Luddy School of Informatics, Computing, and Engineering

Phase 4- 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

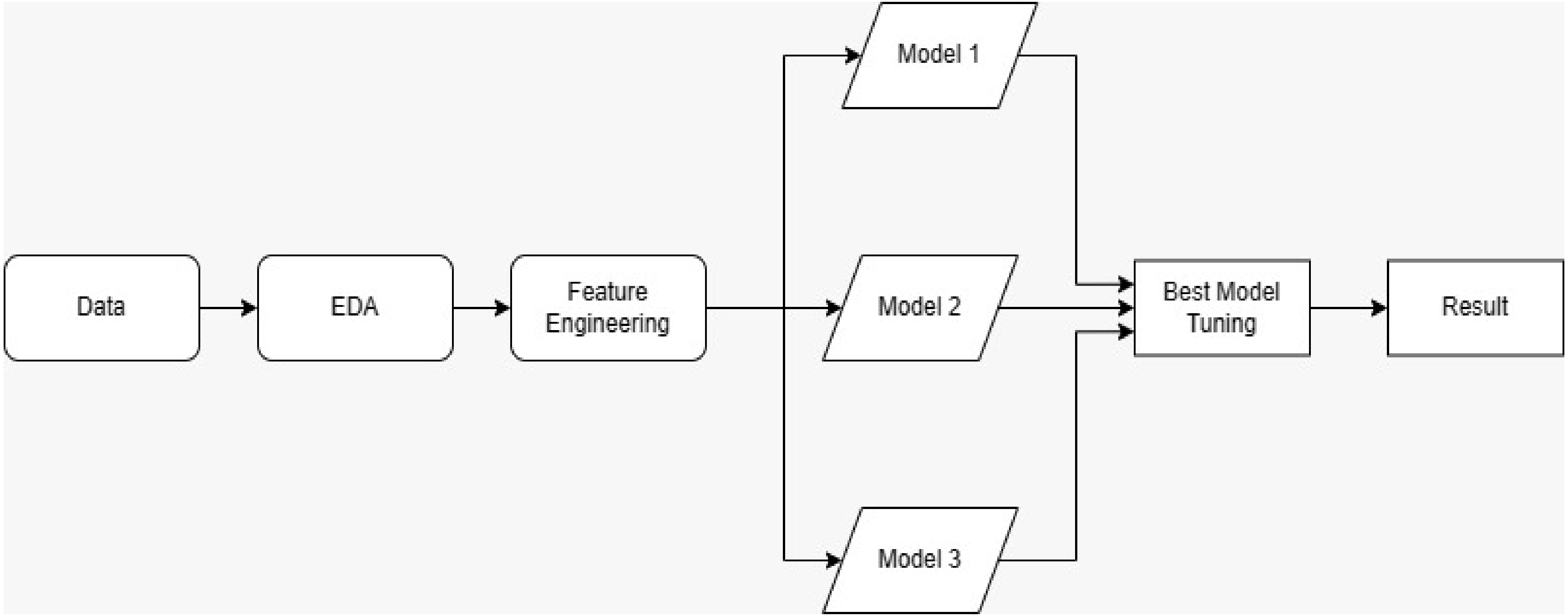
- Project Description
- EDA and Visual EDA Summary
- Overview of Pipelines Implemented
- Discussion and Results
- Conclusion
- Challenges

Project Description

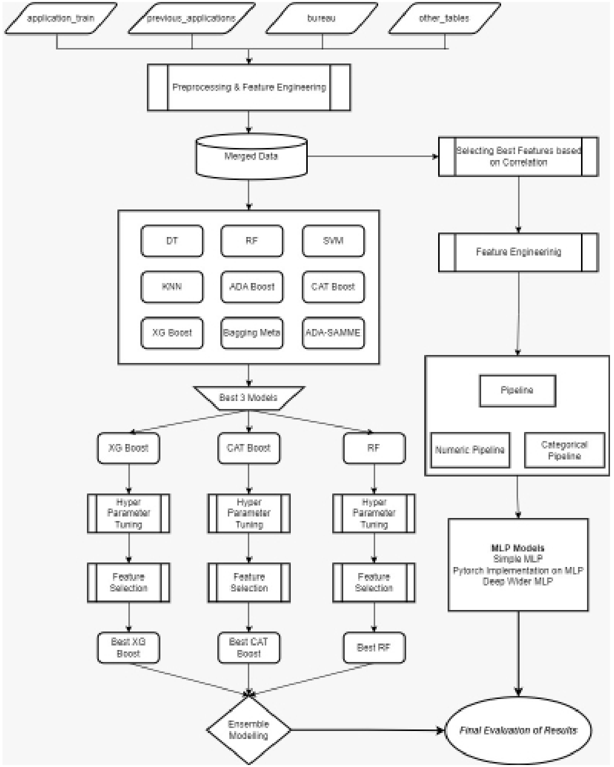
During this project phase, we obtained data from the Kaggle competition "Home Credit Risk Analysis" and conducted Exploratory Data Analysis to examine and understand the dataset. We generated various visualizations for most input features related to the "Target" variable to identify individuals at the 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, feature selection and hyper-parameter tuning on Neural Network models such Multi layer perceptron and submitted best model for Kaggle submission

PHASE 4 WORKFLOW



ENTIRE PROJECT WORKFLOW



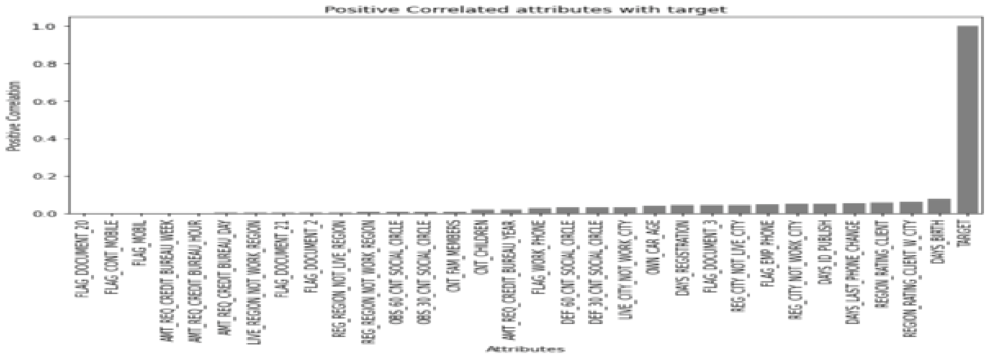
Exploratory Data Analysis.

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”.

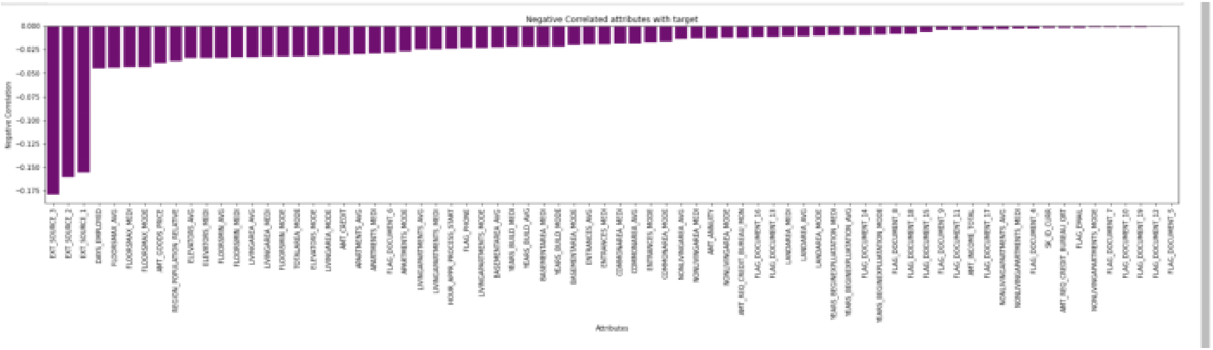
Visual Exploratory Data Analysis

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



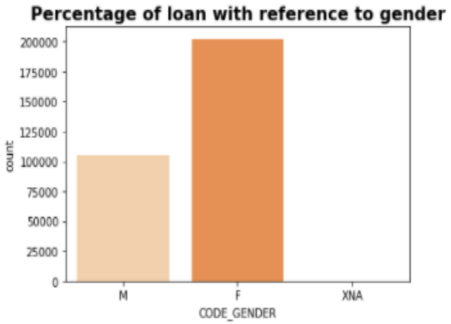
The graph depicts the column features which are Positively correlated based on target

Visual Exploratory Data Analysis

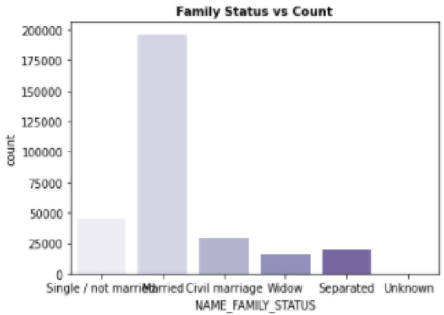


The graph depicts the column features which are Negatively correlated based on target.

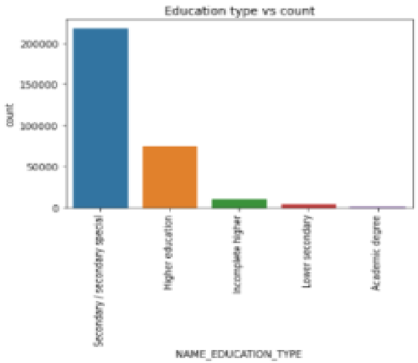
Visual Exploratory Data Analysis



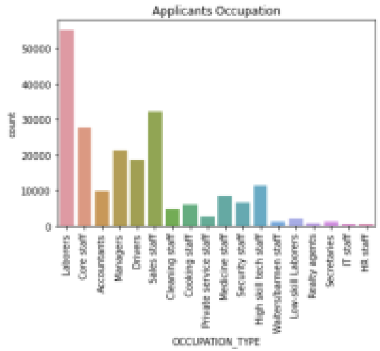
Based on Gender



Based on Family Status



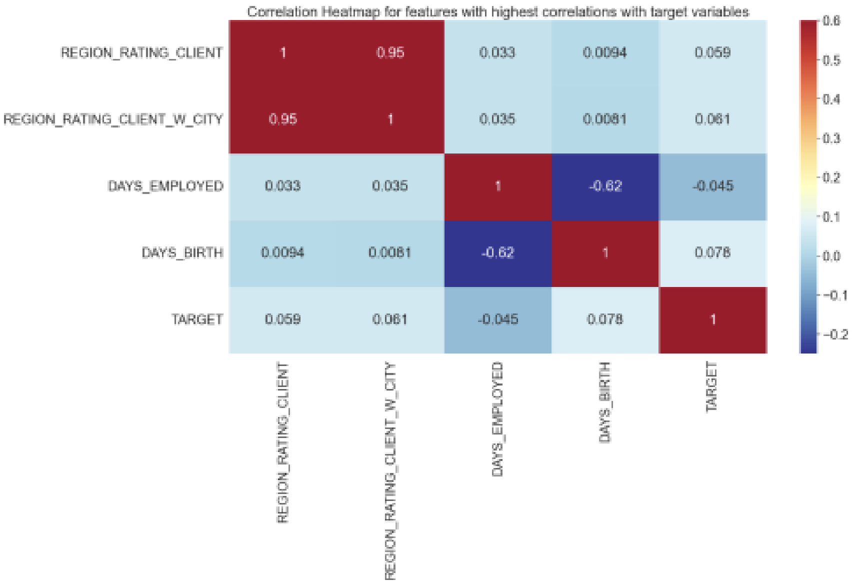
Based on Education Type



Based on Occupation

Visual Exploratory Data Analysis

```
Text(0.5, 1.0, 'Correlation Heatmap for features with highest correlations with target variables')
```



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

Neural Networks

- Multilayer perceptron

Machine Learning

- Workflow for the project is discussed and distribution of responsibilities to work on HCDR project.
- We did neural network with full features from data dict from and also selected from this filtering x>0
- Metrics used in the problem are **F1 Score, Recall, Precision Score, AUC**
- Machine Learning pipelines outlined for our case

Discussion and Results

exp_name	learning_rate	epochs	Train Time (sec)	Test Time (sec)	Train Acc	Test Acc	Train AUC	Test AUC	Train F1	Test F1
Model 1 All	0.01	1000	5.0025	3.6912	0.6909	0.6828	0.6909	0.6828	0.6903	0.6832
Model 1 All	0.01	1000	5.0025	3.6912	0.6909	0.6828	0.6909	0.6828	0.6903	0.6832
Model 1 All	0.01	1000	5.0025	3.6912	0.6909	0.6828	0.6909	0.6828	0.6903	0.6832
Model 1 selected	0.01	1000	2.297	2.2334	0.6902	0.6814	0.6902	0.6814	0.6896	0.6816
Model 2 Enhanced all	0.01	1000	27.099	28.2518	0.999	0.6346	0.999	0.6349	0.999	0.6661
Model 2 enhanced 2	0.001	50	1.4786	1.407	0.7411	0.6806	0.7411	0.6807	0.7501	0.6925
Model 2 enhanced and selected	0.001	50	1.4156	1.4354	0.7364	0.6826	0.7364	0.6826	0.7413	0.6904
Model 2 change learning rate and epochs and selected	0.0005	50	1.4849	1.4059	0.7101	0.6816	0.7101	0.6817	0.7165	0.6915
Model 3 deepwide all	0.001	50	3.6939	3.6335	0.7561	0.6805	0.7561	0.6804	0.7491	0.6738
Model 3 deepwide selected	0.001	50	3.6692	3.6169	0.7576	0.6806	0.7576	0.6807	0.7722	0.7029
Model 4 Hyper Parameter Tuning	Variable	20	Nan	Nan	0.7476	0.6761	0.7489	0.6843	0.7478	0.6772

Conclusion

In this project, we aimed to predict the probability of default for Home Credit clients using historical data. We hypothesized that machine learning models with custom features could accurately predict default risk. In Phase 4, we tested Multi-Layer Perceptron (MLP) models and found that Model 2 and Model 3 showed strong performance, with test accuracies of 0.6806 and test F1 scores of 0.6925 and 0.7029, respectively.

Our work demonstrates the importance of feature engineering and hyperparameter tuning for optimizing model performance. Future improvements can include experimenting with hyperparameters, regularization techniques, and model architectures, enhancing feature selection, increasing dataset size, and utilizing advanced ensemble methods to improve lending decisions.

Challenges

- One of the main challenges was dealing with imbalanced data, as there were far more non-default cases than default cases.
- hyperparameter tuning was a significant challenge,
- Working on colab and mac environment was challenging , need to change code to run on different hardware
- Working with deep learning needed lot of GPU resources otherwise it takes a lot cpu time even with powerful processors

Thank you

INDIANA UNIVERSITY BLOOMINGTON