Credit Score Analysis

Arpan Banerjee
Saurabh Kankekar
Suyash Mohta

Business Case

Predictive Model Development

- Forecasts credit scores using historical financial data
- Helps stakeholders predict customer creditworthiness.

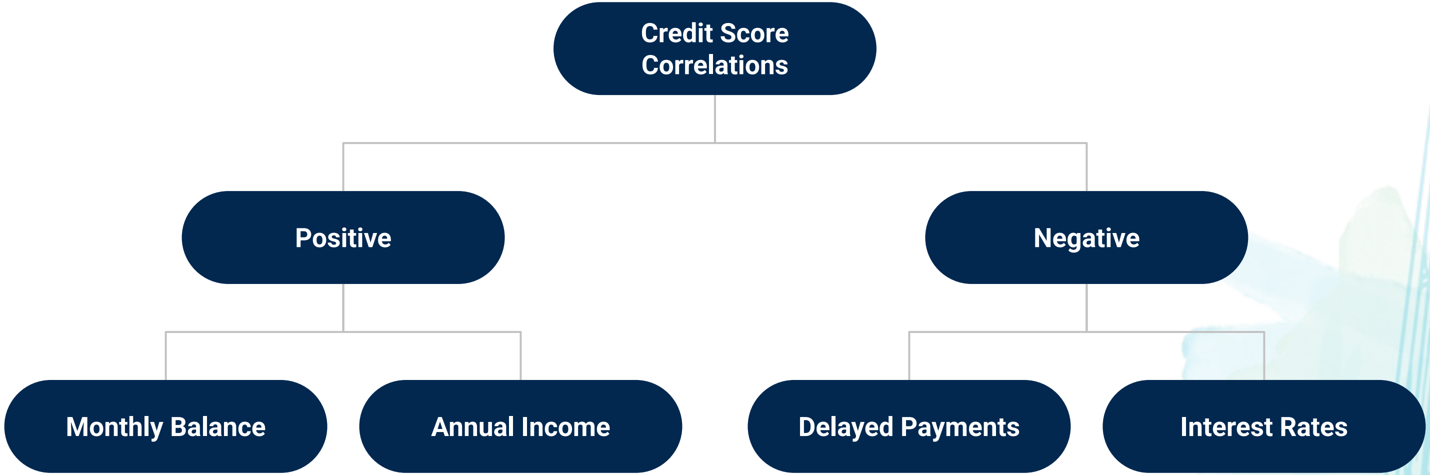
Categorization of Credit Scores

- Segments into Bad, Average, and Good
- Simplifies risk evaluation for lenders

Value Proposition for Stakeholders

- Supports data-driven decisions to mitigate risk
- Empowers lenders with actionable insights.

Key Findings from EDA

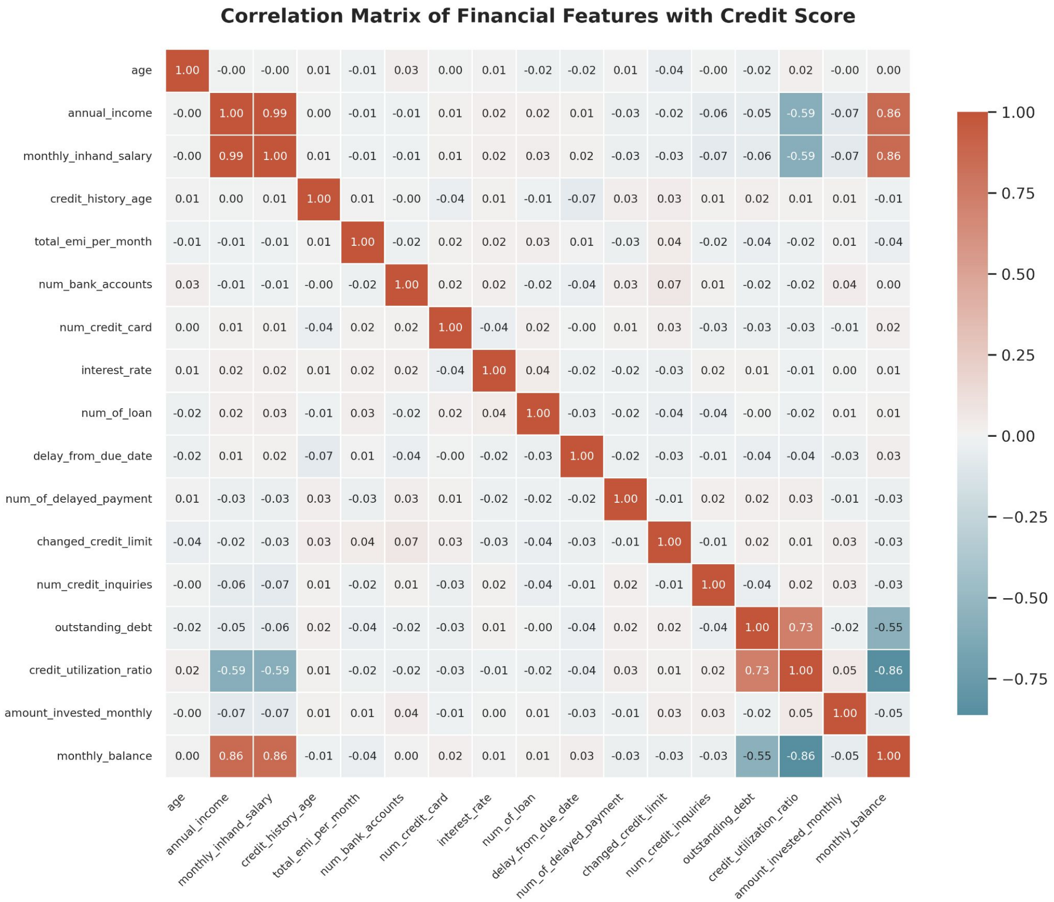


Data Cleaning and Transforming

- Imputation of missing values
- 8 rows of data per customer aggregated into single rows
- Used Average for numerical columns
- Used Mode for categorical columns

Correlation Matrix

- monthly_inhand_salary (correlation = 0.86 with annual_income)
- amount_invested_monthly (correlation = 0.03 with credit_score)



Metrics

- **Weighted Cohen's Kappa**

  Measures inter-rater agreement while accounting for the degree of disagreement using weights.

- **F1 score**

  A balance between precision and recall, useful for evaluating classification models, especially on imbalanced data.

Cohen's Weighted Kappa (κ_W):

$$\kappa_W = 1 - \frac{\sum w_{ij} \cdot fo_{ij}}{\sum w_{ij} \cdot fe_{ij}}$$

F1-Score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Model Selection & Evaluation

Preprocessing → **Fitting Models** → **Selecting Best Model** → **Model Inference**

Preprocessing
- Dropping irrelevant columns
- One-hot encoding the categorical
- Scaling the numerical features

Fitting Models
- Hyperparameter Tuning
- Cross Validation

Selecting Best Model
- Based on overall balance of F1-score & Weighted Kappa Score

Model Inference
- Permutation importance
- PDP & ICE Plots

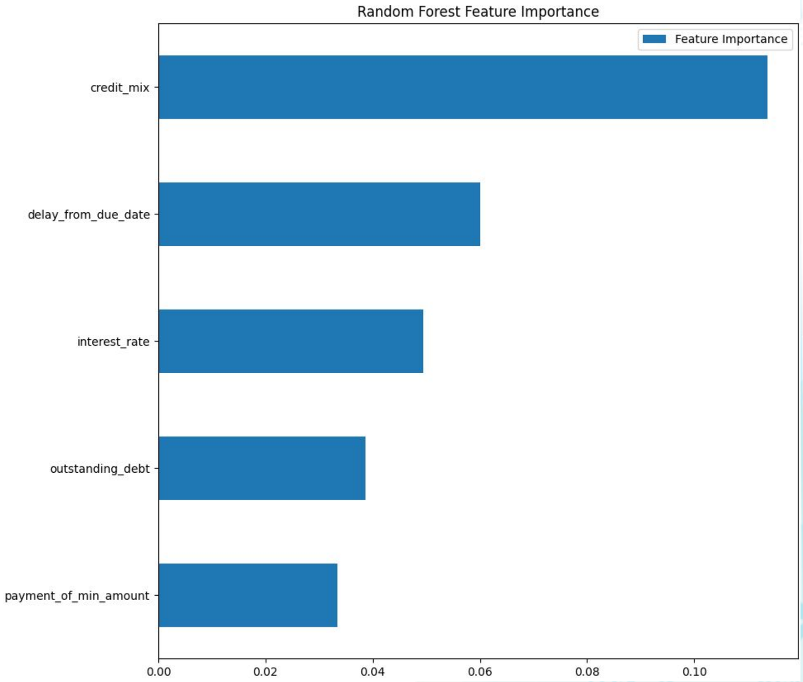
	F1 Score	Weighted Kappa
Logistic Regression	0.709565	0.597992
Random Forest	0.767050	0.683618
XGBoost	0.758447	0.670593
CatBoost	0.762795	0.661908

Model Inference

Permutation Importance

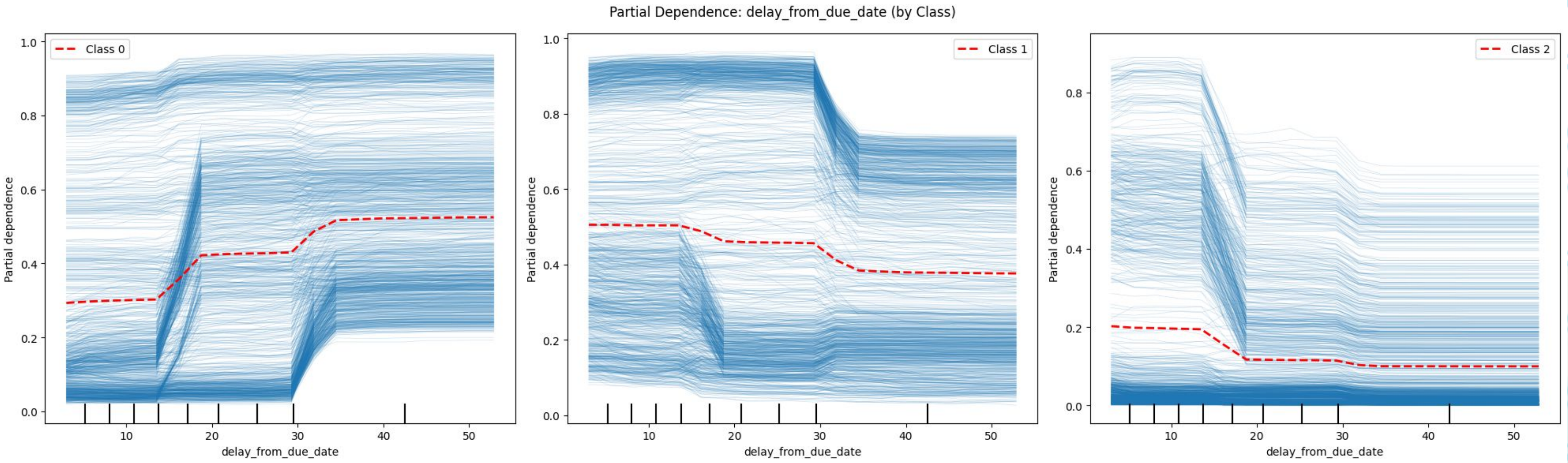
Feature's impact on model performance by randomly shuffling its values and observing the performance drop

- Credit_mix (Bad, Standard, Good)
- Delay_from_due_date
- Interest_rate
- Outstanding_debt
- Payment_of_min_amount (Yes, No)



Model Inference

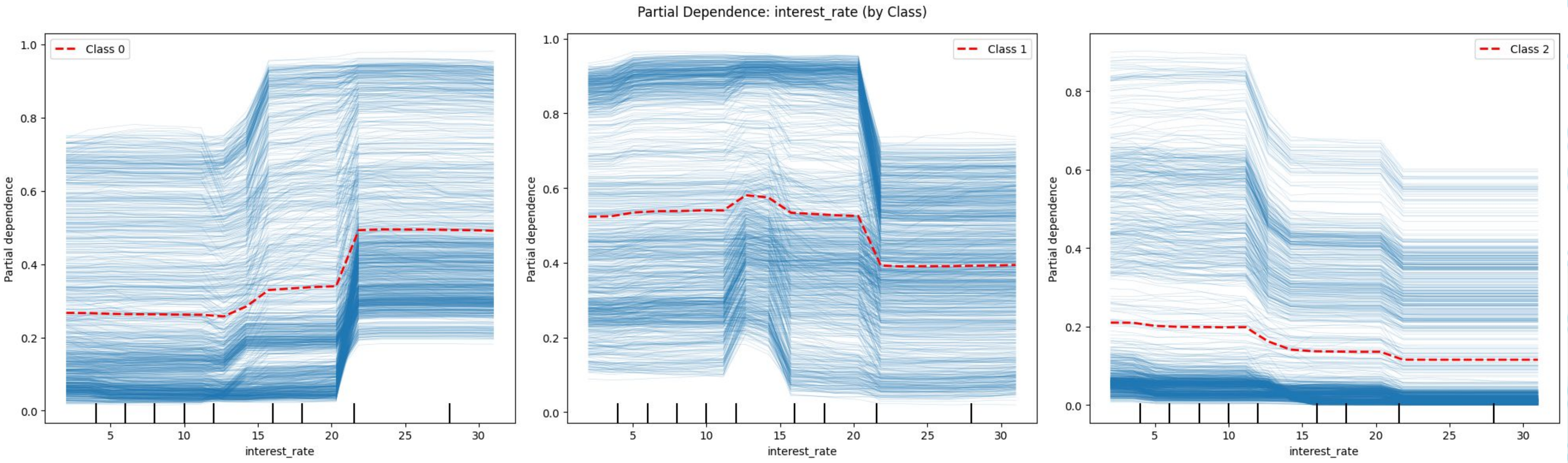
- Partial Dependence Plots
- Individual Conditional Expectation Plots



More the delay, more probability of having lower credit score

Model Inference

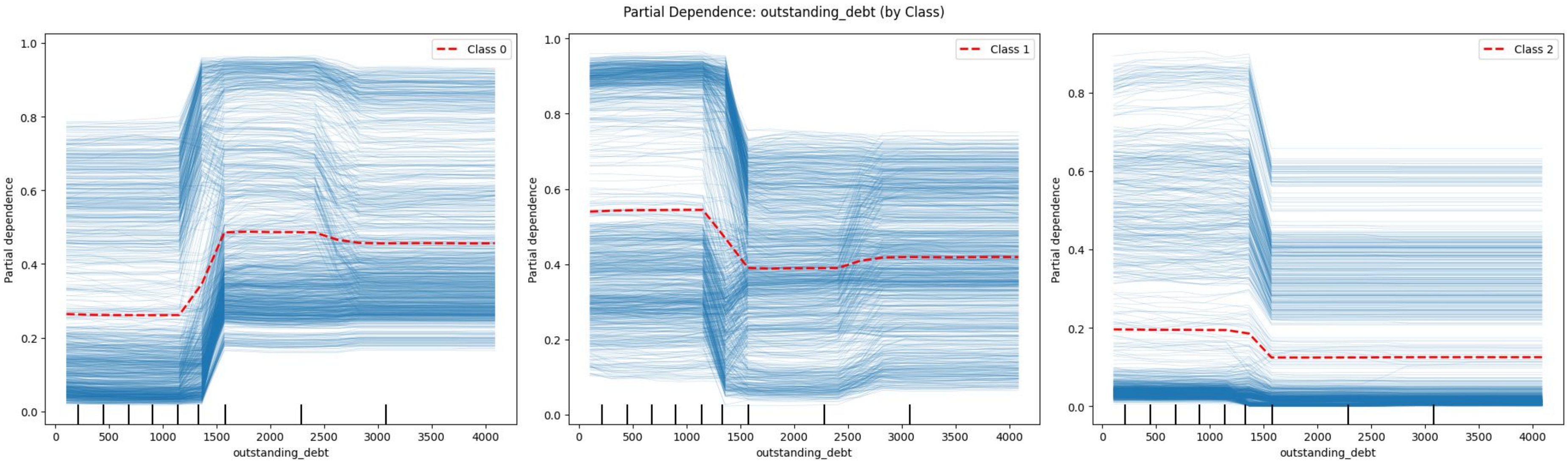
- Partial Dependence Plots
- Individual Conditional Expectation Plots



Lower the interest rate, more probability of having high credit score

Model Inference

- Partial Dependence Plots
- Individual Conditional Expectation Plots

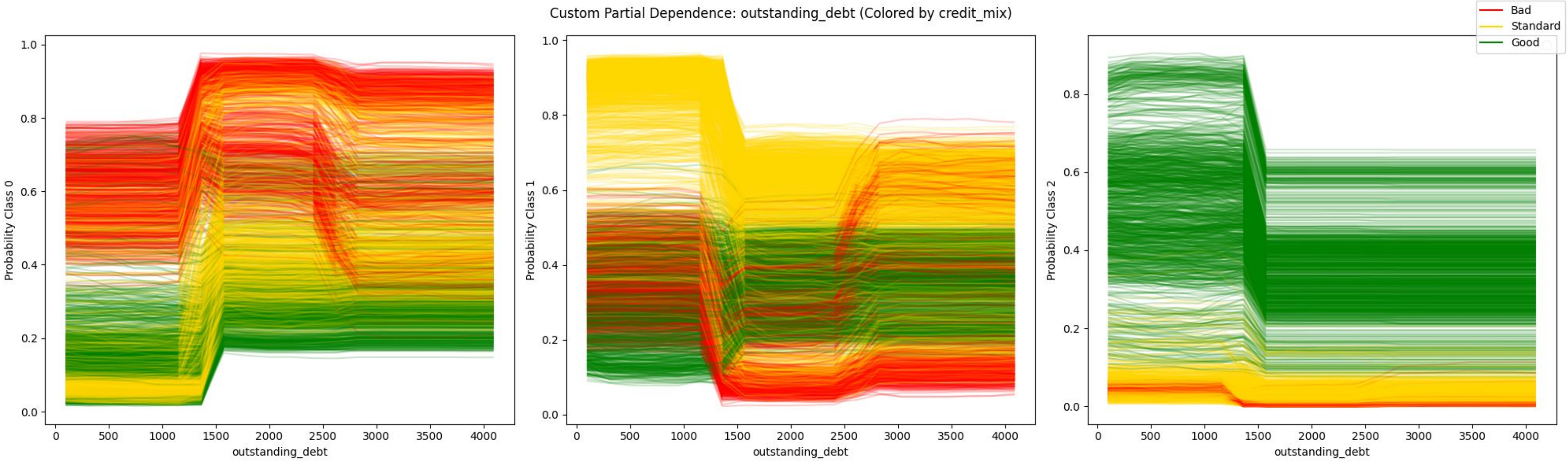


More the outstanding debt, more probability of having lower credit score

UCDAVIS

Model Inference

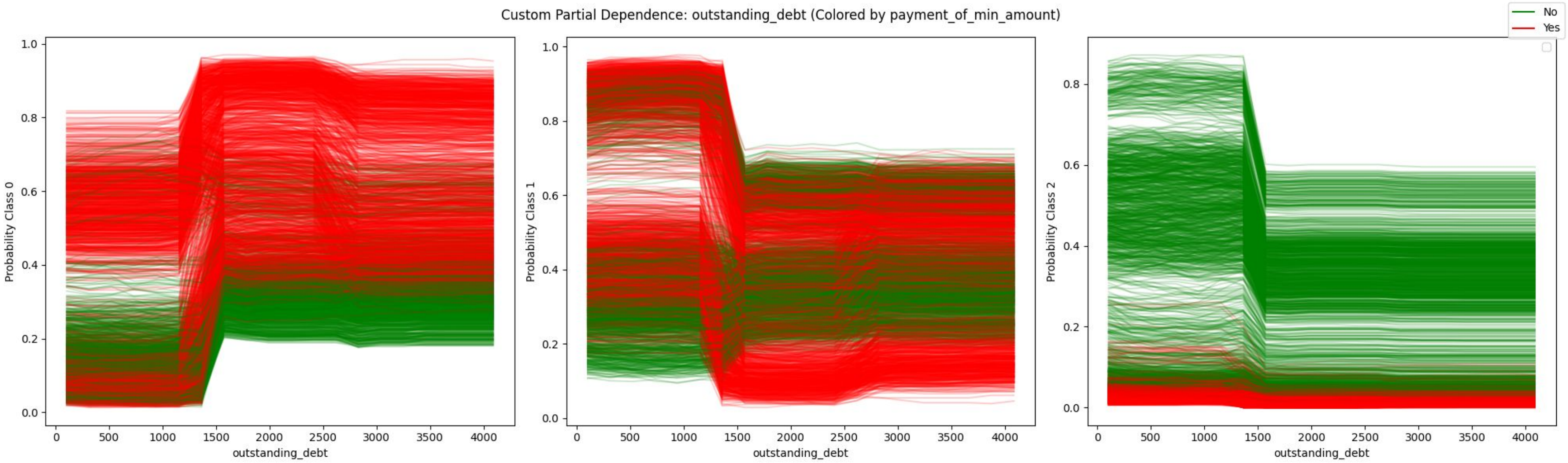
- Custom Individual Conditional Expectation Plots



Having a Good Credit Mix is critical, however large outstanding debt weighs down probability of good credit score

Model Inference

- Custom Individual Conditional Expectation Plots



Paying more than min_amount_due is important, however large outstanding debt weighs down probability of good credit score

Limitations & Future Work

- **Causal Inference:** Exploring causal inference techniques to understand the direct drivers of creditworthiness and improve model interpretations.

- **Model Generalizability:** Testing & Training the model on diverse data to ensure it generalizes well to new populations and situations.

- **Ethical Considerations:** As the model influences credit decisions, it's important to address potential biases and ensure fairness