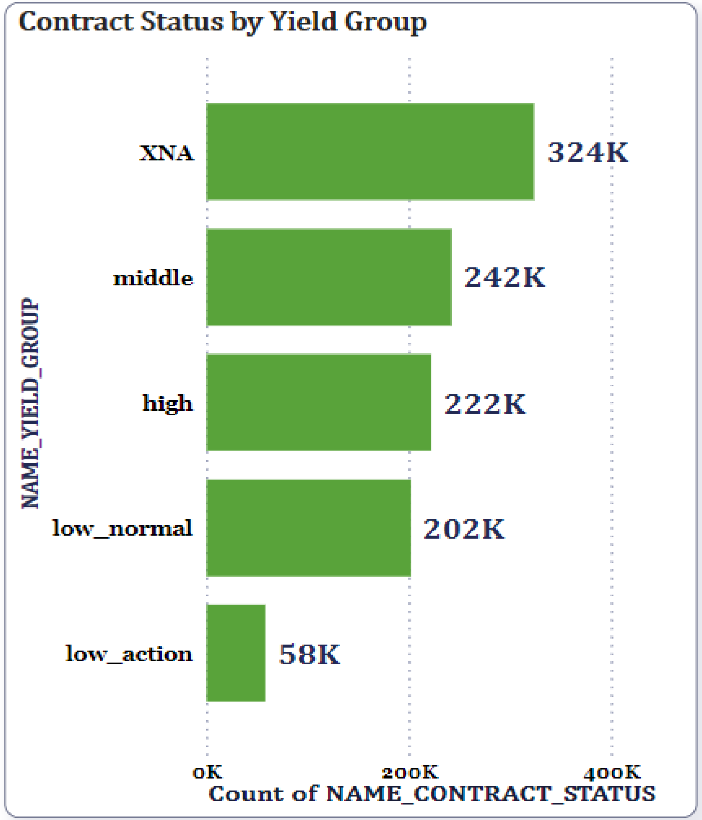
Contract Status by Portfolio



Previous Data

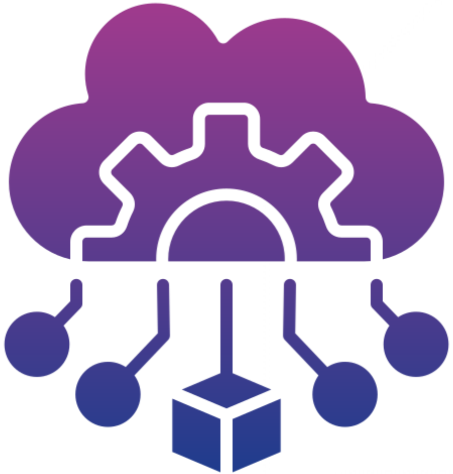






STEPS THAT CAN BE TAKEN TO REDUCE DEFAUTERS AND INCREASE INCOME

- 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

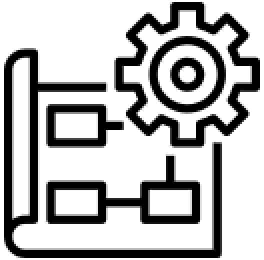
- To develop predictive models that accurately identify defaulters and non-defaulters in the dataset.

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

Data Cleaning

Removed columns with more than 40% missing values to improve data quality and model performance.

Null Value Treatment

Imputed or removed rows with missing values to ensure the completeness and reliability of the dataset.

Outlier Treatment

Identified and handled outliers to reduce their impact on model accuracy and predictions.

Categorical Encoding

Converted categorical features into numerical codes for compatibility with machine learning algorithms.

Feature Scaling

Standardized features using StandardScaler to ensure each feature contributes equally to the model.

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)

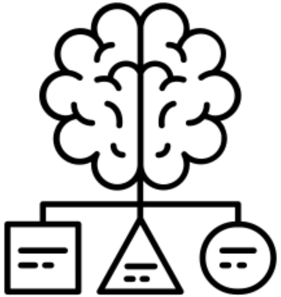
- **Advantages**:
  - Effective in high-dimensional spaces.
  - Works well for both linear and non-linear classification using kernel functions.
- **Disadvantages**:
  - Sensitive to the choice of kernel and regularization parameters.
  - Computationally expensive for large datasets.

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	Accuracy	Precision	Recall	F1 Score	Support
Model 1	66.19%	0.67	0.66	0.66	3566 (Defaulters)
Model 2	60.95%	0.62	0.63	0.62	3566 (Defaulters)
Model 3	60.55%	0.61	0.62	0.62	3566 (Defaulters)
Model 4	66.19%	0.67	0.66	0.66	3566 (Defaulters)
Model 5	64.87%	0.66	0.65	0.65	3566 (Defaulters)
Model 6	66.55%	0.67	0.68	0.68	3566 (Defaulters)

Model 6:

- Highest accuracy (66.55%) and F1 score (0.68); best for identifying defaulters with a good balance between precision and recall.

Models 1 & 4:

- Similar performance with 66.19% accuracy and F1 score (0.66); effective but may miss some defaulters.

Models 2 & 3:

- Lower accuracy (60.95% and 60.55%) and F1 scores (0.62); less effective in distinguishing defaulters.

Conclusion:

- Model 6 is the most reliable for risk analytics, crucial for minimizing missed defaulters while balancing precision and recall.

Final Conclusion

Model Performance:

- The Random Forest classifier was the most effective, achieving an accuracy of 66.55% and an F1 score of 0.68.

Importance of Metrics:

- Balancing precision and recall is crucial in risk analytics to minimize the financial impact of misclassification.

Comparison with Other Models:

- Other models performed adequately but were less effective in accurately identifying defaulters.

Risk Mitigation:

- Using robust algorithms like Random Forest improves financial decision-making and enhances default prediction accuracy.

Future Directions:

- Further exploration of hyperparameter tuning, feature engineering, and alternative modeling techniques could improve predictive performance.

Top features affecting model

- **Employment Duration and Birth Date:**
  - DAYS_EMPLOYED, DAYS_BIRTH
- **Financial and Credit Data:**
  - AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_MON, AMT_GOODS_PRICE, AMT_CREDIT
- **Demographic Information:**
  - NAME_HOUSING_TYPE, CODE_GENDER, NAME_CONTRACT_TYPE, FLAG_OWN_CAR
- **Social and Family Circles:**
  - OBS_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, CNT_FAM_MEMBER

Thank you!