Home Credit Risk Prediction

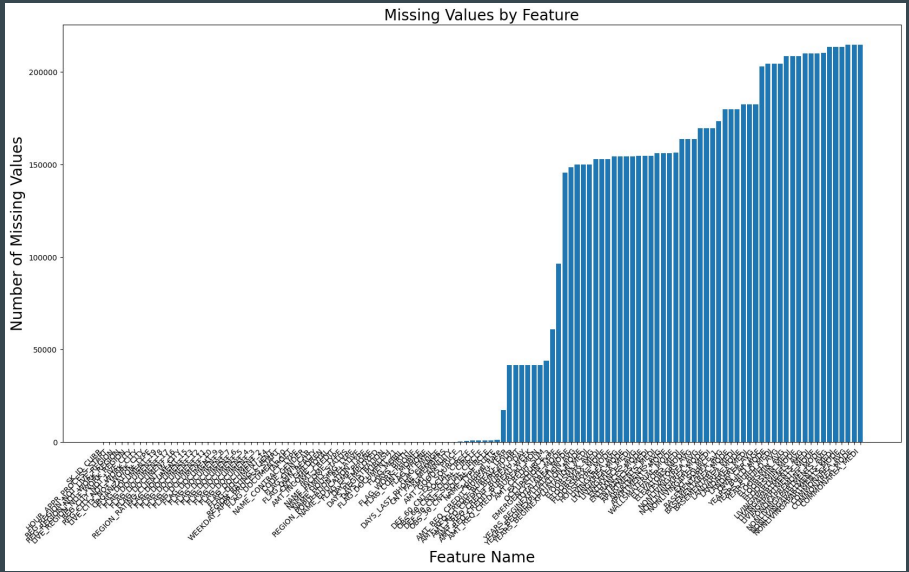
Zachary Galante

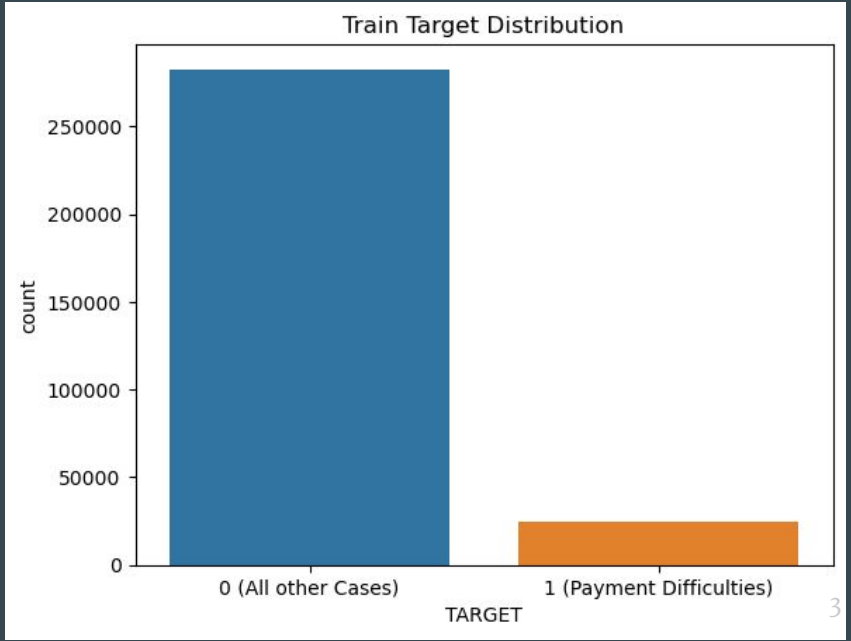
Overview

- Predict a client's repayment abilities for a loan. (Target)
    - Will a client have difficulties with their payments?
- Allows underserved communities to safely get loans.
- Leverage data from multiple external sources
    - Previous reported loans
    - Previous applications
    - Credit card payments

EDA

- 1 pre-split main application file from Home Credit
  - Training set (307,511 x 122) Test Set (48,744 x 121)
- 6 related tables with external data on clients.
- Class Imbalance for target variable
- Missing values for many features





EDA Cont.

Bottom 5 correlated variables

Column Name	Correlation
Ext_Source3	-0.179
Ext_Source2	-0.160
Ext_Source1	-0.155
Days_Employed	-0.045
Floorsmax_Avg	-0.044

Top 5 correlated variables

Column Name	Correlation
Days_Birth	0.078
Region_Rating_Client_W_City	0.061
Region_Rating_Client	0.059
Days_Last_Phone_Change	0.055
Days_Id_Publish	0.051

Data Cleaning

- Missing Values
  - Kept columns that included 70% or more of the data
  - Numeric variables → Mean
  - Categorical variables → Most Frequent Value
- SMOTE
  - Resampled to a 35/65 ratio with the minority class
- Encoding
  - Label Encoder for categorical variables

Feature Engineering

1. Average Credit of Previous Loans
   - Average of previous credit loans reported to the credit bureau.
2. Previous Applications with the company
   - Approval Rate
   - Number of applications
3. Average difference of loan payments
   - Installment Payment - Amount Paid
4. Max Credit
   - The highest credit balance of a client

Model Building

- K Fold Cross Validation
  - 5 folds

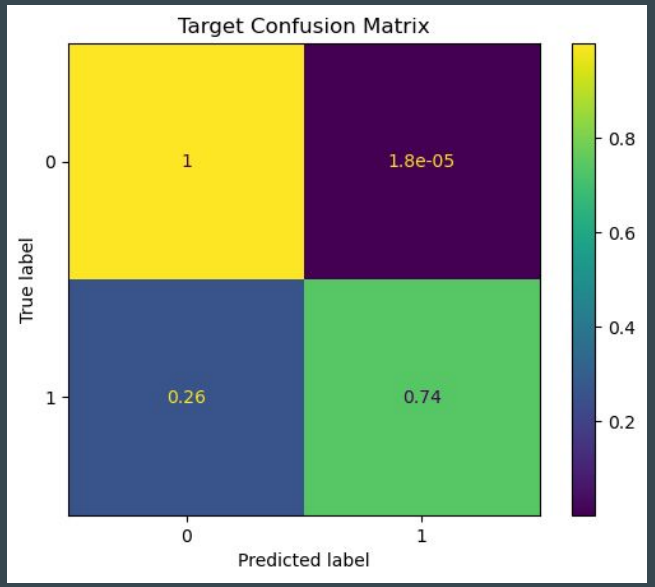
Feature Selection:

- LASSO (L1 Regularization)
  - Removed 2 features (Flag_Document_10, Flag_Document_12)

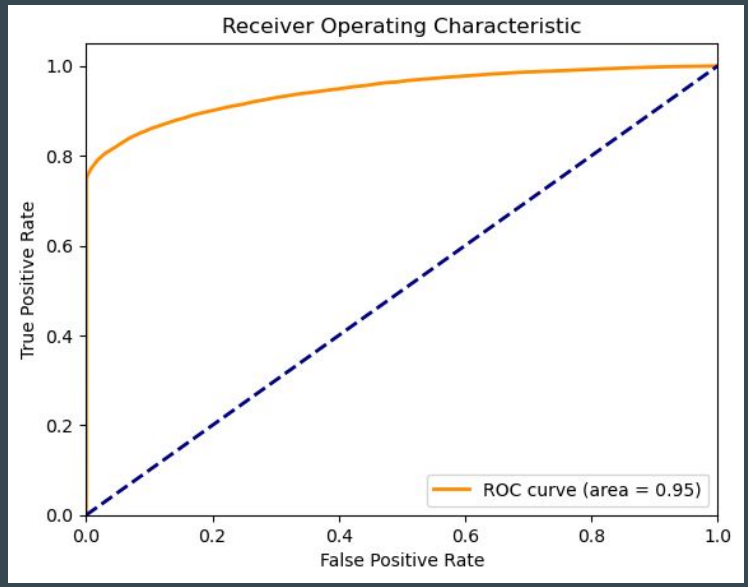
Model Comparison

Model Name	Accuracy Score	F1 Score	AUC	Training Time Seconds
Logistic Regression	0.74	0.00	0.63	3.47
Decision Tree	0.86	0.74	0.83	7.70
Random Forest	0.93	0.85	0.95	230.37
Neural Network	0.74	0.53	0.50	181.59

Best Performing Model: Random Forest



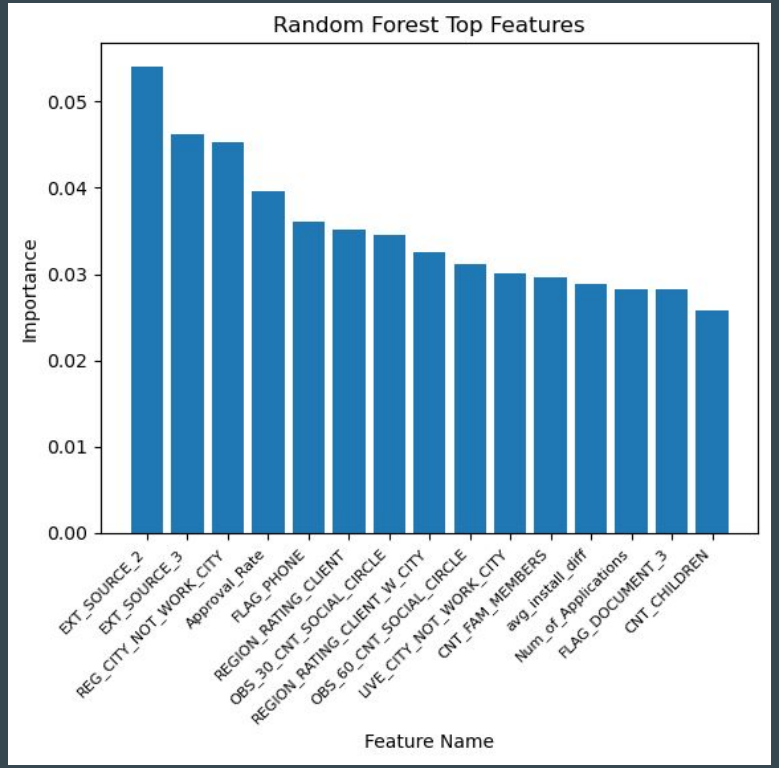
Metric	Score
Accuracy	0.94
F1 Score	0.85
AUC	0.95



Kaggle Test Score: 0.687

Recommendations

- Continue to look at previous applications
    - Approval Rate
- Look at the clients social circle
    - Social Media info
    - Previous applications for members of that circle
- Explore where the client works
    - Differences in addresses



Limitations and Future Work

- Limitations
  - SVM and Grid Search took long
  - Lasso didn't have promising results
- Future Work
  - Feature Engineering
    - Graph based features
      - Understanding relationships in client's social circles.
  - More Data
  - More resources

Thank You!