

Comprehensive Analysis Of Debtor Payment Behavior and Default Risk Prediction For Fines Victoria



By

Jay Sangani


Prabhath Ummiti

Master of Business Analytics, Monash University



Meet Our Team DAaS


 **John Gehman**
Managing Technical Specialist.


 **Hana Basyoni**
Data Scientist

 **Cameron Bolton**
Data Analyst

 **Lanny Tieu**
Project Officer

 **James Farnell**
Data Scientist

 **Fiona Lu**
Data Analyst

 **Allanagh O'Donnell**
Manager, Priority Projects

 **DAaS and DGS Teams**

Tools and Workflow



Jira

Task management platform.



GitLab

Version control and collaboration.



Azure

Database management.



SharePoint

Document sharing and storage.



Postico

Data access and collaboration.



Visual Studio Code

Integrated development environment
(IDE).



Python

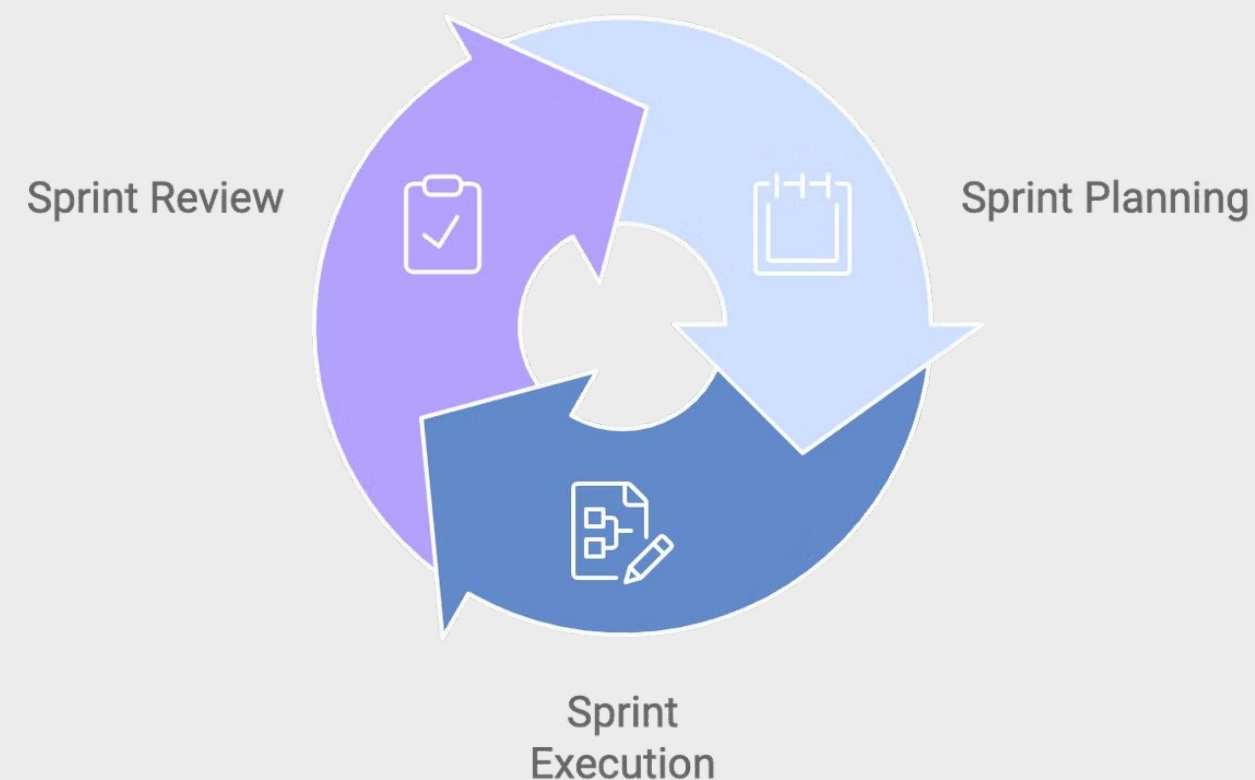
Programming language for data analysis.



SQL

Query language for database
management.

Agile Framework



Project Overview

DAaS under DGS

Works on consulting projects for government agencies, delivering cost-effective solutions that streamline processes and achieve results at 50% of the cost of outsourcing to private companies.

Fines Victoria

Oversees the management and collection of fines across the state, ensuring compliance, managing payment arrangements, and striving to enhance the system for timely collections and financial stability.

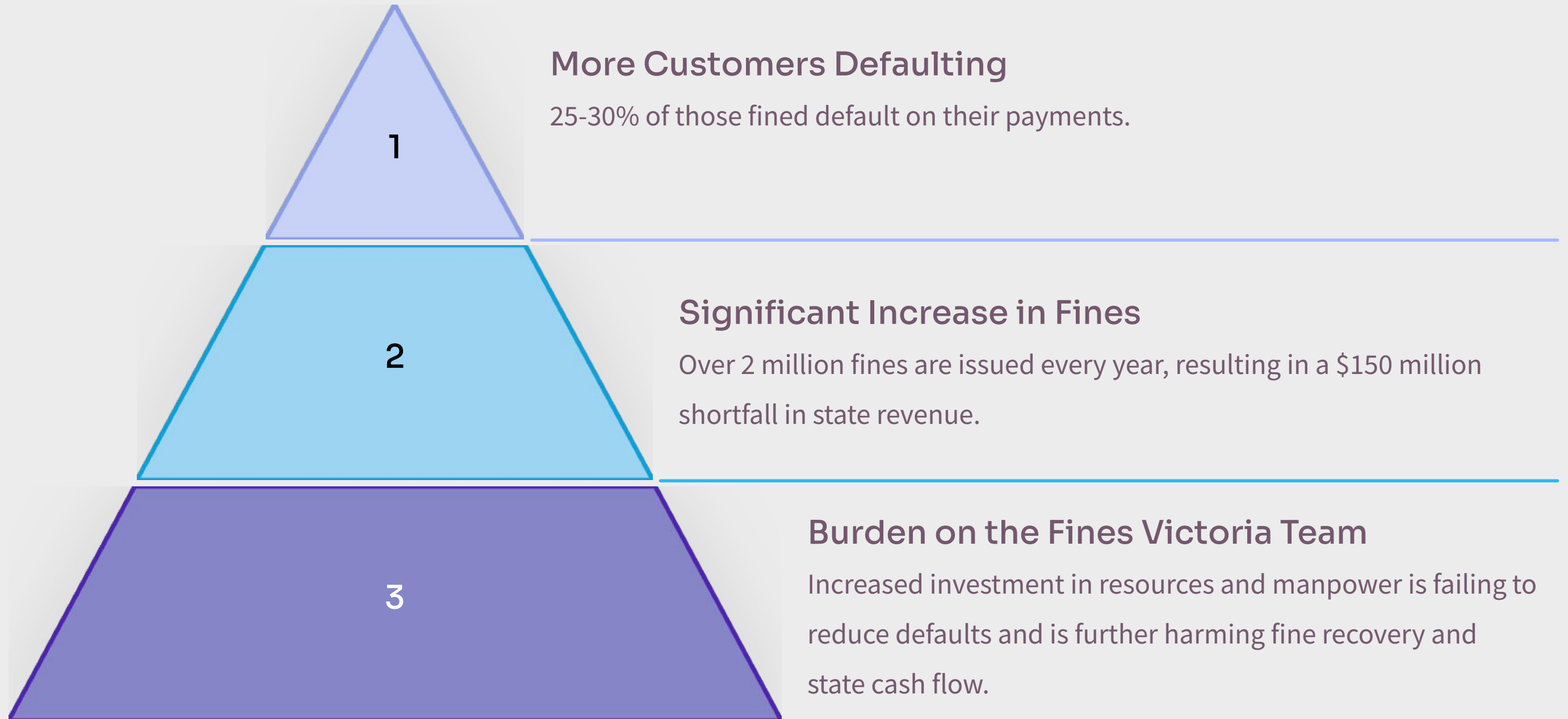
The Challenge

Fines Victoria faces growing challenges with an increasing number of debtors defaulting on payments, which causes financial strain. Without predictive insights, they struggle to take proactive action.

The Project

Aims to develop a predictive model for debtor defaults using historical data, equipping Fines Victoria with actionable insights to better manage and reduce the risk of defaults.

The Problem



The Solution



Streamlining Processes

- Restructure Fines Victoria's processes to optimize efficiency and minimize delays.
- Ensure a smoother and more effective debt collection system.

Boosting Collections

- Predict high-risk debtors to focus resources on targeted interventions.
- Lead to increased debt collection and a more sustainable financial outlook.

Optimizing Resources

- Manage a larger volume of debtors with reduced staffing requirements.
- Free up resources for other critical initiatives.

Financial Stability

- Help the Victorian government maintain a more stable financial position.
- Reduce the need to offset Fines Victoria's financial burdens.

Our Methodology

1

Planning

Identify key variables and establish the model approach.

2

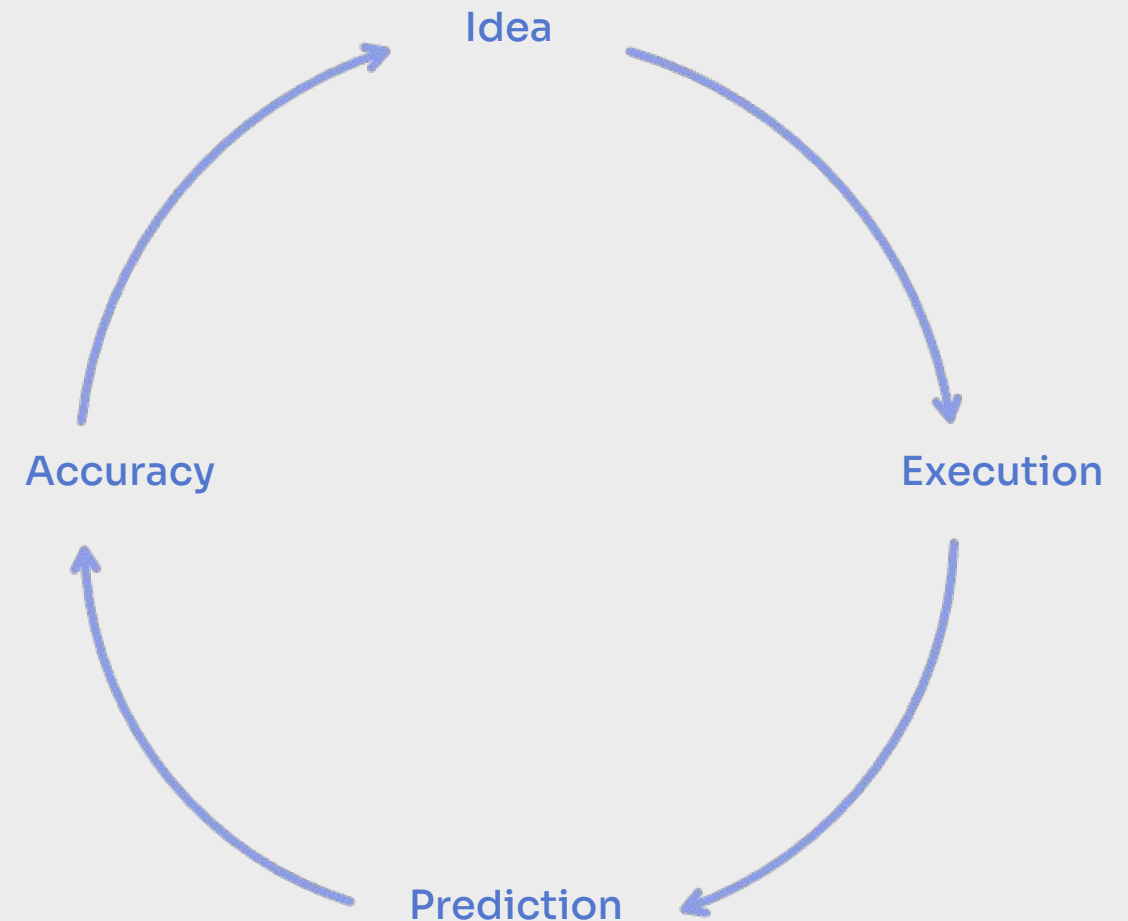
Implementation

Develop and execute the models, processing relevant data.

3

Optimization

Refine and enhance models based on performance evaluation.



About The Data

1

Data Collection

- **Source:** Datasets provided by Fines Victoria.
- **Hashing:** Dataset was hashed to ensure debtor privacy and confidentiality

2

Data Storage and Security

Storage: Securely stored in Azure.

Access: Analyzed using **Postico**.

Security: Ensured by not pushing the dataset to GitLab.

3

Dataset Overview

Rows: 4.3 million entries.

Columns: 52 columns.

Data Types: Numeric, categorical, and datetime columns.

Initial Data Analysis (IDA)

Missing Values

Identified and handled missing values across relevant columns.

Data Types

Verified correct data types (e.g., dates, numeric).

Outliers

Detected and flagged potential outliers in payment-related variables.

Basic Statistics

Summarized key metrics such as mean, median, and distribution.

Data Cleaning

Addressed inconsistencies like invalid date ranges or negative balances.

What made us decide on fitting the model?

Stakeholder Discussion: In a meeting with stakeholders at **Fines Victoria**, the key challenge identified was the difficulty in predicting debtor defaults, which was impacting their collection process.

Problem Identification: The team realized that understanding and predicting **debtor behavior** could help address these issues.

Decision to Model: We decided to train a predictive model based on **debtors' past behavior**, aiming to help **Fines Victoria** improve their strategies for managing debts and reducing the number of defaulters.

What are we trying to predict with the model?

Primary Objective:

Predict whether a debtor will default based on past payment behavior.

Focus on PA_Status:

The model targets the "Cancelled - Defaulted" status in the PA_Status variable to identify patterns that lead to defaults.

Outcome:

Accurate prediction of defaulters will enable Fines Victoria to take early intervention measures, improving collection processes and reducing financial losses.

First model

First Model Setup: The initial model was trained with **52 variables** without much refinement or evaluation of data quality.

Overfitting Issue:

Training Set: Perfect classification with an accuracy of **100%**.

Test Set: Nearly perfect accuracy of **99.9%**.

Such results typically indicate **overfitting**, where the model has memorized the training data but lacks generalization ability.

Key Insight: While both precision and recall are perfect, this model is too tightly fitted to the specific training data, making it unreliable for unseen cases in real-world applications.

Next Step: We recognized the need for **feature selection, engineering**, and other data improvements to prevent overfitting and achieve better generalization.

```
Test Set Confusion Matrix:
[[4009  0]
 [  4 987]]

Test Set Classification Report:
              precision    recall  f1-score   support

      0       1.00      1.00      1.00     4009
      1       1.00      1.00      1.00      991

   accuracy       1.00
  macro avg       1.00
weighted avg       1.00

Test Set Accuracy Score:
0.9992

Training Set Confusion Matrix:
[[16180  0]
 [  0 3820]]

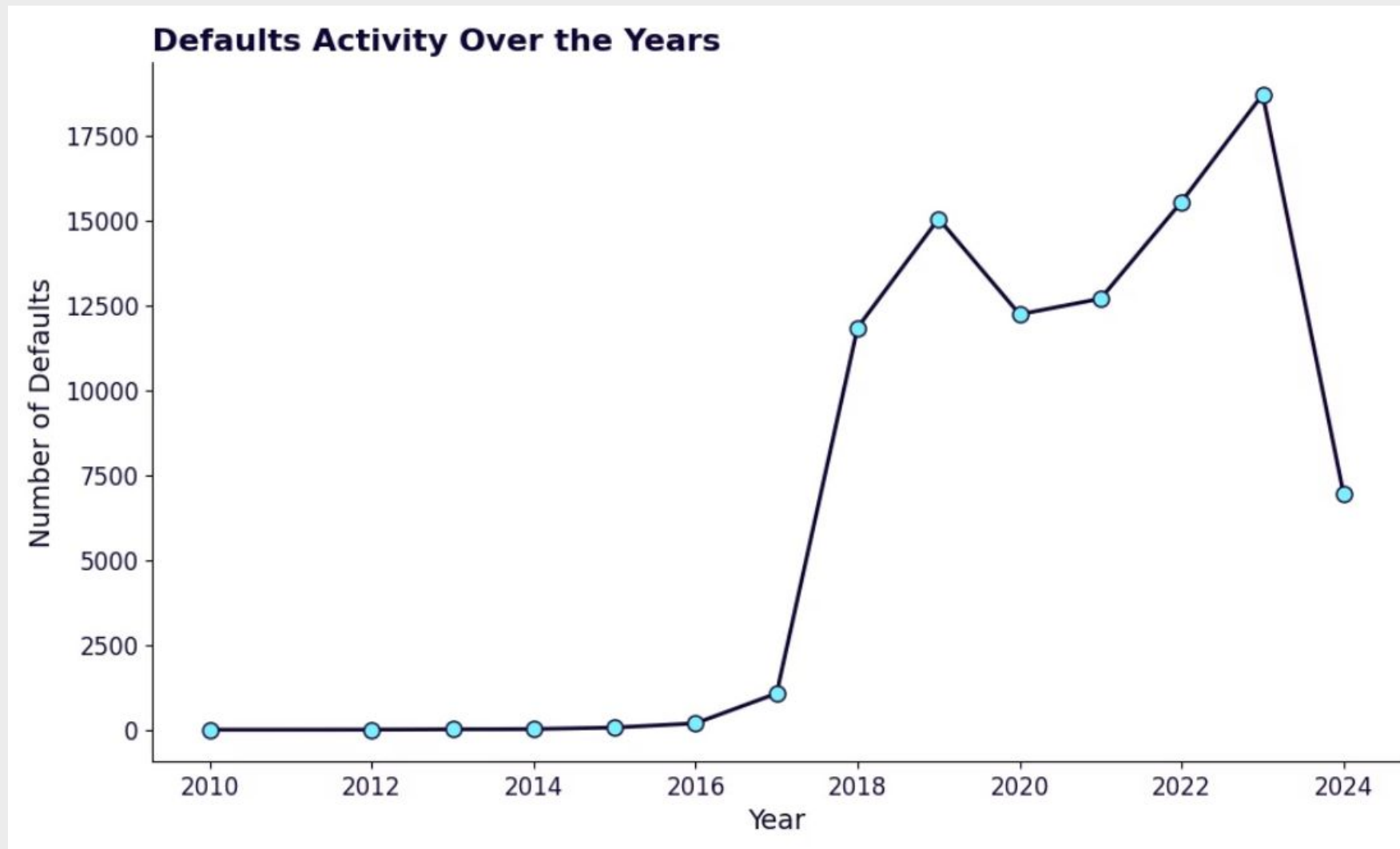
Training Set Classification Report:
              precision    recall  f1-score   support

      0       1.00      1.00      1.00    16180
      1       1.00      1.00      1.00     3820

   accuracy       1.00
  macro avg       1.00
weighted avg       1.00

Training Set Accuracy Score:
1.0
```

Time Series Plot



Observation:

A sharp rise in defaults starts in 2018, peaking in 2023 before a decline.

Insight:

The spike from 2018 onward indicates a critical period for analyzing defaults.

Decision:

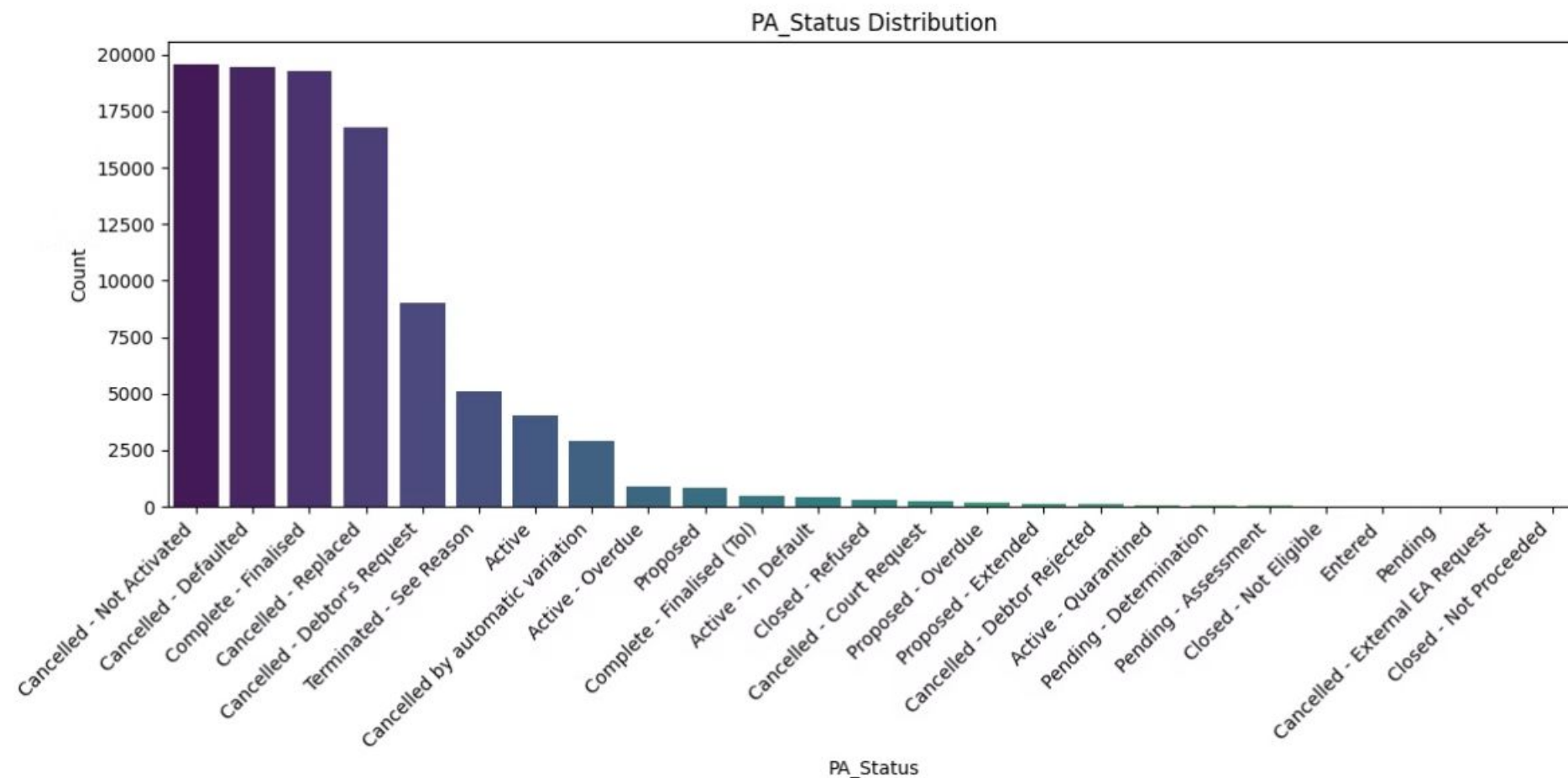
We trimmed the dataset from 2018 to focus on significant default activity for a more precise analysis.

EDA: Feature Reduction

Introduction to PA_Status: The payment arrangement status (PA_Status) is a critical variable in determining whether debtors have defaulted. The distribution of this status gives insights into the overall dataset and the proportion of defaulters versus non-defaulters.

Insights:

- The majority of debtors in the dataset have non-default statuses, with "Cancelled - Defaulted" representing a smaller portion.
- This imbalance highlights the importance of ensuring correct default predictions.



EDA: Correlation Matrix of Key Features

Content:

Correlation Insights: Exploring correlations between key numerical features helps identify relationships between variables that might influence default predictions.

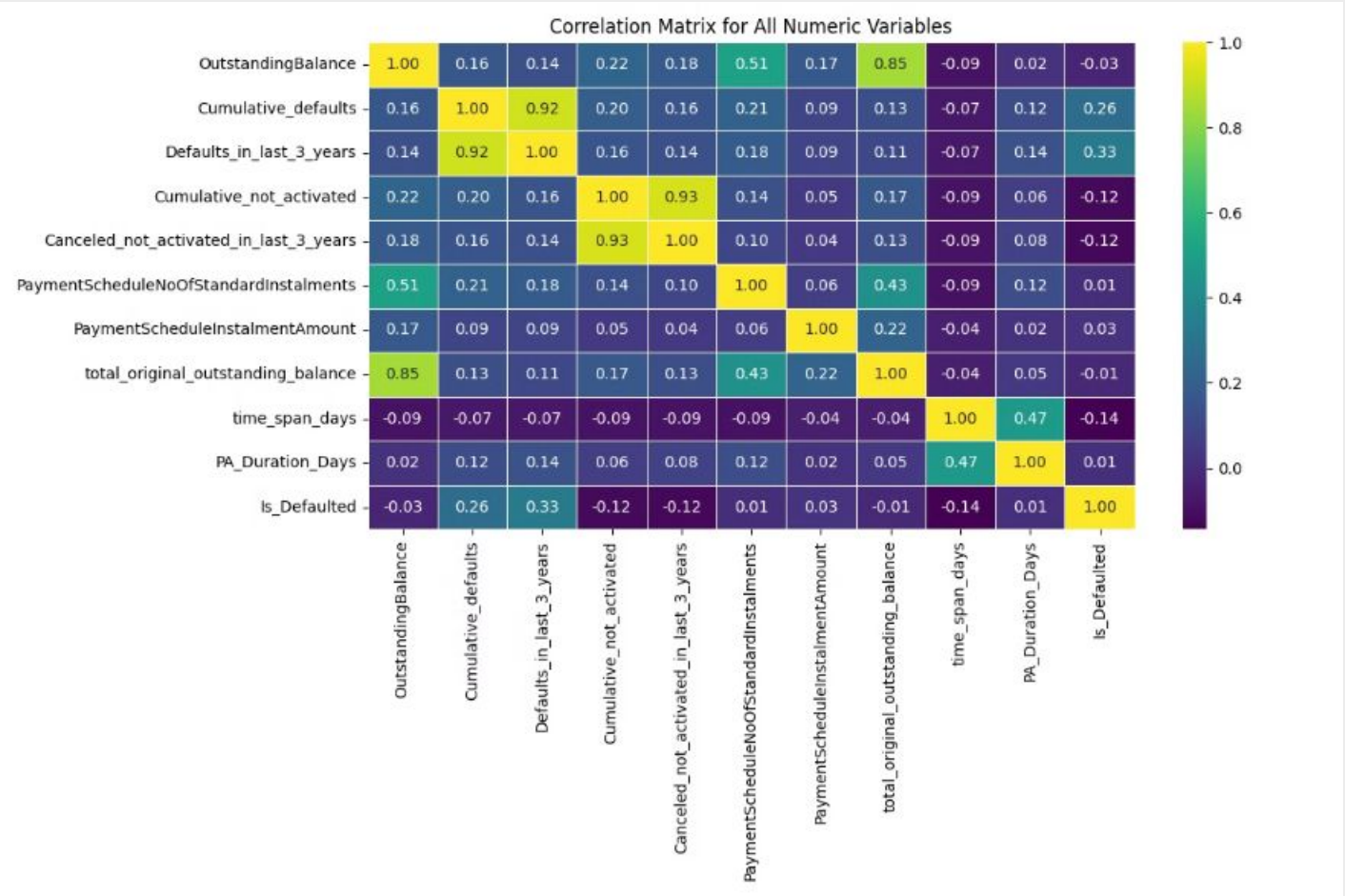
Key Correlations:

- OutstandingBalance and Cumulative_defaults show moderate positive correlations.
- PA_Duration_Days has correlations with PaymentScheduleInstalmentAmount and OutstandingBalance, indicating that the duration of a payment arrangement could impact the total outstanding balance.

Visual:

Heatmap: Correlation matrix of all numeric variables with detailed annotations.

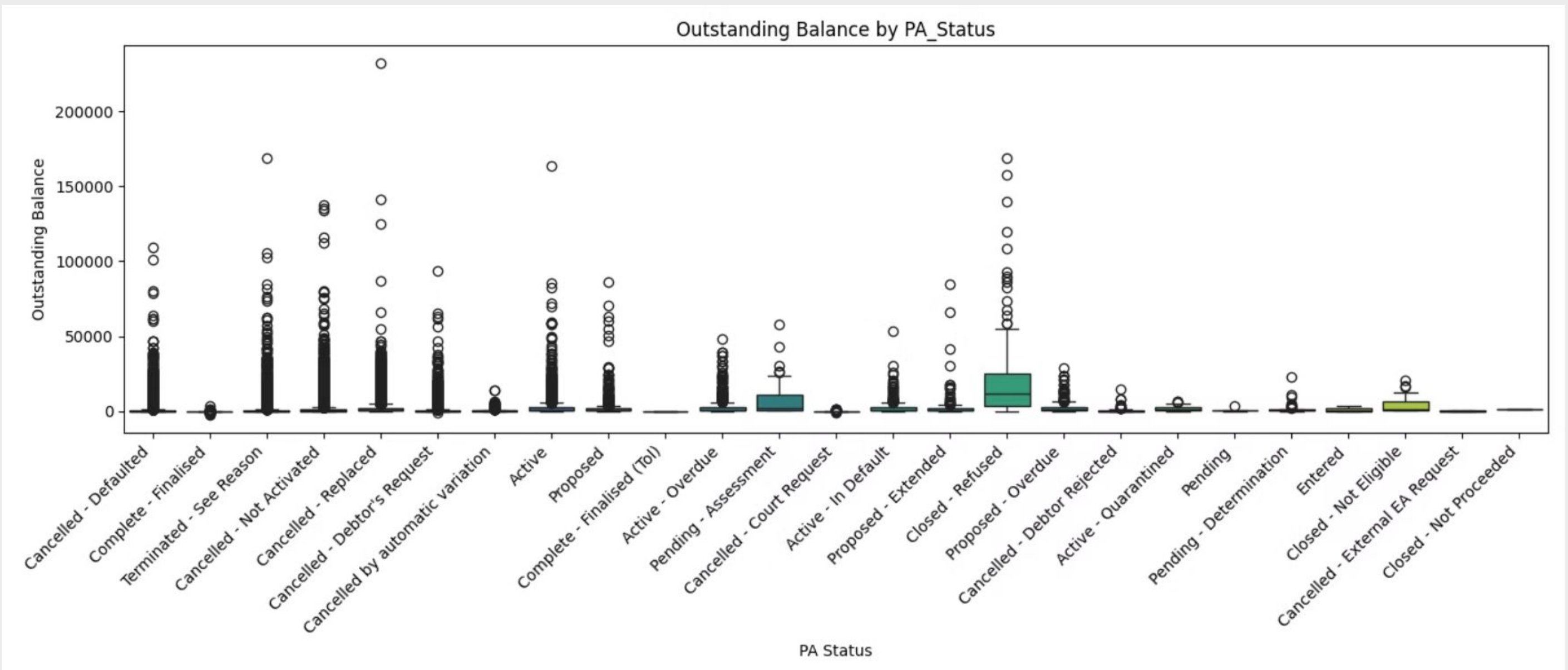
- This heatmap allows quick identification of strongly correlated features for model consideration.



EDA: Outstanding Balance by PA_Status

Wide Range of Balances: Significant variation in outstanding balances across different statuses.

High Outliers: Some statuses (e.g., defaults, canceled) show high outliers, indicating large unpaid amounts.



Modeling and Focus on Default Detection

Model Overview: The RandomForest model shows an overall accuracy of 86%.

Key Insight: We are less concerned about the overall accuracy, as the primary goal is to improve the model's ability to detect **defaults (class 1)**.

Confusion Matrix Analysis:

Correctly classified **1,588** defaulters, but missed **2,229** actual defaults.

- Precision for defaulters (class 1) is 0.74, while recall is 0.42, indicating room for improvement in identifying more defaulters.

Next Step: Feature engineering will be applied to enhance the model’s default detection capability.

Confusion Matrix:					
[[15629 554]					
[2229 1588]]					
Classification Report:					
	precision	recall	f1-score	support	
0	0.88	0.97	0.92	16183	
1	0.74	0.42	0.53	3817	
accuracy			0.86	20000	
macro avg	0.81	0.69	0.73	20000	
weighted avg	0.85	0.86	0.84	20000	
Accuracy Score: 0.86085					

Key Feature Analysis through Pairplots and Boxplots

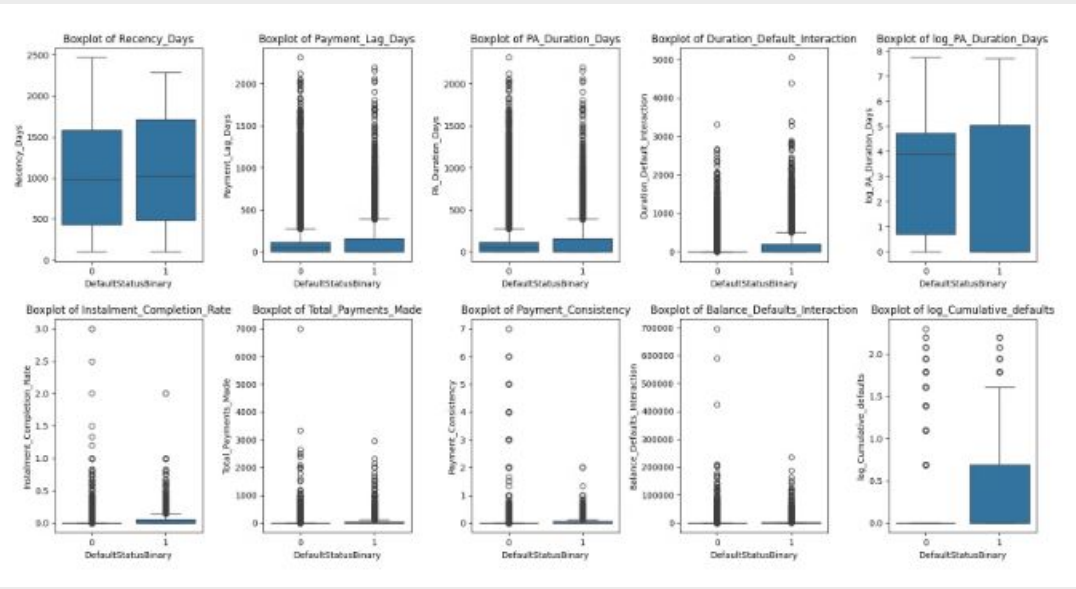
Pairplot Analysis: The pairplot visualizes relationships between key variables such as OutstandingBalance, PaymentScheduleInstalmentAmount, Cumulative_defaults, and PA_Duration_Days across the defaulted and non-defaulted groups.

Boxplot Analysis: The boxplot illustrates the differences in OutstandingBalance across the PA_Status categories, revealing that higher balances are more prevalent in default cases.

Visual:

Pairplot: Showcasing relationships between key variables and their interaction with default status.

Boxplot: Distribution of OutstandingBalance across PA_Status categories, emphasizing the difference between defaulted and non-defaulted debtors.



Results after feature engineering

Improvement in Precision: The precision for defaulters (class 1) increased to **0.74**, meaning when the model predicts a default, it's more likely to be correct.

Consistent Overall Accuracy: The model's overall accuracy remains strong at **0.86085**.

Persistent Recall Issue: The recall for defaulters dropped to **0.42**, indicating the model still fails to identify a significant number of actual defaulters.

Key Concern: Misclassification of defaulters as non-defaulters remains a major issue that needs to be resolved to improve default prediction.

Confusion Matrix:					
[[14927 1256]					
[1423 2394]]					
Classification Report:					
	precision	recall	f1-score	support	
0	0.91	0.92	0.92	16183	
1	0.66	0.63	0.64	3817	
accuracy			0.87	20000	
macro avg	0.78	0.77	0.78	20000	
weighted avg	0.86	0.87	0.86	20000	
Accuracy Score:					
0.86605					

Model Performance

After Tuning

Tuning:

- **SMOTE:** Oversampled the minority class (defaulters) to handle class imbalance.
- **Class Weighting:** Penalized the model for misclassifying defaulters with a 1:3 weight ratio.
- **Threshold Adjustment:** Tuned the classification threshold to 0.4 for better recall of defaulters.

Results:

- **Recall for Defaulters:** Improved to **1.00**, correctly identifying all defaulters.
- **Precision for Defaulters:** Dropped to **0.49**, but recall was prioritized.
- **Accuracy:** Decreased to **79.88%**, but the goal of predicting defaulters accurately was achieved.

Confusion Matrix:					
[[12173 4010]					
[13 3804]]					
Classification Report:					
	precision	recall	f1-score	support	
0	1.00	0.75	0.86	16183	
1	0.49	1.00	0.65	3817	
accuracy			0.80	20000	
macro avg	0.74	0.87	0.76	20000	
weighted avg	0.90	0.80	0.82	20000	
Accuracy Score:					
0.79885					

What we solved and how it is beneficial to the team?

Improved Default Prediction: We enhanced the ability to predict which debtors are likely to default, allowing for more accurate identification of those at risk.

Actionable Insights: The model now provides useful insights that can help Fines Victoria take early action on high-risk debtors, improving debt collection strategies.

Business Impact: By predicting defaulters more effectively, Fines Victoria can focus efforts on key cases, reduce financial losses, and manage debts more efficiently.

Overall Value to the Project: The solution offers better decision-making tools that will lead to improved processes, contributing to the overall success of Fines Victoria's debt recovery initiatives.

Limitations and Challenges

1

Limited Data on Defaulters

The dataset had a limited number of defaulters, which impacted the ability to train the model effectively for predicting defaults.

2

Prediction Trade-off

While we improved the model's ability to predict defaulters, it led to a decrease in the accuracy of predicting non-defaulters. The model requires further development to achieve balanced accuracy for both classes

3

Real-World Application Challenges:

The model, while predictive, may face challenges in real-time application due to dynamic changes in debtor behavior that may not be captured in the historical data.

Conclusion

- Successfully built a predictive model to identify potential debtor defaults based on historical data.
- Achieved notable improvement in defaulter prediction through feature engineering and tuning techniques.
- However, trade-offs were observed in predicting non-defaulters, highlighting the need for further optimization.
- The model serves as a valuable tool for Fines Victoria to take proactive actions, though ongoing refinement is required to balance prediction accuracy across both defaulters and non-defaulters.

Thank You

We are open for discussion