Home Credit Default Risk Analysis

Daryle Bilog
Joe Sarnello
Sanskriti Bhargava
Vinay Kumar Vascuri

Introduction

Business Problem:

- Identify underserved individuals
- Assess repayment ability
- Predict high-risk clients

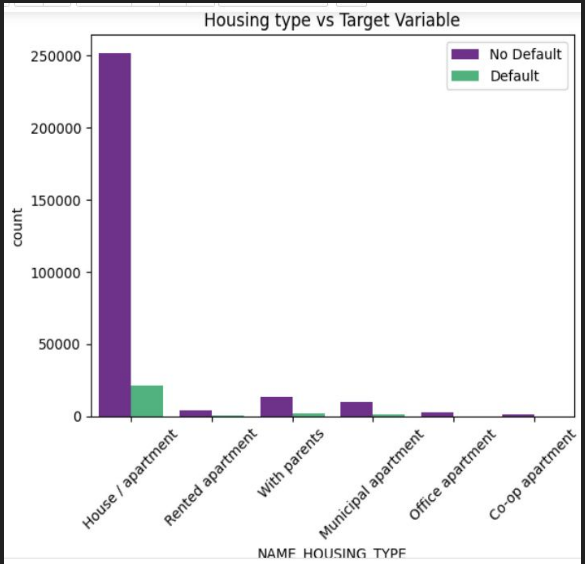
Objective

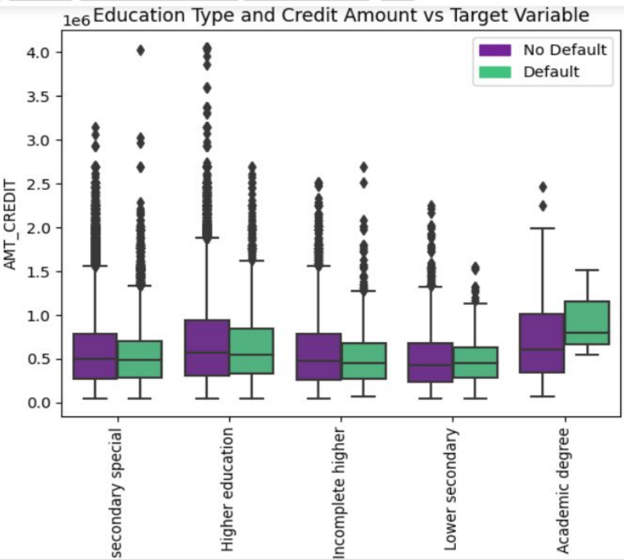
Build a model to:

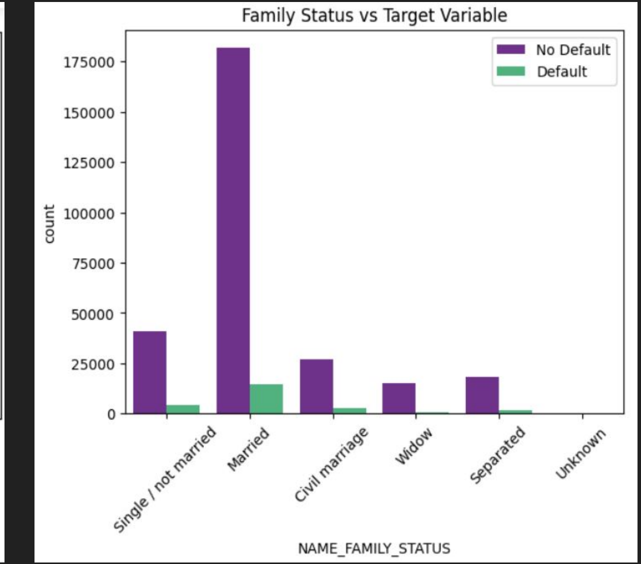
- Effectively extend loans to creditworthy individuals and decline those with insufficient repayment ability.

- Accurately assess and suggest the Home credit group regarding the applicant’s loan repayment capability.

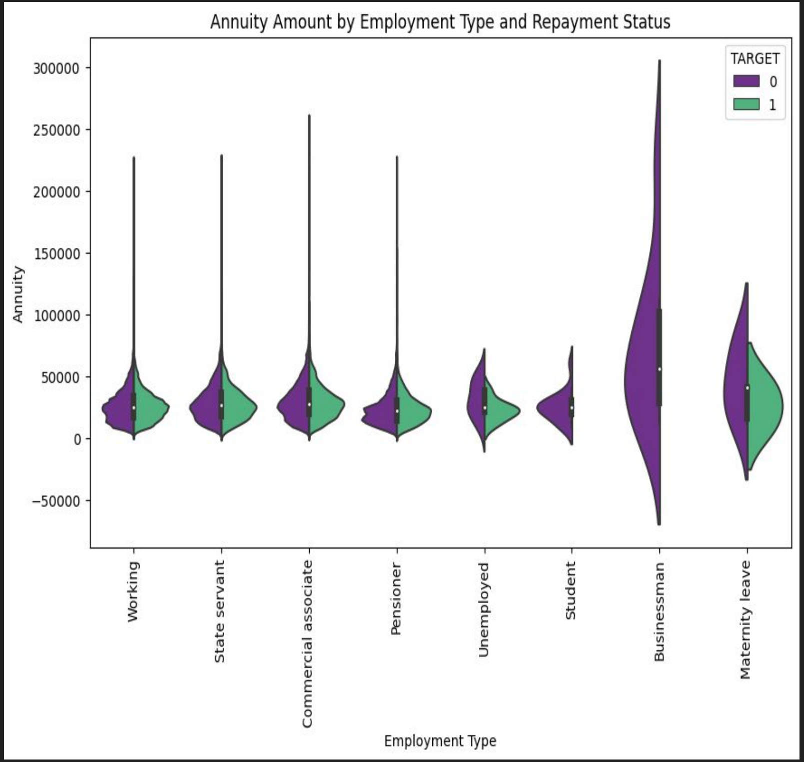
Feature-Target Visualization

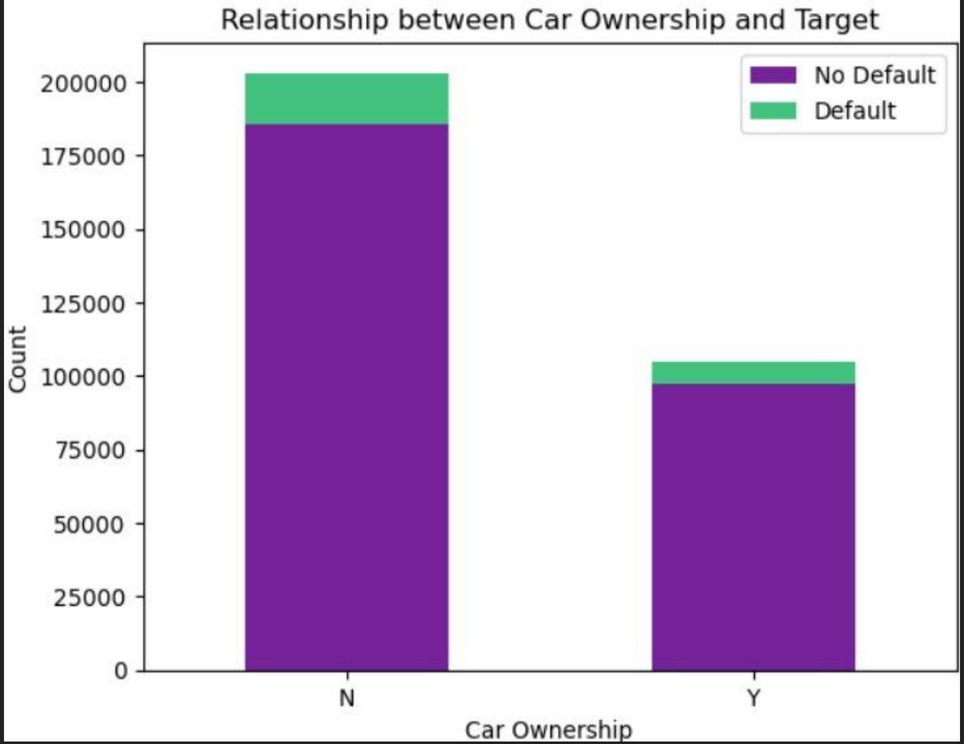

Housing type vs Target Variable


Education Type and Credit Amount vs Target Variable


Family Status vs Target Variable

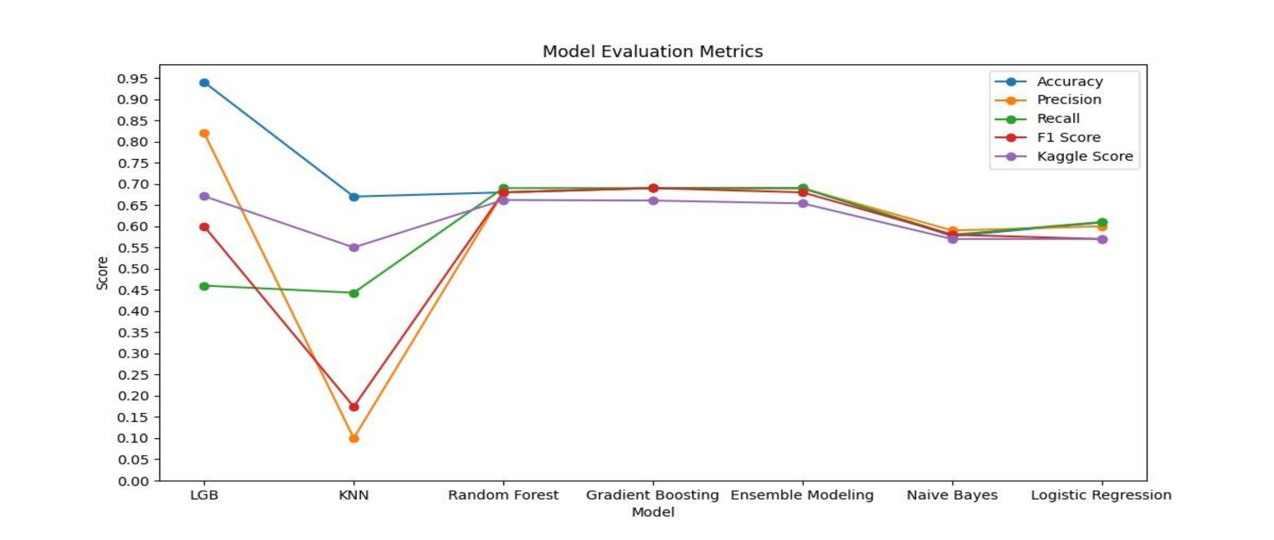
Feature-Target Visualization



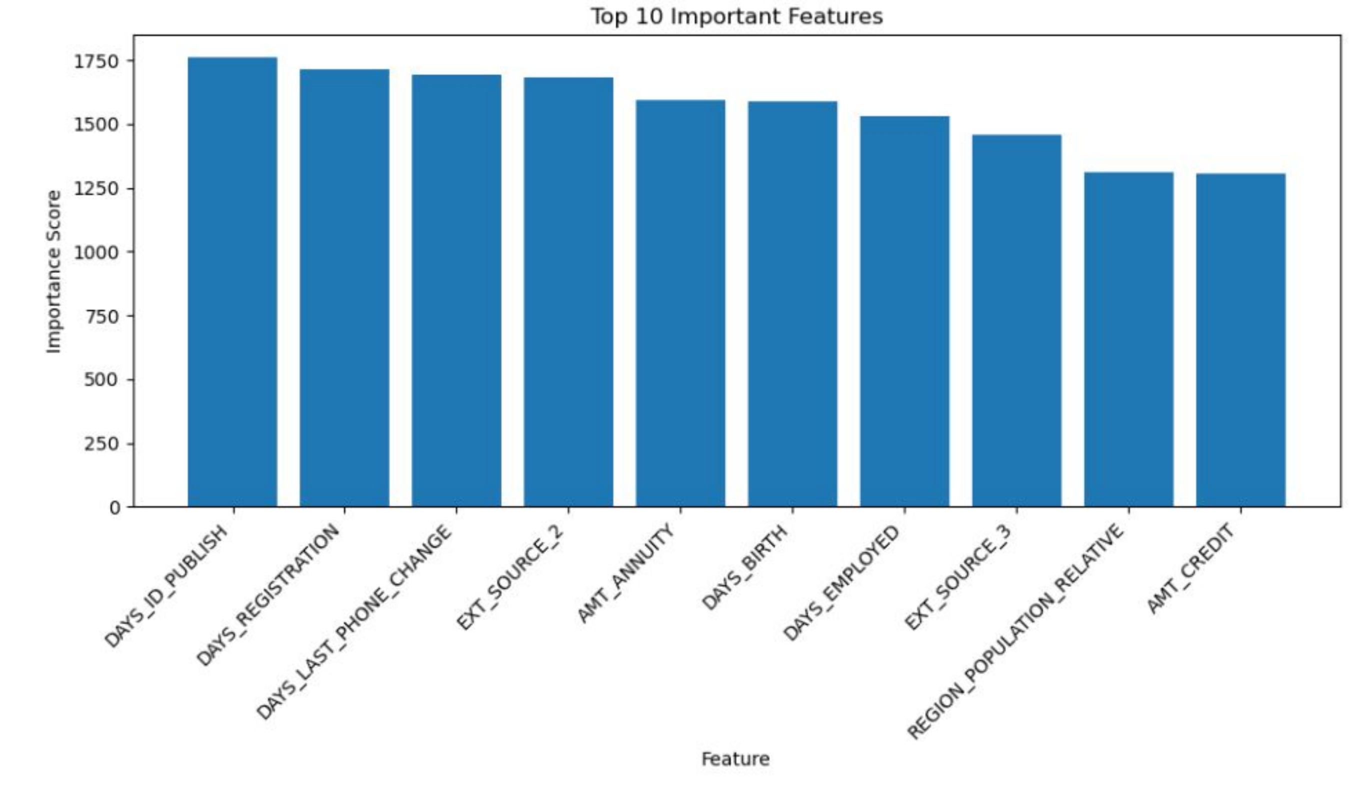


Models Implemented

1. Light Gradient Boosting
2. KNN
3. Random Forest
4. Gradient Boosting
5. Ensemble Modeling (Random Forest, Gradient Boosting, Logistic Regression)
6. Naive Bayes
7. Logistic Regression



Important Features



Light Gradient Boosting Machine Model(LGB)

Important Features:

DAYS_ID_PUBLISH

DAYS_REGISTRATION

DAYS_LAST_PHONE_CHANGE

Less Likely to Default

Married

Higher Education

Businessmen

Owns a car

Owns a house or an apartment

Thank you!

Any Questions?

Data Cleaning Steps (Appendix part 1)

- Dropped columns with 30% or more of their values being null
- Dropped 4 rows that had “XNA” as their value in the Code_Gender column
- Dropped the 1 row with a value of 117000000 in the AMT_Income_Total column
- Replaced the remaining nulls in the data:
  - Numerical columns: Replaced nulls with the median value of the column
  - Categorical columns: Replaced nulls with the most frequent value of the column
- Used Winsorization to handle outliers / normalize the data

Model Building (Appendix part 2)

1. Undersampling using Random Under Sampler
2. Oversampling using SMOTE
3. Train Test Split
4. Model Fitting/Ensemble Modelling
5. Hypertuning
6. Evaluation

Comparison of Models (Appendix part 3)

	LGB	KNN (K=3)	Random Forest	Gradient Boosting	Ensemble Modelling	Naive bayes	Logistic Regression
Accuracy	94%	67%	68%	69%	69%	57.7%	60.9%
Precision	82%	10%	68%	69%	69%	59%	60%
Recall	46%	44.31%	69%	69%	69%	58%	61%
F1 score	60%	17.42%	68%	69%	68%	58%	57%
Kaggle Score	67.1%	55%	66.2%	66.07%	65.4%	57%	57%

Results (Appendix part 4)

1. Top Performers: LGB and Random Forest
2. Competitive model: Gradient Boosting
3. Ensemble Approach
4. Baseline Models: Logistic Regression and Naive Bayes
5. Underperforming model: KNN