Rakamin x Home Credit Indonesia

30-12-2024

Final Task Project-based Internship

Loan Repayment Success Prediction

Fitria Zusni Farida

Github

PBI Rakamin x Home Credit Indonesia

Overview

Goal

Machine Learning Modeling

Exploratory Data Analysis (EDA)

Model Evaluation

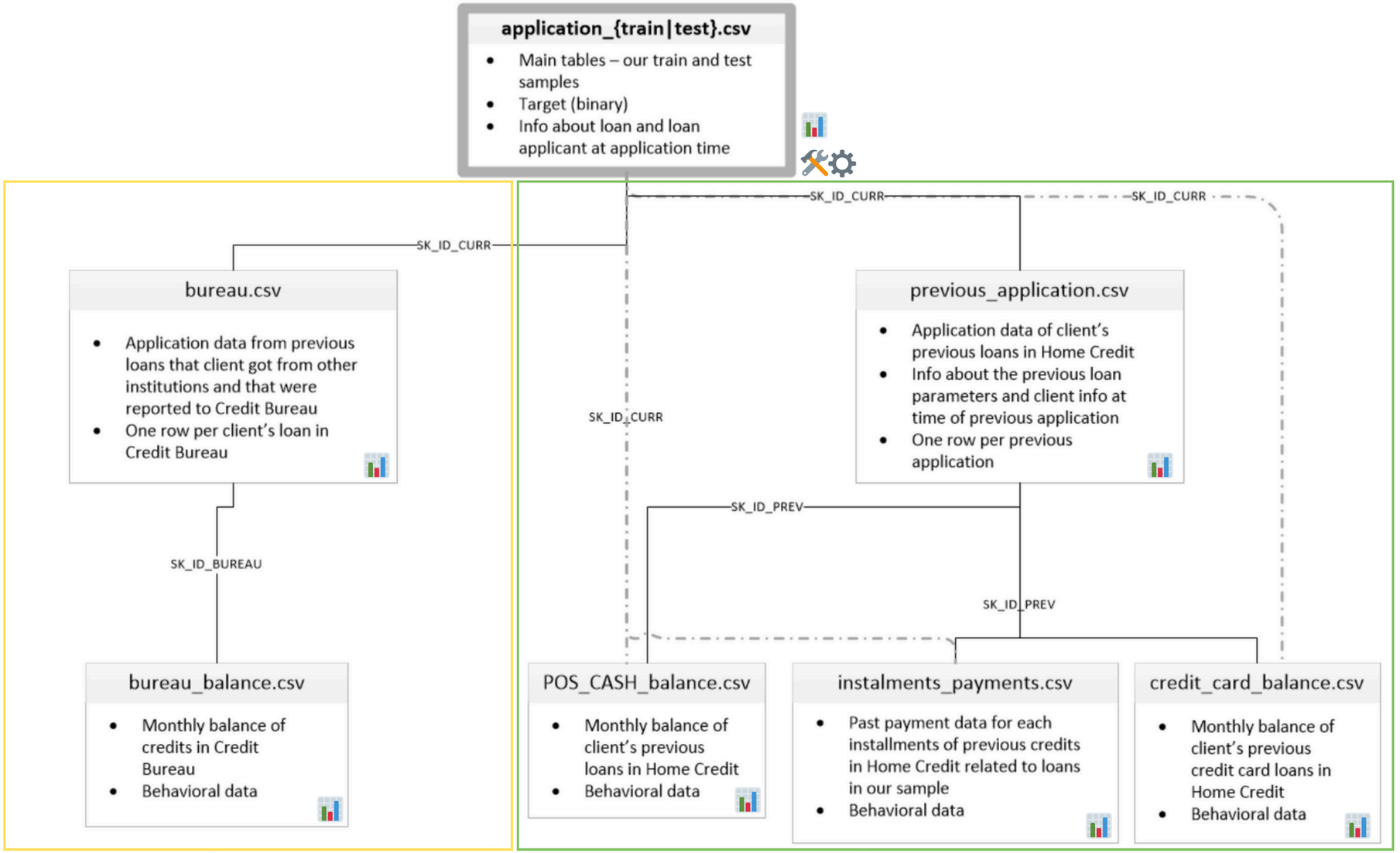
Data Preprocessing

🎯Goal

Predicting Loan Repayment Success and Minimizing Risk

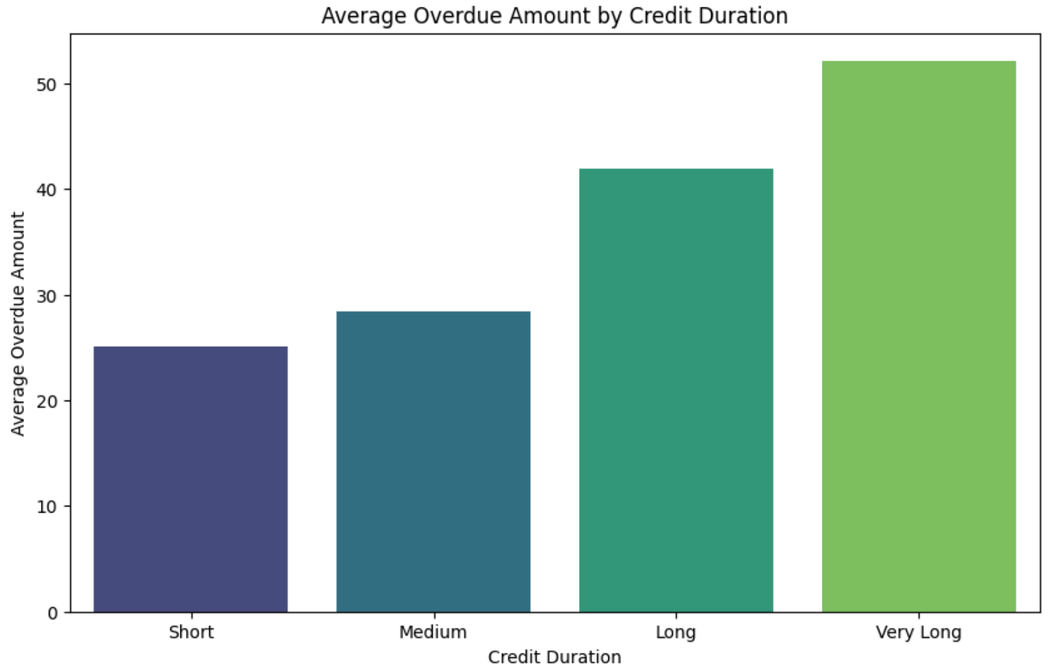
The goal of this research is to analyze and predict customer behavior related to loan repayments, with a focus on ensuring that creditworthy customers are not turned down and that loans are structured in a way that maximizes the likelihood of successful repayments. This will help optimize loan offerings, reduce defaults, and enhance financial inclusivity.

Datasets


application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data
Credit Bureau
Previous Application

📁Credit Bureau

Application data from previous loan that clients got from other institutions and that were reported to Credit Bureau.



🔍**Insights**:

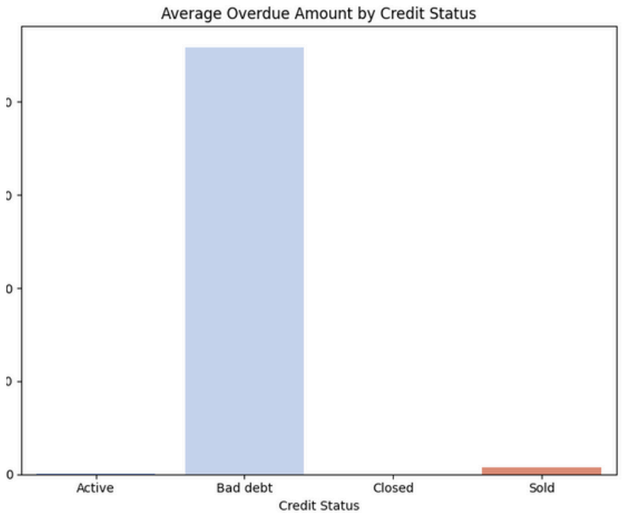
- Longer credit duration showed higher overdue amounts, indicating risk.

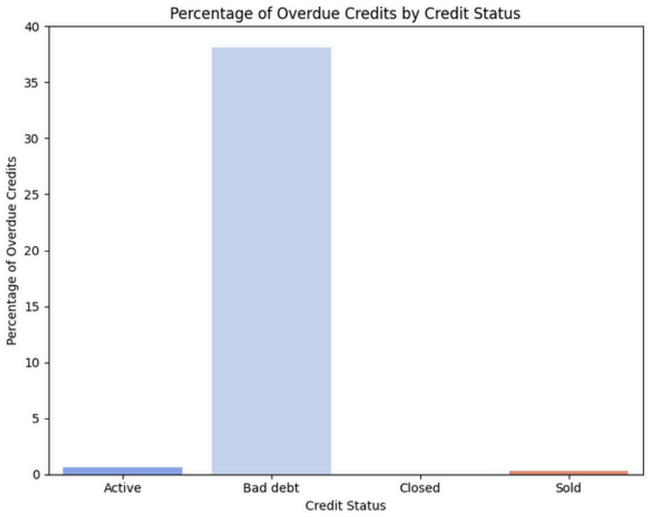
💡**Recommendations:**

- **Stricter Risk Assessment for Long-Term Loans**: Introduce more stringent creditworthiness checks for applicants seeking long-term loans.
- **Dynamic Credit Limits**: Consider implementing dynamic repayment structures, like higher initial payments or balloon payments toward the end.
- **Periodic Financial Health Checks**: Regularly assess the borrower's financial health during the loan tenure to detect early warning signs of default.
- **Offer Financial Planning Support:** Provide resources or consultations to help borrowers plan long-term repayments effectively.

📁Credit Bureau

Application data from previous loan that clients got from other institutions and that were reported to Credit Bureau.





🔍Insights:

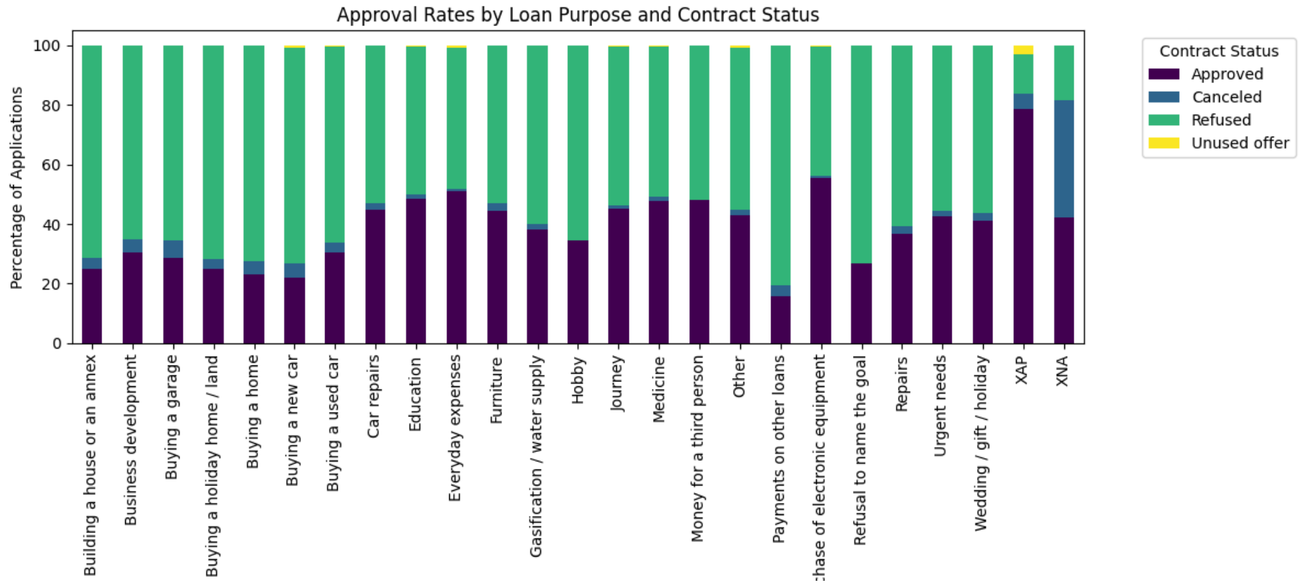
- Bad debt significantly has high percentage of overdue credit among other credit status (active, closed, and sold).

💡Recommendation:

- **Early Intervention Programs:** Proactively monitor signs of overdue payments and engage borrowers early to prevent defaults.
- **Restructure Loan Terms for At-Risk Accounts:** Offer renegotiated repayment terms to reduce overdue risk.
- **Focus on Active Loan Management:** Shift resources to actively monitor and manage at-risk borrowers before they fall into bad debt.
- **Limit Credit Access to High-Risk Borrowers:** Implement tighter credit controls for applicants with a history of overdue debts.

📁Previous Application

Application data from clients previous loans in Home Credit.



🔍**Insights**:

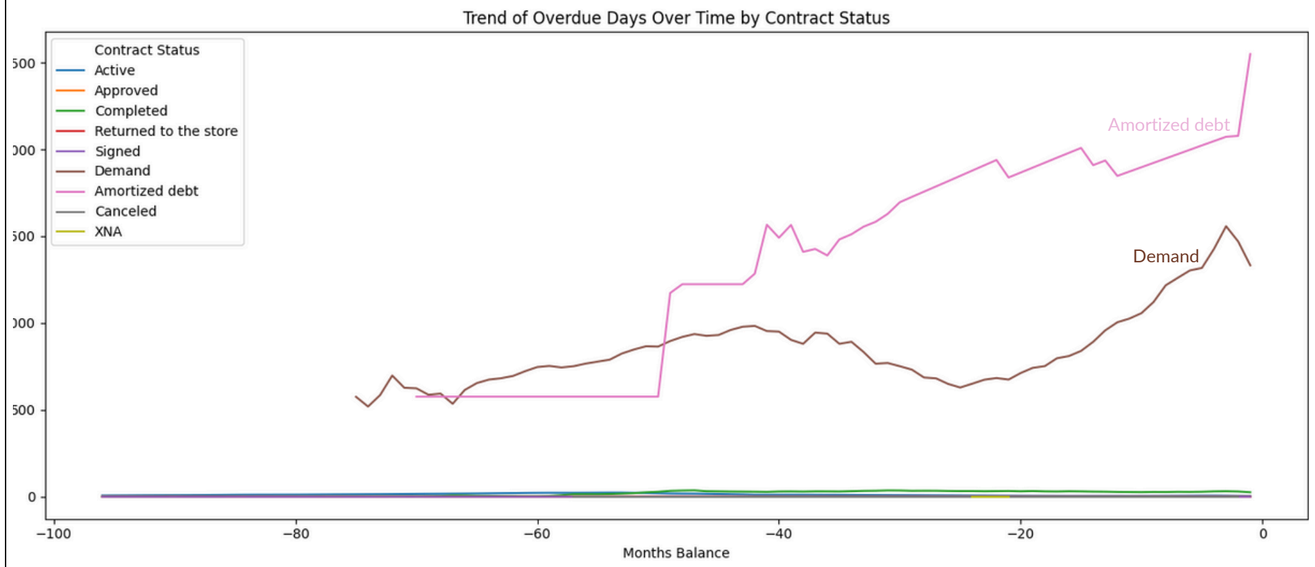
- Highest Approved Loan Purpose: XAP (likely an internal code for a specific loan product or purpose).
- Highest Refused Loan Purpose: Payment on Other Loans (indicating concerns about existing debt obligations).

💡**Recommendations**:

- For XAP Loans: Identify key factors contributing to their high approval rates and replicate these criteria across other loan purposes where feasible.
- For Payment on Other Loans: Conduct a deeper risk analysis to understand why these applications are frequently refused. Consider offering tailored financial counseling or restructuring options to improve approval rates.

📁Previous Application

Application data from clients previous loans in Home Credit.



🔍**Insights**:

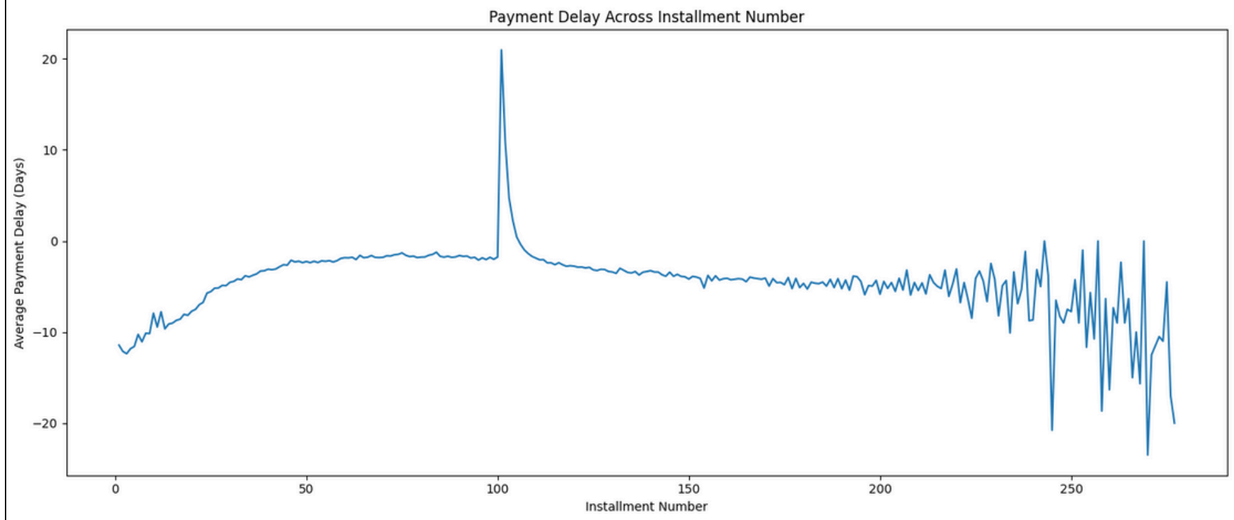
- Amortized debt shows the largest increase in overdue days (SK_DPD) over time, followed by Demand loans.
- This indicates that amortized loans, which are repaid in fixed installments, may have higher overdue risks as time progresses compared to other loans.

💡**Recommendation**:

- Loan repayment plans for amortized debts may need closer monitoring and intervention (e.g., restructuring or financial counseling) to prevent defaults, especially in the later stages of repayment.

📁Previous Application

Application data from clients previous loans in Home Credit.



🔍**Insights**:

- At approximately 100 installments, there was a sudden and sharp increase in overdue days (DAYS_ENTRY_PAYMENT - DAYS_INSTALMENT), peaking at around 20 days overdue before sharply decreasing.
- Between 200 and 250 installments, the trend became highly unstable, showing large fluctuations and even dipping into negative values, indicating early payments or irregular payment behavior.

💡**Recommendation**:

- Investigate the root cause of the spike at 100 installments to understand if it's tied to policy changes or borrower behavior.
- For the 200–250 installment range, further analysis is needed to determine if the issue stems from borrower financial stress or systematic data recording issues.

Data Preprocessing

Application Data
(122 features x 307,511 samples)
Info about loan and loan applicants.

1. EDA for feature selection
2. Handle missing values
3. Handle duplicated data
4. Handle outliers
5. Feature transformation
   (Dimensionality reduction, Encoding, Standardization, and Scaling)
6. Train/Test split
7. Handle class imbalance

modeling

EDA for feature selection

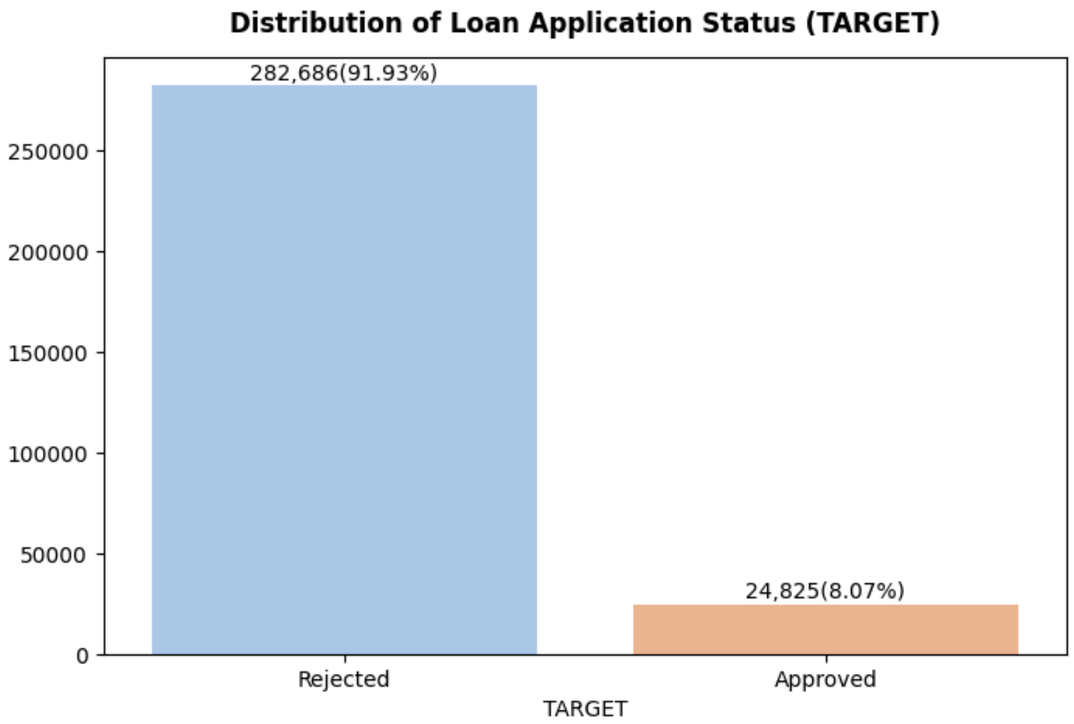
1. Target Analysis

🔍Insights

- There is significant class imbalanced with 91.93% clients' application were rejected and only 8.07% client's application were approved.

💡Actions

- Use resampling techniques (oversampling or undersampling)



EDA for feature selection

2. Descriptive Analysis (numerical features)

	count	mean	std	min	25%	50%	75%	max	skewness	kurtosis	variance	outliers (%)
AMT_INCOME_TOTAL	307511.0	168797.919297	237123.146279	25650.00000	112500.000000	147150.00000	202500.000000	1.170000e+08	391.559654	191786.554381	5.622739e+10	4.564064
AMT_CREDIT	307511.0	599025.999706	402490.776996	45000.00000	270000.000000	513531.00000	808650.000000	4.050000e+06	1.234778	1.934041	1.619988e+11	2.133907
AMT_ANNUITY	307499.0	27108.573909	14493.737315	1615.50000	16524.000000	24903.00000	34596.000000	2.580255e+05	1.579777	7.707320	2.100684e+08	2.440238
AMT_GOODS_PRICE	307233.0	538396.207429	369446.460540	40500.00000	238500.000000	450000.00000	679500.000000	4.050000e+06	1.349000	2.431916	1.364907e+11	4.789422
REGION_POPULATION_RELATIVE	307511.0	0.020868	0.013831	0.00029	0.010006	0.01885	0.028663	7.250800e-02	1.488009	3.260065	1.913043e-04	2.735512

Statistics descriptive of numerical features (displayed only 5 features)

🔍Insights

- There are features with **highly skewed distribution** (eg. AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, REGION_POPULATION_RELATIVE, etc)
- There are features with **very low variance** (eg. REGION_POPULATION_RELATIVE, BASEMENTAREA_AVG, YEARS_BEGINEXPLUATATION_AVG, etc)

💡Actions

- Handle highly skewed features by winsorizing to remove outliers and feature transformation (standardization and scaling).
- Drop low variance features as they don't give significant information to the model.

EDA for feature selection

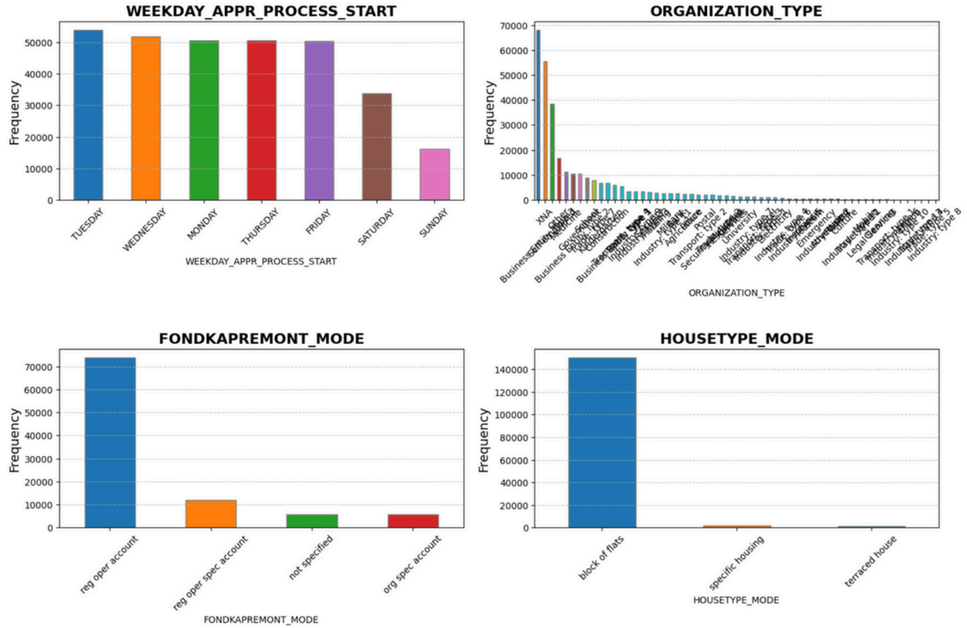
2. Descriptive Analysis (categorical features)

🔍Insights

- There exist time-based features.
- There exist features with significant cardinality.
- There exist features with significant imbalance.

💡Actions

- Drop time-based features to reduce complexity.
- Do context-based transformation to reduce the number of dimension of the feature.



Data distribution for each category of categorical features
(displayed only 4 features)

EDA for feature selection

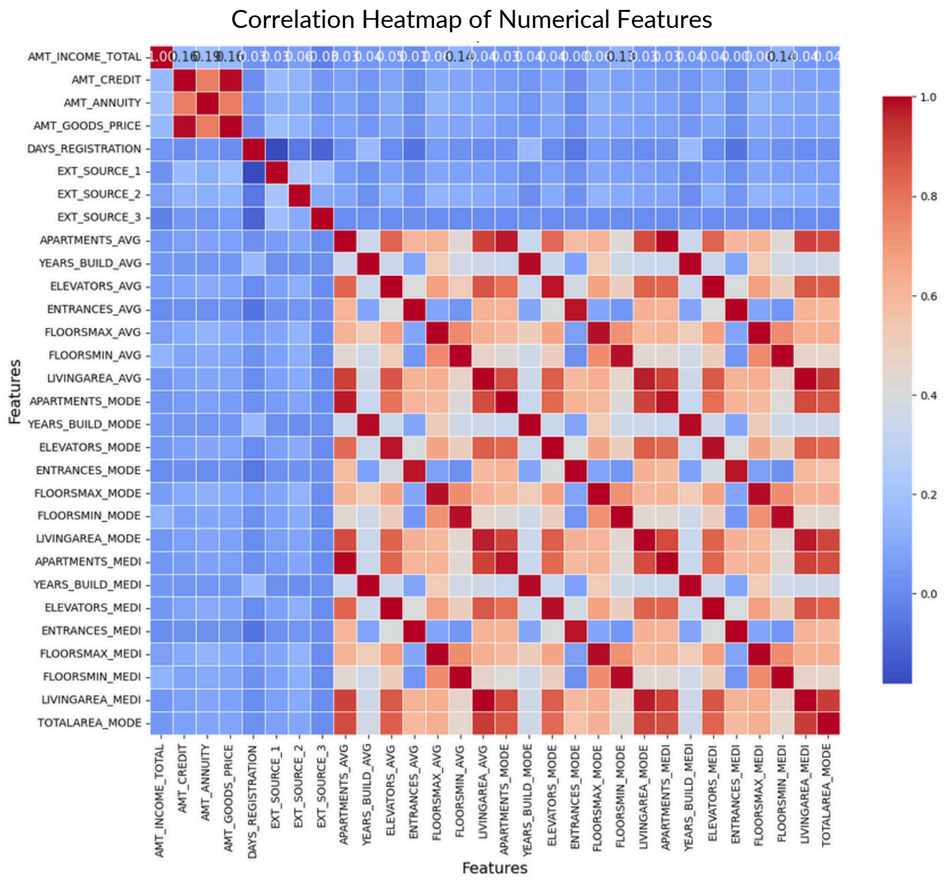
3. Correlation Analysis

🔍Insights

- There exist pairs of features with high correlation (0.7)

💡Actions

- Drop or do feature engineering to combine multiple features into one new feature.



Handle missing values

🔎**Insights**

- There exist feature with high number of missing values (>18%).
- There exist feature with low number of missing values (<=18%)
- Some features don’t have missing values.

🛠️**Actions**

- Drop feature with high number of missing values (>18%).
- Fill the missing value with the median value for feature with low number of missing values (<=18%)

	Feature	Missing Count	Missing Percentage (%)
0	EXT_SOURCE_3	60965	19.825307
1	EXT_SOURCE_2	660	0.214626
2	AMT_GOODS_PRICE	278	0.090403
3	AMT_ANNUITY	12	0.003902
4	CNT_FAM_MEMBERS	2	0.000650

Features with missing values

Handle duplicated data

🔎**Insights**

- There are only 9 duplicated data.

🛠️**Actions**

- Drop the duplicated data and keep the first data.

Drop duplicates

```
# Remove duplicates
new_train_app_df = new_train_app_df.drop_duplicates(keep='first')

# Detect duplicate rows based on the 'features' list
duplicate_rows = new_train_app_df.duplicated(keep='first')

# Count the number of duplicates
num_duplicates = duplicate_rows.sum()
print(f"Number of duplicate rows: {num_duplicates}")
```

```
Number of duplicate rows: 0
```

Outliers handling

🔎**Insights**

- CNT_CHILDREN, CNT_FAM_MEMBERS, AMT_ANNUITY, AMT_GOOD_PRICE have outliers, with <5% of them are below lower threshold (Q1-1.5*IQR) or above upper threshold (Q3+1.5*IQR).

🛠️**Actions**

- Do winsorizing technique, to cap the data with outliers with the threshold.

	outlier_count	outliers (%)
CNT_CHILDREN	4272	1.389259
CNT_FAM_MEMBERS	4007	1.303081
AMT_ANNUITY	7504	2.440309
AMT_GOODS_PRICE	14728	4.789562
EXT_SOURCE_2	0	0.000000

Outliers in the numerical features

Feature Transformation

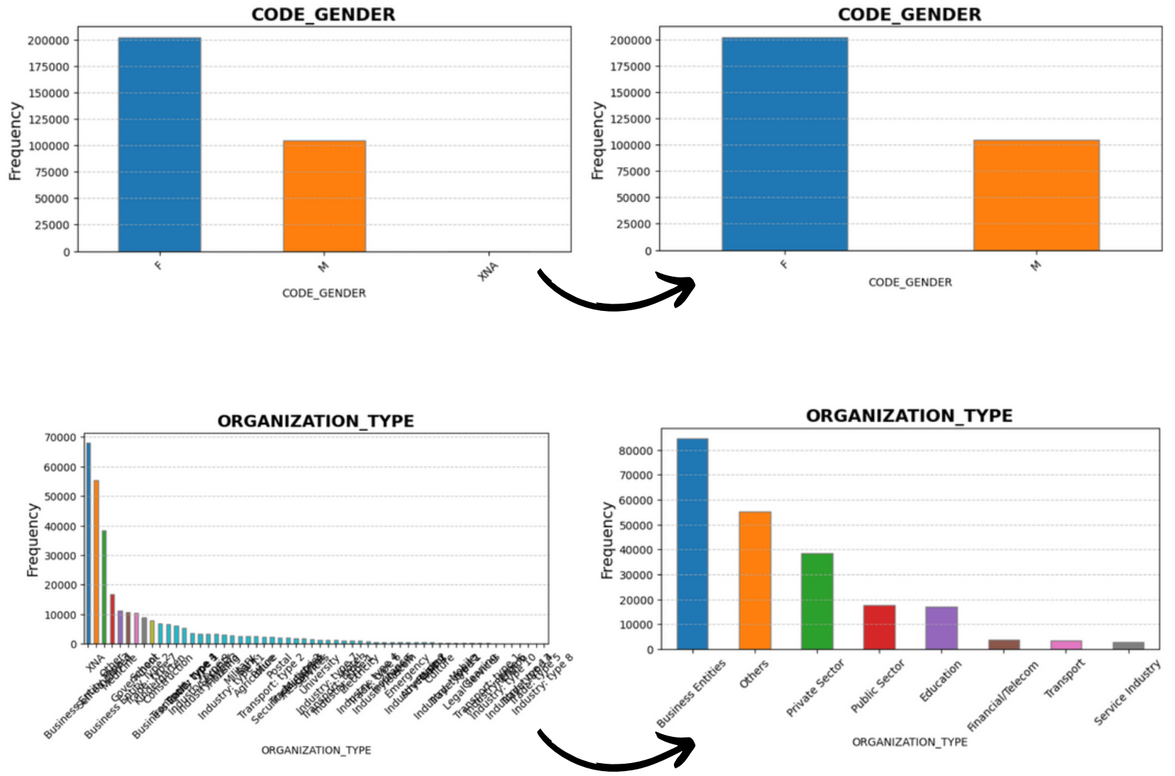
1. Dimensionality reduction (high cardinality categorical data)

🔎Insights

- CODE_GENDER has three categories (F, M, XNA), with XNA significantly underrepresented.
- ORGANIZATION_TYPE has 23 categories with imbalance distribution across categories.
- NAME_INCOME_TYPE and NAME_EDUCATION_TYPE have significant imbalance across categories.

🛠️Actions

- CODE_GENDER: replace XNA with the modus of the data (F)
- ORGANIZATION_TYPE: perform context-based grouping. That is, grouping multiple categories into one categories.
- NAME_INCOME_TYPE and NAME_EDUCATION_TYPE: group rare categories into one category callend ‘Other’.



Dimensionality Reduction of features

Feature Transformation

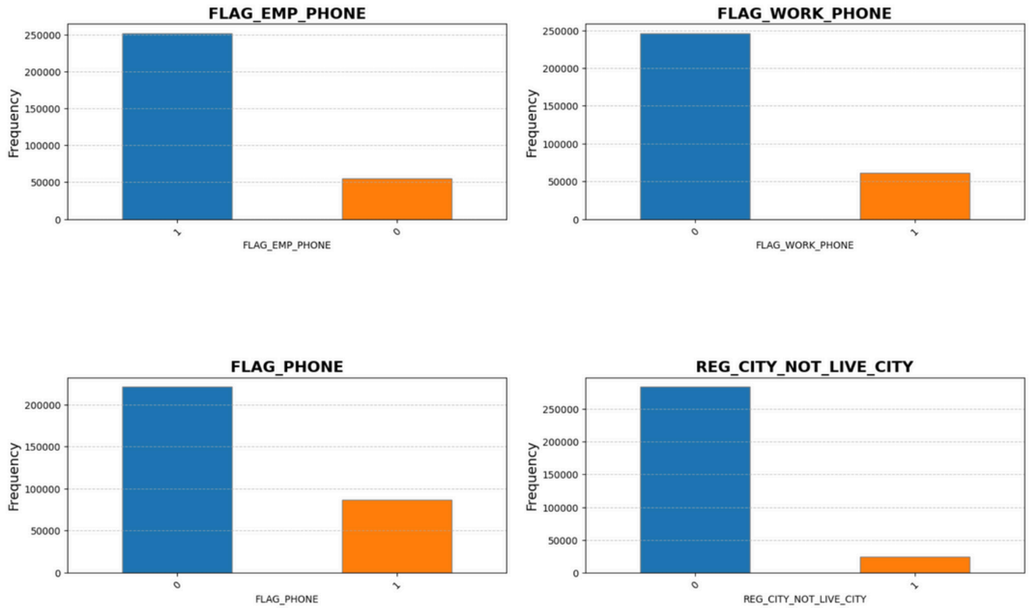
2. Feature Encoding (categorical data)

🔎Insights

- **Features with two categories**: CODE_GENDER, NAME_CONTRACT_TYPE, FLAG_OWN_CAR, FLAG_OWN_REALTY, FLAG_EMP_PHONE, FLAG_WORK_PHONE, FLAG_PHONE, REG_CITY_NOT_LIVE_CITY, REG_CITY_NOT_WORK_CITY, LIVE_CITY_NOT_WORK_CITY, FLAG_DOCUMENT_3, FLAG_DOCUMENT_6
- **Features with more than two categories**: NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, NAME_HOUSING_TYPE, ORGANIZATION_TYPE

🛠️Actions

- Binary label encoding for features with two categories.
- One-hot encoding for features with more than two categories.



Feature encoding of categorical data
(showed only 4 features)

Feature Transformation

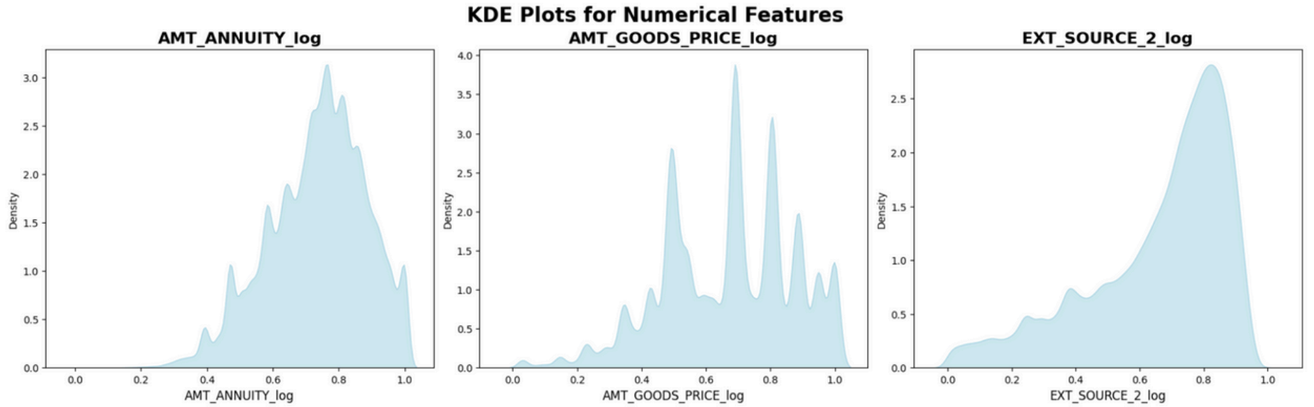
3. Standardization and Scaling (numerical features)

🔎Insights

- **Features with skewed distribution**: AMT_ANNUITY, AMPT_GOODS_PRICE, EXT_SOURCE_2

🛠️Actions

- Normalize the distribution using log-p transformation (p=1).
- Perform Min-Max scaling for better robustness of the model.



Train/Test split

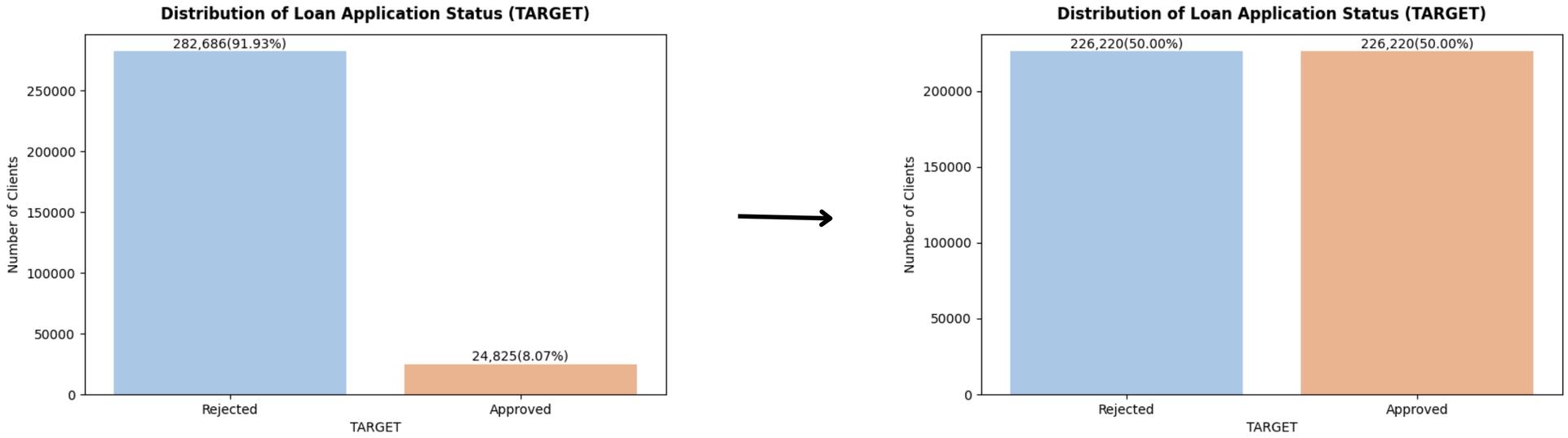
Training set
452,440 samples
80%

Testing set
61,500 samples
20%

The final dataset contains 39 features:

'NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_PHONE', 'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY', 'LIVE_CITY_NOT_WORK_CITY', 'FLAG_DOCUMENT_3', 'FLAG_DOCUMENT_6', 'CNT_CHILDREN', 'CNT_FAM_MEMBERS', 'NAME_INCOME_TYPE_Other', 'NAME_INCOME_TYPE_Pensioner', 'NAME_INCOME_TYPE_State servant', 'NAME_INCOME_TYPE_Working', 'NAME_EDUCATION_TYPE_Incomplete higher', 'NAME_EDUCATION_TYPE_Other', 'NAME_EDUCATION_TYPE_Secondary / secondary special', 'NAME_FAMILY_STATUS_Married', 'NAME_FAMILY_STATUS_Separated', 'NAME_FAMILY_STATUS_Single / not married', 'NAME_FAMILY_STATUS_Widow', 'NAME_HOUSING_TYPE_House / apartment', 'NAME_HOUSING_TYPE_Municipal apartment', 'NAME_HOUSING_TYPE_Office apartment', 'NAME_HOUSING_TYPE_Rented apartment', 'NAME_HOUSING_TYPE_With parents', 'ORGANIZATION_TYPE_Education', 'ORGANIZATION_TYPE_Financial/Telecom', 'ORGANIZATION_TYPE_Others', 'ORGANIZATION_TYPE_Private Sector', 'ORGANIZATION_TYPE_Public Sector', 'ORGANIZATION_TYPE_Service Industry', 'ORGANIZATION_TYPE_Transport', 'AMT_ANNUITY_log', 'AMT_GOODS_PRICE_log', 'EXT_SOURCE_2_log'

Handle class Imbalance



Synthetic Minority Oversampling Technique (SMOTE) for training set

⚙️Machine Learning Modeling

1. **Models**

   📈Logistic Regression

   🌳Decision Tree

   🕸️K-Nearest Neighbors (KNN)

⚙Machine Learning Modeling

2. Model Evaluation

	Accuracy		Precision		Recall		F1-score	
Model	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**
Logistic Regression	0.66233	0.67107	0.66738	0.13650	0.64723	0.56511	0.65718	0.219891
Decision Tree	1.0	0.812	**1.0**	0.115	**1.0**	0.192	**1.0**	0.144
KNN	0.87859	0.68856	0.82285	0.10636	0.96493	0.37780	0.88824	0.16599

Training and test performances measured by accuracy, precision, recall, and F1-score.

⚙️Machine Learning Modeling

3. Evaluation and recommendations

📈**Logistic Regression:**

- Training and test accuracy are relatively close, but test precision and F1 score are low.
- Issue: Model struggles to identify positive cases correctly.
- Recommendation: Tune hyperparameters.

🌳**Decision Tree:**

- Training accuracy is perfect, but test performance is significantly lower, indicating severe overfitting.
- Recommendation: Prune the tree, limit its depth, or use ensemble methods like Random Forest or Gradient Boosting.

🕸️**KNN:**

- Training accuracy is high, but test accuracy and precision are low.
- Issue: Overfitting and poor generalization to unseen data.
- Recommendation: Tune hyperparameters, particularly n_neighbors.

	Accuracy		Precision	
Model	Training	Test	Training	Test
Logistic Regression	**0.66233**	**0.67107**	0.66738	0.13650
Decision Tree	1.0	0.812	1.0	0.115
KNN	0.87859	0.68856	0.82285	0.10636

	Recall		F1-score	
Model	Training	Test	Training	Test
Logistic Regression	0.64723	0.56511	0.65718	0.219891
Decision Tree	1.0	0.192	1.0	0.144
KNN	0.96493	0.37780	0.88824	0.16599

Training and test performances measured by accuracy, precision, recall, and F1-score

⚙️Machine Learning Modeling

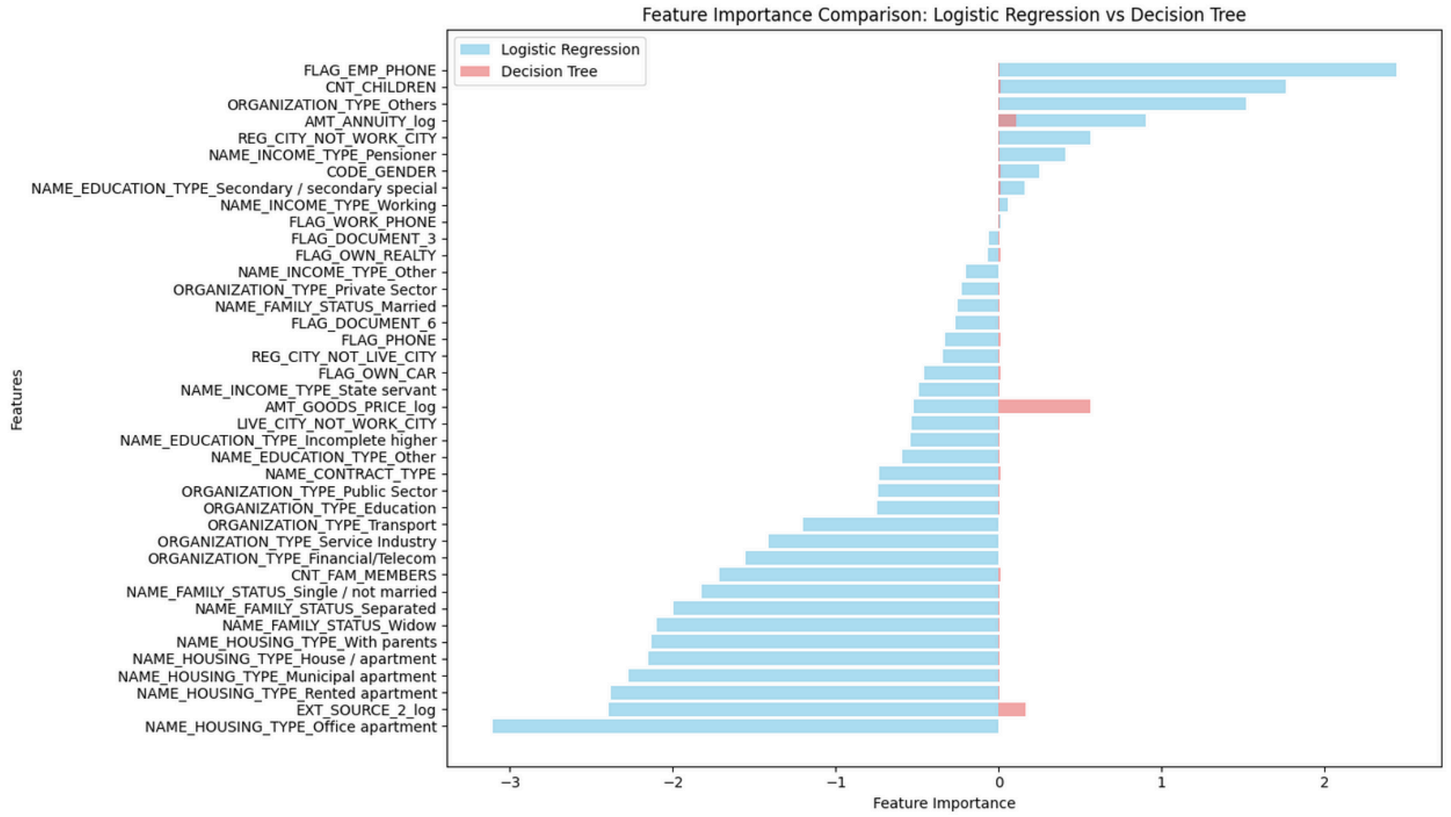
4. Feature importance

🔍Insights

- **Features with the most importance:** NAM_HOUSING_TYPE, EXT_SOURCE_2, FLAG_EMP_PHONE, CNT_CHILDREN, NAME_INCOME, and ORGANIZATION_TYPE.
- **Features with the least importance:** FLAG_WORK_PHONE, FLAG_DOCUMENT_3, and FLAG_OWN_REALTY.

💡Recommendation

- Consider to remove the least important features to reduce model complexity.



30-12-2024

Thank you!