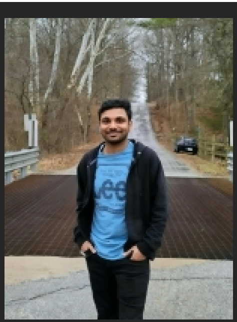
Luddy School of Informatics, Computing, and Engineering

Phase 2- HOME CREDIT DEFAULT RISK- GROUP 06



Deepak Kasi Nathan
(dekasi@iu.edu)

Sai Sumanth Muvva
(saimuvva@iu.edu)

Viswa Suhaas Penugonda
(vpenugon@iu.edu)

Teja Naidu Chintha
(tnchinth@iu.edu)

INDIANA UNIVERSITY BLOOMINGTON

Presentation overview

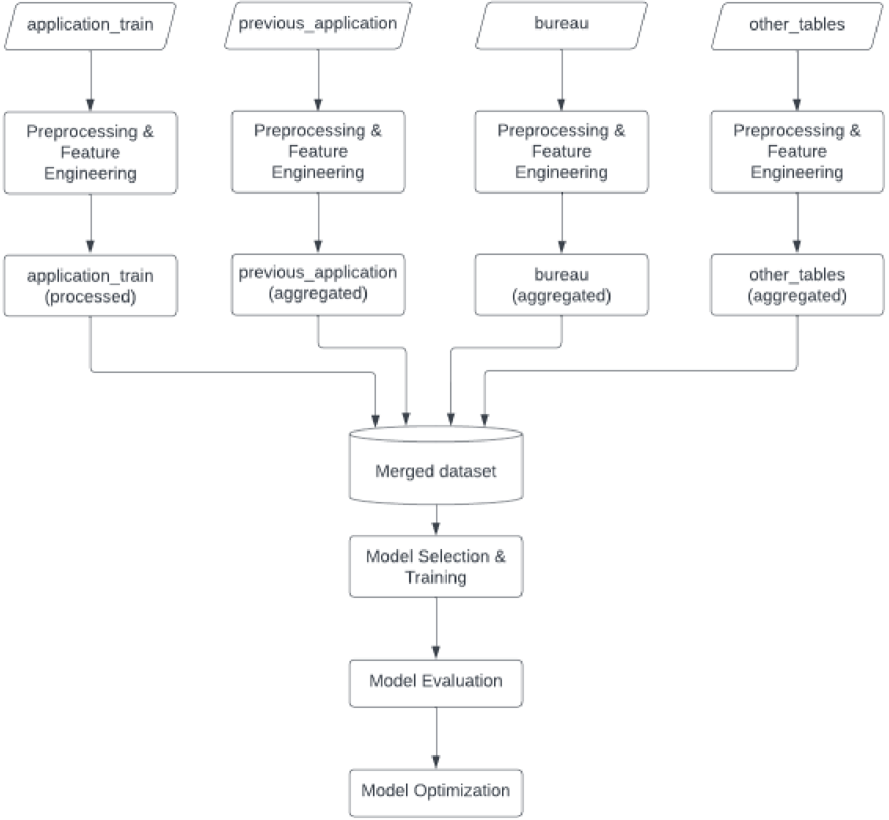
- Project Description
- EDA and Visual EDA Summary
- Overview of Pipelines Implemented
- Discussion and Results
- Conclusion
- Next Steps (4P’s)

Project Description

During this project phase, we obtained data from the Kaggle competition "Home Credit Risk Analysis" and conducted Exploratory Data Analysis to examine and understand the dataset. We generated various visualizations for most input features related to the "Target" variable to identify individuals at the highest risk.

Performed feature engineering on the tables (bureau, installments_payments, credit_card_balance, and previous_application). In this process, we created new features by grouping data based on its primary key and applying the mean as an aggregate function on several crucial columns related to the domain. We employed a column transformer to consolidate all features for use in the pipeline. We applied feature engineering and hyper-parameter tuning on basic models such as Logistic Regression, Lasso Regression, SGD and compared their performances.

WORKFLOW



application_train

previous_application

bureau

other_tables

Preprocessing & Feature Engineering

Preprocessing & Feature Engineering

Preprocessing & Feature Engineering

Preprocessing & Feature Engineering

application_train (processed)

previous_application (aggregated)

bureau (aggregated)

other_tables (aggregated)

Merged dataset

Model Selection & Training

Model Evaluation

Model Optimization

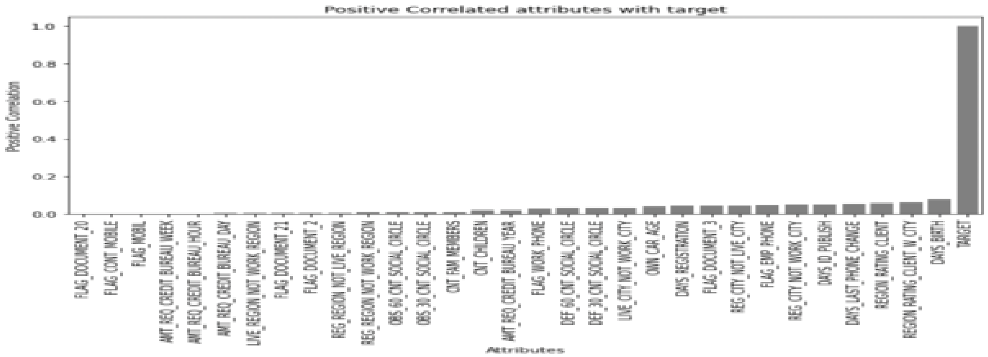
Exploratory Data Analysis.

We were able to describe the key characteristics of the data set using statistical graphics and other data visualization approaches with the help of exploratory data analysis. We explored the following on the dataset:

1. The Data types and General Statistics of data.
2. Number of Missing values (percentage of the missing values.)
3. Numerical and Categorical Data.
4. We also visualized the missing data for each dataset
5. Correlation of the numerical data with the “Target column”.

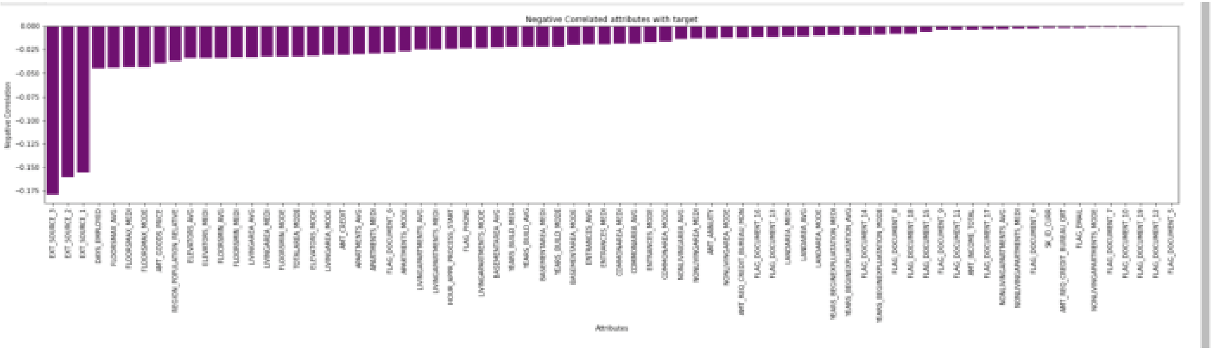
Visual Exploratory Data Analysis

Visual Exploratory data Analysis performed on Categorical values to understand their significance in data



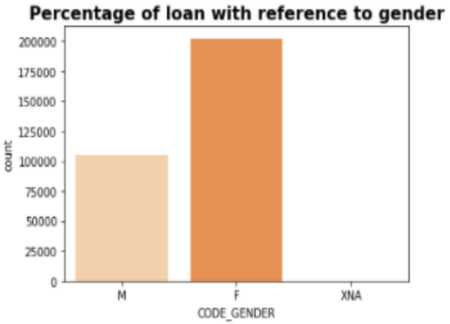
The graph depicts the column features which are Positively correlated based on target

Visual Exploratory Data Analysis

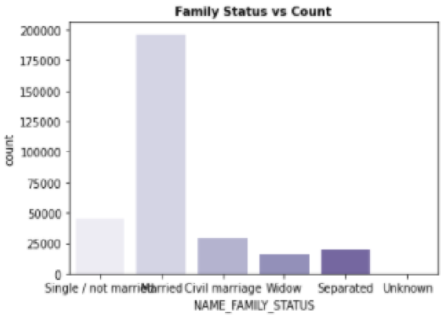


The graph depicts the column features which are Negatively correlated based on target.

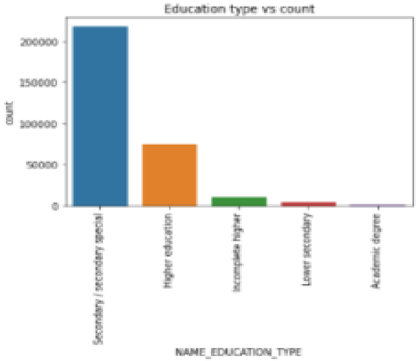
Visual Exploratory Data Analysis



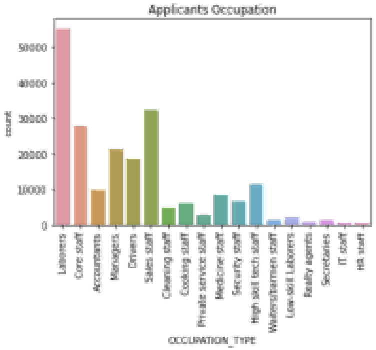
Based on Gender



Based on Family Status



Based on Education Type



Based on Occupation

Feature Engineering

1. Cat transformer was used to encode categorical features and num transformer was used to scale numerical features.
2. Feature union was used to combine these transformers and create a data pipeline for consistent preprocessing.
3. Highly correlated features were removed to reduce dataset dimensionality and improve model performance.
4. These techniques help to ensure efficient data processing and improve the performance of machine learning algorithms

Machine Learning Model

- Logistic Regression
- Lasso Regression
- Stochastic Gradient Descent Regression

Machine Learning

- Workflow for the project is discussed and distribution of responsibilities to work on HCDR project.
- Four algorithms were selected for this project including baseline logistic regression, SGD regressor, Lasso regression with full batch dataset and also with undersampled dataset
- Metrics used in the problem are **F1 Score, Confusion matrix, Recall, Precision Score, AUC**
- ✓ Machine Learning pipelines outlined for our case

Discussion and Results

Model	Basis	Test Accuracy	Test AUC	F1 Score
Baseline Logistic Regression	Full Batch Gradient Descent	91.94%	74.36%	0.0272
Logistic Regression	15 Features	91.59%	73.55%	0.0120
Logistic Regression	With Undersampling	77.21%	73.82%	0.3192
Lasso Regression	With Undersampling	75.60%	70.89%	0.0
SGD Lasso Regression	With Undersampling	66.14%	60.52%	0.4166
Logistic Regression	With Undersampling	77.92%	60.86%	0.3783

Conclusion

Our project focused on predicting the probability of default for Home Credit clients using machine learning techniques. we employed baseline machine learning pipelines which includes logistic regression, lasso regression, SGD with feature engineering, hyperparameter optimization, and undersampling. We evaluated their performance using key metrics. The best accuracy (91.94%) was achieved using baseline logistic regression with full batch gradient descent; however, its low F1 score (0.0272) suggests potential imbalanced class performance.

Future work involves further experimentation with other algorithms like SVM, KNN, GBM's like XGBoost, and neural networks, feature engineering techniques, and sampling methods, as well as incorporating domain-specific knowledge and expanding the dataset. Our project lays the foundation for Home Credit to make more informed lending decisions and promote financial inclusion for underserved populations.

4 P’s

Past	Present	Planned	Problems
Understanding and loading the Data	Performed EDA and Visual EDA	More Feature Engineering and feature selection	Merging 8 huge datasets
Running the baseline code	Feature Engineering	Hyperparameter Tuning and Ensemble methods	Implementing various models on huge dataset
Analysis of ML Algorithms and metrics	Performed aggregation of features and merged the tables	Implementing Neural Networks	Training models taking lot of time
Design basic pipelines	Ran the baseline pipeline models (Ex. Lasso, LogisticReg) and got the scores	Analysis of Loss functions And comparing the best performing models	Plotting graphs for huge datasets crashing kernel sometimes (Due to more RAM usage)

Thank you

INDIANA UNIVERSITY BLOOMINGTON