Project Machine Learning Fundamentals

Members: Anna Abrahamyan, Zahir Ahmad, Grisel Quispe
Dataset: Credit Card Approvals
https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data

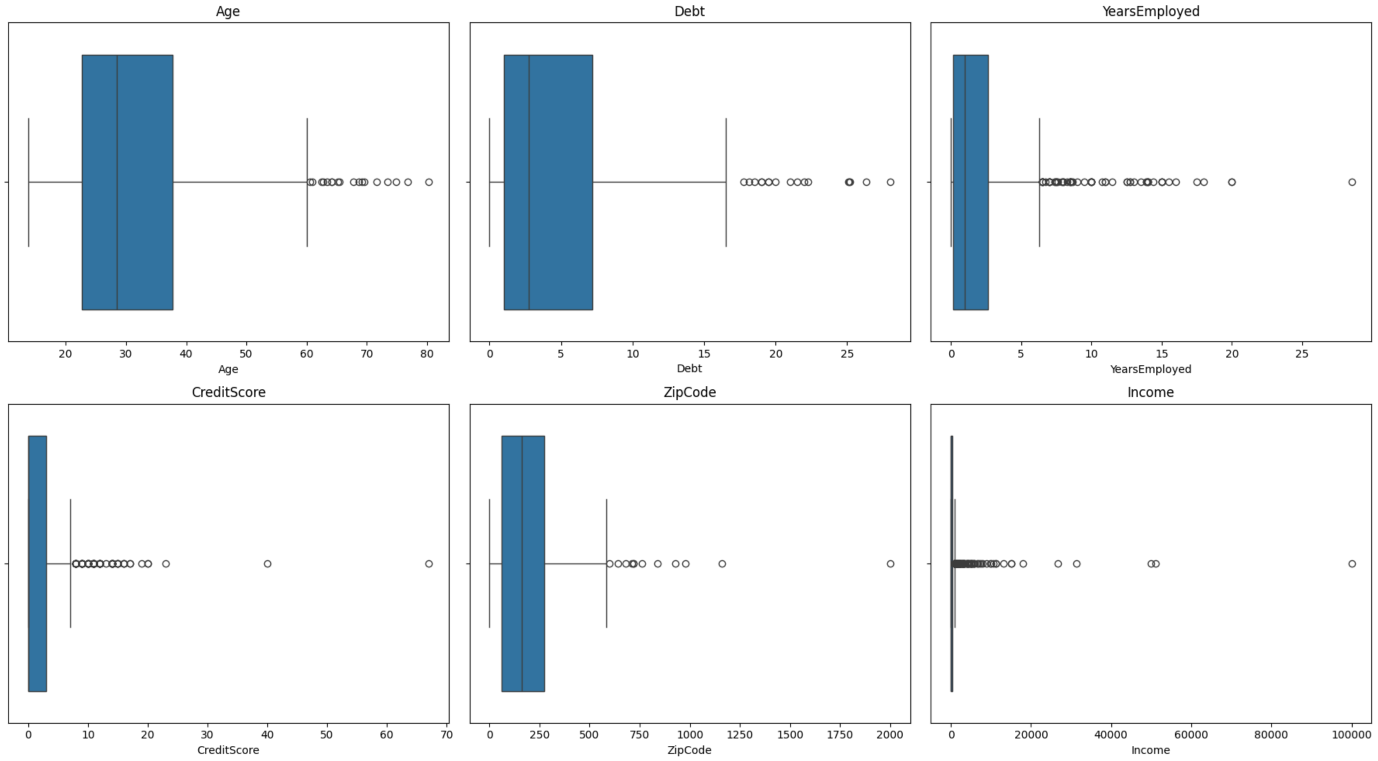
This dataset is sourced from the public domain and pertains to the business domain with 15 features and one target variable. These features enable us to predict whether a credit card application will be approved. This dataset is of particular interest to us as machine learning practitioners, as we can leverage our expertise to implement a Support Vector Machine Model and observe its performance with real-world data in a business context.

Dataset: Comprises 690 entries with 16 columns in a multidimensional feature space with numerical and categorical variables. Also presents:

- Imbalance class distribution
- Presence of outliers
- High-dimensionality feature space.
- Noisy features.
- Support a classification task (with categorical labels)
- Have at least 3 features

Features: Gender, Age, Debt, Married, BankCustomer, Industry, Ethnicity, YearsEmployed, PriorDefault, Employed, CreditScore, DriversLicense, Citizen, ZipCode, Income, Approved.
Target Variable: Approved, is binary, indicating that this dataset supports a classification task.

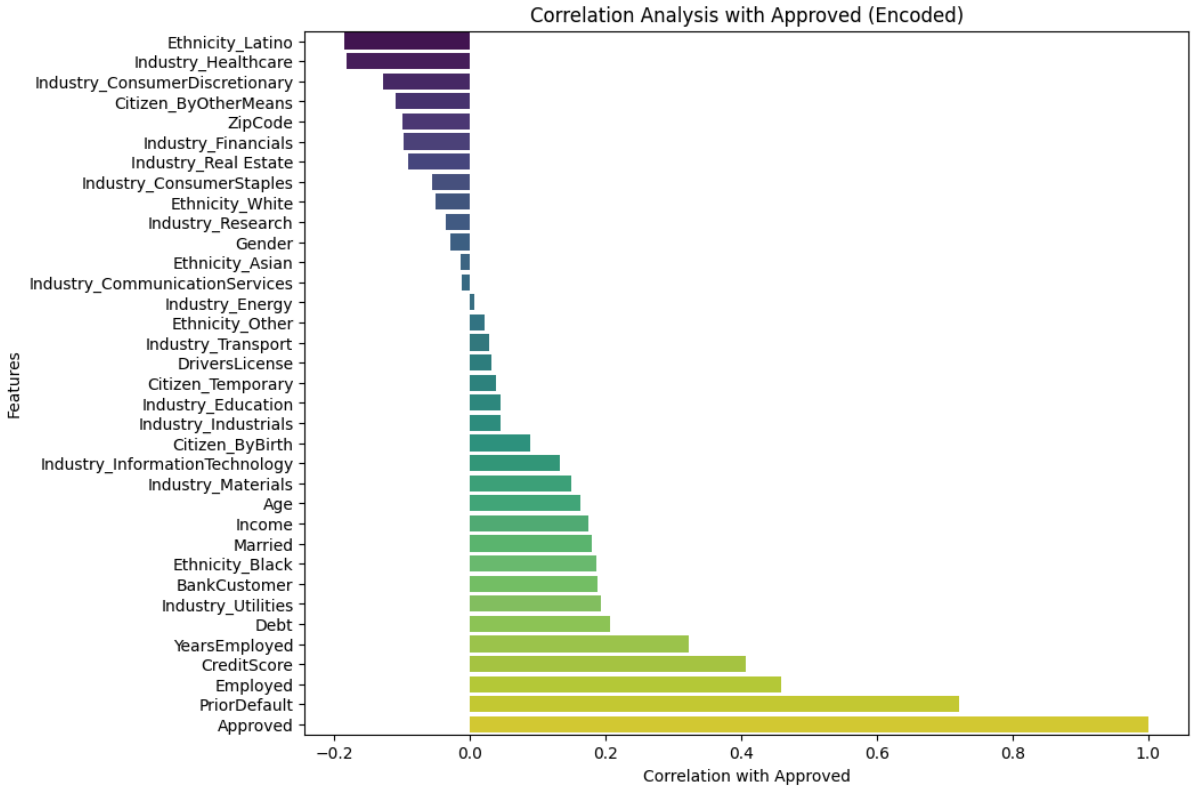
Outliers for numerical columns



Presence of Outliers: The box plots for numerical columns indicate the presence of outliers in several features:

-Age, Debt, YearsEmployed, CreditScore, and Income have outliers, suggesting that some entries have unusually high values compared to the rest of the data. -ZipCode also shows a wide distribution, but as a categorical feature represented numerically, its "outliers" might actually represent less common categories.

Correlation Analysis



Feature Correlation Analysis reveals the varying degrees of association between features and the target variable, Approved. Remarkably, PriorDefault emerges as the feature with the strongest positive correlation to approval status, underscoring its potential significance in classification tasks. Following closely are Employed, CreditScore, and YearsEmployed, exhibiting notable positive correlations. Conversely, features such as ZipCode, Gender, and DriversLicense exhibit minimal correlation with the target variable, implying their limited predictive utility.

We found this dataset interesting and useful for our project to implement SVM.