Leveraging Machine Learning to Predict Loan Defaults

Phase 2: EDA and Baseline Pipelines

Group 4

Maria P. Aroca, Cassie Cagwin, Lexi Colwell, Nasheed Jafri

Project Overview

- **Purpose:** The project aims to help lending institutions improve loan approval processes by predicting an applicant's ability to repay loans using a combination of Telco and Transactional data.
- **Problem:** Many loan applications are rejected due to insufficient credit history, forcing applicants to turn to predatory lenders. This project seeks to reduce unjust rejections while minimizing the risk of defaults.
- **Goal:** Build a predictive system to identify potential defaulters using applicant data.

Project Overview: Data

- The dataset from the **Home Credit Default Risk** project comprises a rich and diverse collection of data sources:
  - **Application Data:** Demographic and financial details for current loan applications (307k rows, 122 columns).
  - **Credit Bureau Data:** 1.7M records of clients' past credits and 27M monthly balances from other institutions.
  - **Credit Card Balances:** 3.8M records of monthly balances for previous Home Credit credit cards.
  - **POS & Cash Loans:** 10M monthly snapshots of repayment histories for POS and cash loans.
  - **Installment Payments:** 13.6M records of repayment histories, including missed installments.
  - **Previous Applications:** 1.7M historical records of past loan applications with Home Credit.

Phase 2 Work: Key Tasks Completed

Data Download and EDA

We loaded and explored the data. We reviewed summary statistics and filled nulls. We made visualizations including histograms and correlation matrices.

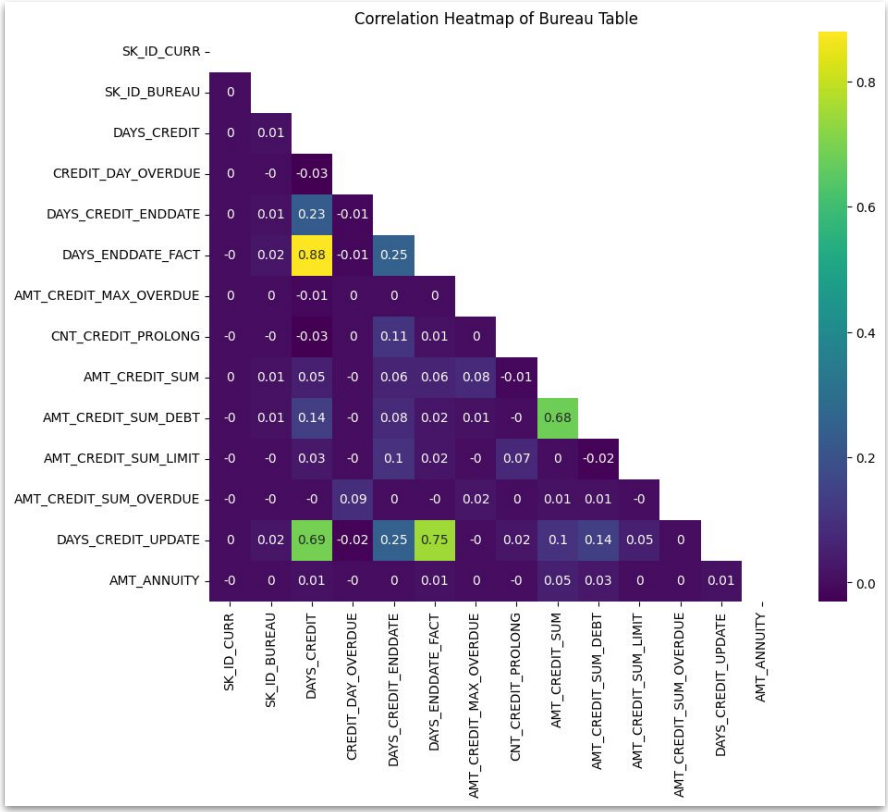
Feature Engineering

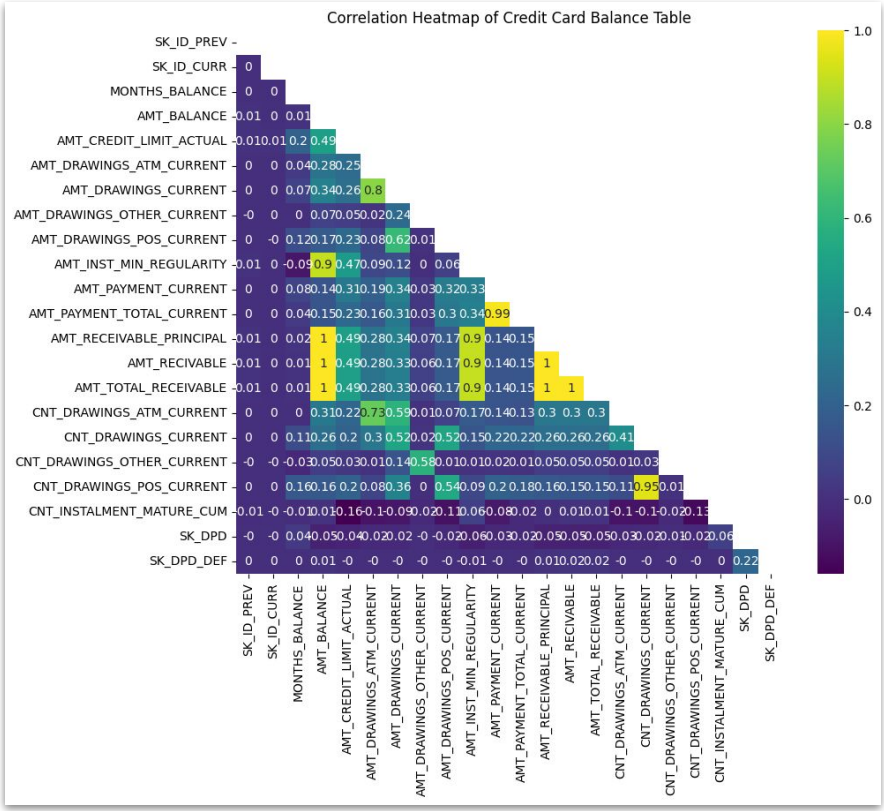
We aggregated numerical features grouped by SK_ID_CURR and then scaled them. We OneHotEncoded most categorical values. We then combined all the tables into a single dataframe.

Baseline Pipeline Dev

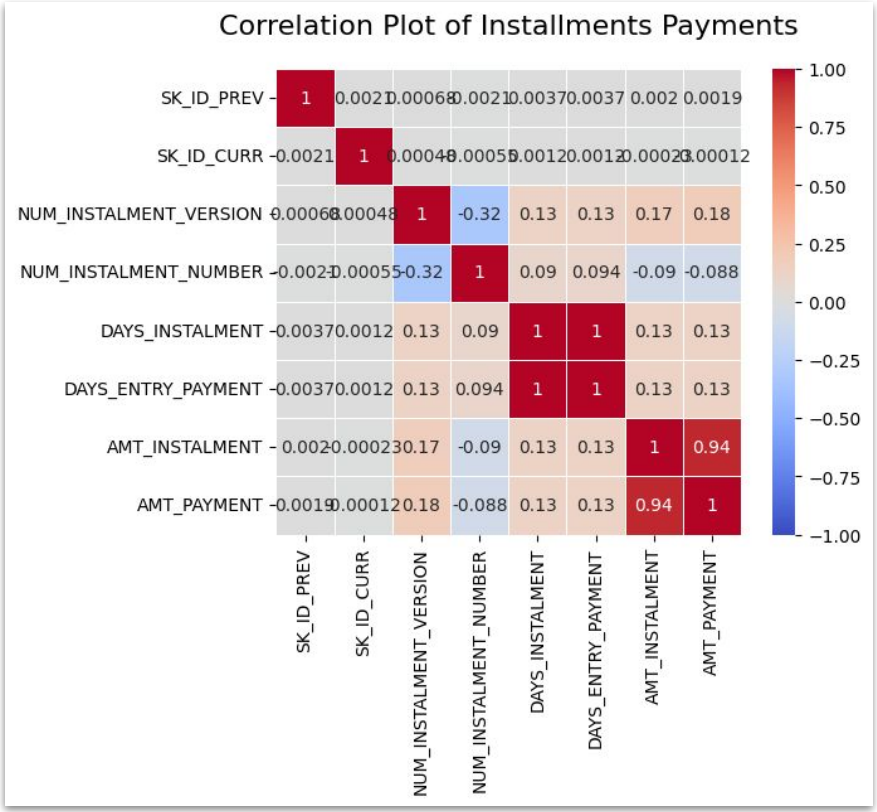
After Data Engineering, we designed implemented a Feature Engineering pipeline as well as a simple baseline pipeline that includes **Logistic Regression.**

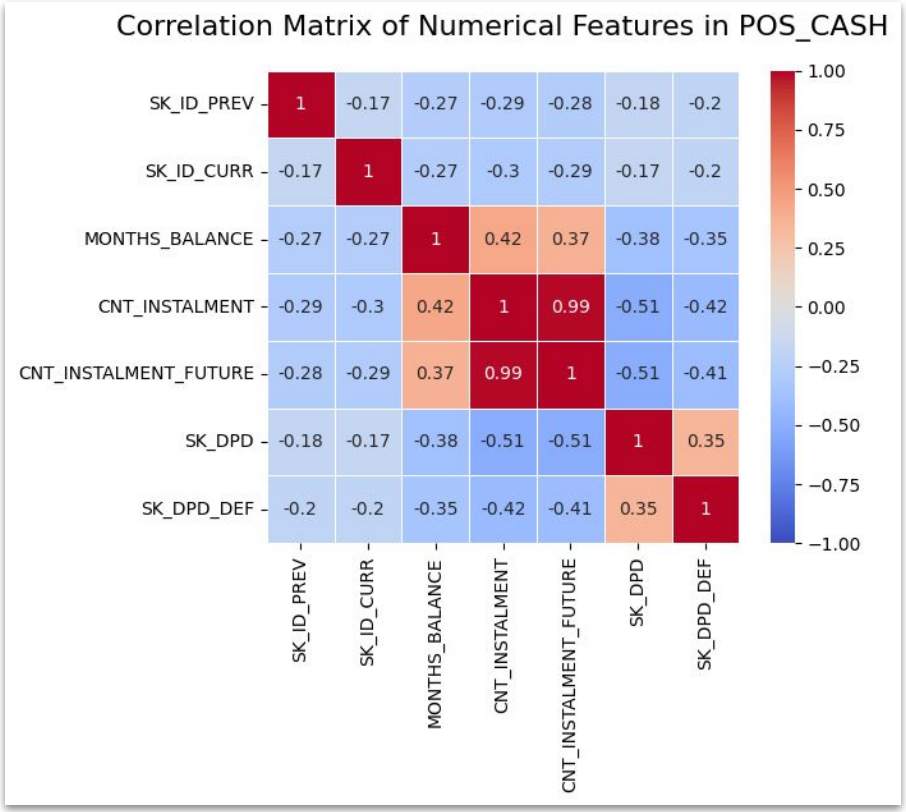
Phase 2 Work: Key insights from EDA and experiments



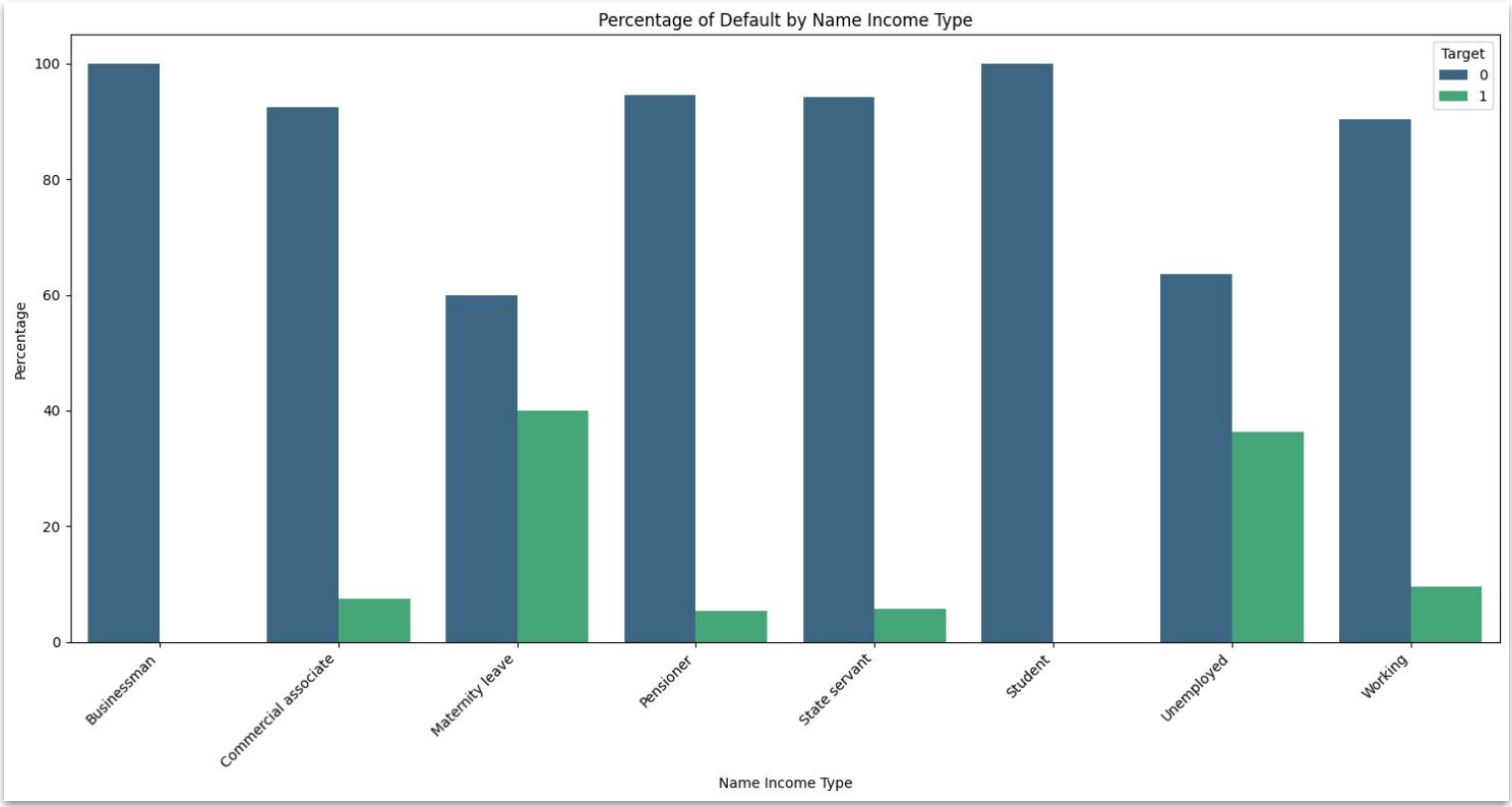


Phase 2 Work: Key insights from EDA and experiments





Phase 2 Work: Key insights from EDA and experiments



Phase 2 Work: Metrics Used

01	**Accuracy**	Proportion of correctly predicted samples out of the total samples
02	**Precision**	Proportion of true positive predictions out of all positive predictions
03	**Recall**	Proportion of actual positives correctly identified
04	**F1 Score**	Harmonic mean of precision and recall, balancing the two metrics
05	**AUC-ROC**	Area under the Receiver Operating Characteristic curve, measuring model's ability to distinguish between classes
06	**Log Loss**	Logarithmic loss penalizes wrong predictions more as they deviate from true class probabilities

Phase 2 Work: Baseline pipeline results

Metric	Baseline
Train Accuracy	0.9196
Valid Accuracy	0.9158
Test Accuracy	0.9204
Train ROC AUC	0.6134
Valid ROC AUC	0.6095
Test ROC AUC	0.6122

Metric	Baseline
Train Precision	0.1852
Valid Precision	0.1111
Test Precision	0.2857
Train Recall	0.0003
Valid Recall	0.0003
Test Recall	0.0008

Metric	Baseline
Train F1	0.0007
Valid F1	0.0006
Test F1	0.0015
Train Log Loss	0.2765
Valid Log Loss	0.2852
Test Log Loss	0.2748

Plans for Phase 3

- Goals for the next phase
  - Revise our features:
    - New features
    - Perform feature selection / dimensionality reduction
  - Develop a more robust modeling pipeline:
    - More types of models
    - Hyper parameter tuning
- Adjustments or new ideas based on Phase 2 findings
  - Upon completion of Phase 2 we recognize the need for further feature engineering
- Stretch goals
  - We would like to try using all of the data
  - We would like to further refine our metrics

Challenges Faced

- Lack of free collaboration platforms with advanced resources
  - We work locally and in Google Colab during Phase 2. We are exploring purchasing GCP VMs to run our Google Colab notebooks on to increase resources.
- Large datasets
  - We will do dimensionality reduction in Phase 3 to improve processing time with little negative impact on model quality.
- Computational limitations
  - We had to reduce some visualizations to stay within our current processing limitations. In Phase 3 we will explore increased processing resources in GCP.
- Merging datasets
  - We were able merge select data from each table, often aggregating the data. We will attempt to include any remaining relevant data in Phase 3.
- Similar features in multiple secondary tables
  - We will check correlations between all secondary table fields to make sure they are not redundant with other features.

Four-P Summary Slide

- **PAST:** The project aims to help lending institutions improve loan approval processes by predicting an applicant's ability to repay loans using a combination of Telco and Transactional data.
- **PRESENT:** EDA on all datasets, Feature Engineering, Baseline pipeline development, Results: basic model had decent accuracy but additional model refinement is needed.
- **PLANNED:** Additional feature engineering, more models (including ensemble), hyper parameter tuning.
- **PROBLEMS:** Large dataset size causing computers and Google Colab to crash, common features in secondary tables showing redundant data, difficulty finding a useful collaboration platform.