Comprehensive Analysis of Debtor Payment Behavior and Default Risk Prediction for Fines Victoria

Prabhath Kiran Ummiti

Jay Sangani

In Collaboration with Department of Government Services,

State of Victoria, Australia

Department of Econometrics & Business Statistics

✉ pumm0001@student.monash.edu
✉ jsan0062@student.monash.edu

Student_ID: 33620296 | 32862741

Table of Contents

1 Abstract

The **"Comprehensive Analysis of Debtor Payment Behavior and Default Risk Prediction for Fines Victoria"** aims to enhance debt recovery and compliance rates through predictive analytics. This project is a collaborative effort between the **Department of Government Services (DGS)** and **Monash University's** Master of Business Analytics program, involving the **Data Analytics and Services (DAaS)** team. It addresses the critical issue of rising defaults on fines, which currently contribute to an estimated $150 million annual shortfall for the state of Victoria, Australia.

The project focuses on developing a machine learning model that predicts the likelihood of debtor defaults, enabling **Fines Victoria** to implement proactive interventions. Using historical payment data, the model identifies patterns in debtor behavior and assesses factors that contribute to payment defaults. The model is designed to prioritize early identification of high-risk debtors, allowing for targeted measures that improve compliance and reduce manual recovery efforts.

The findings from this analysis demonstrate the utility of predictive analytics in optimizing resource allocation and supporting data-driven decision-making within public sector operations. By accurately forecasting defaulters, **Fines Victoria** can enhance operational efficiency, allocate resources more effectively, and achieve a higher rate of debt recovery. The project not only addresses immediate challenges but also sets a precedent for scalable analytical solutions that can be adapted for similar compliance-related initiatives across government departments.

Future efforts may include integrating more detailed debtor data, exploring alternative modeling approaches, and refining feature engineering to further improve accuracy. The project underscores the broader potential of data analytics in public service, illustrating how machine learning can drive effective policy implementation and financial stability.

2 Background and Motivation

2.1 Background

2.1.1 Overview of Fines Victoria

Fines Victoria is a Victorian Government administrative body established under the **Fines Reform Act 2014**. It is responsible for managing the administration and enforcement of infringement fines and court fines across the state. The organization oversees various stages of the fines lifecycle, which includes:

- **Issuing infringement notices**: Sent on behalf of enforcement agencies.
- **Collecting payments**: Ensuring fines are paid in a timely manner.
- **Reviewing fines**: Handling eligible fines upon application.
- **Taking enforcement action**: Pursuing individuals who do not resolve their fines.
- **Social justice initiatives**: Administering programs like the **Family Violence Scheme** and the **Work and Development Permit Scheme** to assist vulnerable individuals in managing their fines (Fines Victoria, 2024).

2.1.2 Role and Responsibilities

Fines Victoria manages fines related to:

- Speeding and red-light offenses.
- Tolling violations.
- Unregistered vehicle use.
- Court-ordered fines.
- Marine safety, gambling, and liquor-related fines.
- Other fines registered by enforcement agencies for collection.

2.1.3 Governance and Decision-Making

The administration and enforcement of fines are under the jurisdiction of the **Director of Fines Victoria**, who acts according to statutory functions outlined in the **Fines Reform Act 2014** and the **Infringements Act 2006**. The Director is supported by Fines Victoria in exercising these powers, following specific administrative law requirements that ensure:

- **Good faith actions and proper purposes**.
- **Compliance with legislative procedures**.
- **Consideration of relevant matters while ignoring irrelevant ones**.
- **Reasonable decision-making and objectivity**.

Fines Victoria strives to maintain the integrity of the fines system while considering the impact on individuals, particularly the vulnerable. The organization aims to reduce the number of people progressing to serious enforcement stages, such as sanctions, warrants, or court hearings, by offering clear information and support. As the lifecycle process can be seen in Figure 1.

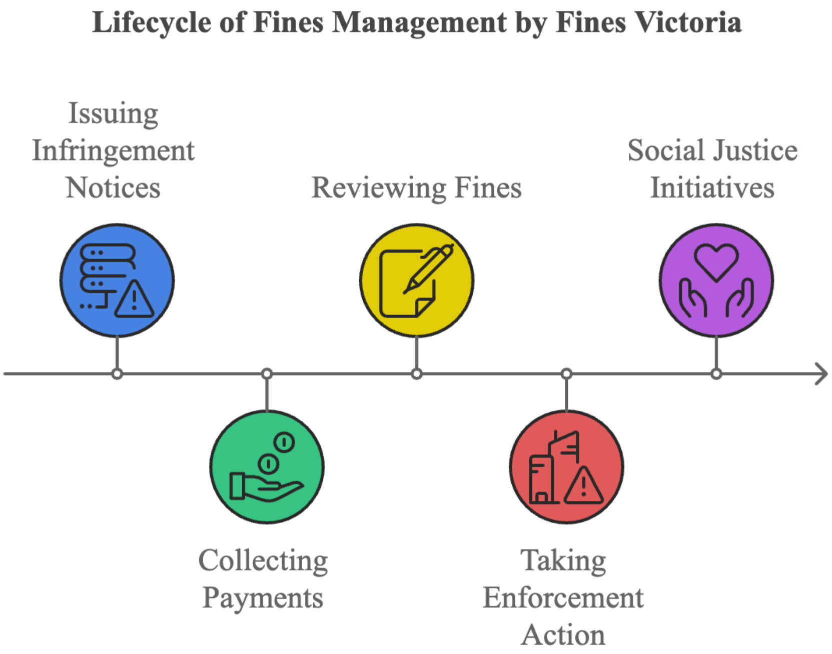

Lifecycle of Fines Management by Fines Victoria
Issuing Infringement Notices
Reviewing Fines
Social Justice Initiatives
Collecting Payments
Taking Enforcement Action

Figure 1: Glimpse of Fines Victoria process

2.2 Project Initiation and Involvement

2.2.1 Our Role in the Project

As part of the **Master of Business Analytics program** at Monash University, we—Prabhath Kiran Ummiti and Jay Sangani—joined the **Data Analytics and Services (DAaS)** team under the **Department of Government Services (DGS)** through the **ETC5543 - Business Creative Activity Unit**. This unit emphasizes practical application of data analytics skills in real-world projects, making it an ideal platform for us to contribute to the "Comprehensive Analysis of Debtor Payment Behavior and Default Risk Prediction for Fines Victoria."

The project aligns with DGS's strategic efforts to improve operational efficiency within the Victorian Government, specifically targeting the high default rates experienced by Fines Victoria. This collaboration provided us with hands-on experience in public sector analytics, allowing us to contribute meaningful insights while enhancing our skills in data analysis and predictive modeling.

2.3 Introduction to DGS and DAaS Partnership

2.3.1 Department of Government Services (DGS)

The **Department of Government Services (DGS)** is a central agency tasked with supporting various government departments through strategic consulting, technical solutions, and process optimization. Its approach focuses on cost-effective solutions, operating at about 50% of the cost compared to private consultancies, by leveraging in-house expertise through the **Data Analytics and Services (DAaS)** team.

2.3.2 Role of DAaS in Government Projects

The **Data Analytics and Services (DAaS)** team within DGS specializes in developing scalable data-driven solutions that enhance decision-making and operational efficiency. It acts as a strategic partner for government agencies, offering insights that drive policy improvements, optimize processes, and support informed decision-making. DAaS's services are user-centric, ensuring that solutions are tailored to meet specific agency needs, as seen in its collaboration with Fines Victoria.

2.3.3 Key Principles of DAaS Operations

- **User-Centric Approach**: Solutions are designed to be intuitive, focusing on the needs of end-users.
- **Collaborative Delivery**: DAaS maintains a unified delivery team with client agencies, ensuring aligned goals and iterative planning.
- **Safe and Ethical Practices**: Emphasis is placed on data privacy, security, and ethical use of analytics across all projects.
- **Quality Assurance**: DAaS aims to develop scalable models that reduce manual efforts and foster continuous improvement.

2.3.4 DaAS-FES Partnership Overview

The **Strategic Partnerships, Analytics, and Innovation (DaAS)** division works in tandem with DAaS to implement large-scale transformation projects, such as the **Fines Enforcement Services (FES)**. This partnership aims to enhance fine management across Victoria by addressing challenges like increased compliance, reduced defaults, and improved resource allocation. The DaAS-FES collaboration aligns with broader government goals of optimizing processes through data-driven insights and improving cash flow management.

2.3.5 Phases of the DaAS-FES Collaboration

- **Discovery Phase (Phase 1)**: Involves initial data assessment, stakeholder consultations, and scope definition to identify pain points within the FES workflow. We participated in this phase by analyzing data characteristics and mapping existing processes (Fines Victoria, 2024).
- **Deep Dive Analysis (Phase 2)**: This phase includes process mapping, feasibility studies, and data modeling to understand patterns and drivers of defaults. Our work in this phase focused on developing a predictive model that accurately identifies debtors at risk of defaulting.
- **Implementation and Optimization (Phase 3)**: In this phase, solutions are integrated into Fines Victoria's systems, business rules are refined, and workforce requirements are modeled to handle increasing demand.

2.4 Project Context and Fines Victoria

2.4.1 Challenges Faced by Fines Victoria

Fines Victoria is tasked with managing a critical aspect of state finance—fine collection and compliance. However, it faces several challenges that affect its effectiveness:

- **High Default Rates:** Currently, 25-30% of fines result in defaults, leading to a significant revenue shortfall of approximately $150 million annually. This not only impacts the state's financial health but also requires substantial resources for recovery efforts, which can be resource-intensive and inefficient.
- **Operational Strain:** The demand for Level-2 (L2) support, responsible for handling complex payment arrangements and defaults, has surged by 40-65% year-on-year. This increase in demand has placed immense pressure on available resources, making it difficult to allocate resources efficiently. The lack of predictive capabilities further

complicates this challenge, as it is hard to identify which debtors are most likely to default.

- **Economic Pressures:** Economic factors, such as rising costs of living, further increase the likelihood of defaults, emphasizing the need for Fines Victoria to adopt a proactive, predictive approach to managing debtors.

2.5 Our Role and Focus in the Project

2.5.1 Developing a Predictive Model

Our primary focus in the **Fines Enforcement Services (FES)** project was to develop a **predictive model** that identifies debtors likely to default. The model analyzes historical payment data and debtor characteristics, aiming to improve recall rates for defaulters and enable timely interventions. By targeting debtors with the "Canceled - Defaulted" status, we sought to provide Fines Victoria with insights that support more effective compliance measures.

2.5.2 Learning and Practical Experience

Engaging in this project allowed us to apply various analytical techniques, including data preprocessing, feature engineering, and model evaluation. Collaborating with the DAaS team deepened our understanding of public sector challenges, providing hands-on experience in developing predictive analytics for real-world applications.

2.5.3 Integration with DAaS's Broader Objectives

The predictive model we developed aligns with DAaS's mission of delivering scalable analytical solutions that enhance government operations. By generating actionable insights, the model supports improved business rules, better resource management, and more efficient compliance strategies for Fines Victoria. We believe this model can be adapted for similar challenges across other departments, reinforcing the effectiveness of in-house consultancy in the public sector.

3 Objectives and Significance

3.1 Analysis Objectives

The "Comprehensive Analysis of Debtor Payment Behavior and Default Risk Prediction for Fines Victoria" aims to transform debt recovery by leveraging advanced data analytics. Our project, as interns within the DAaS team, specifically focuses on predictive modeling, improving compliance, and optimizing resource allocation within Fines Victoria.

3.1.1 Primary Objectives

- **Develop a High-Recall Predictive Model**:
    - The core objective is to create a machine learning model that accurately predicts debtors at risk of default, with a focus on maximizing recall for defaulters to ensure early identification.
    - By specifically targeting the "Cancelled - Defaulted" payment status, the model is designed to provide timely alerts, supporting proactive interventions.
- **Enable Proactive Debt Management**:
    - By identifying high-risk debtors early, Fines Victoria can implement targeted interventions such as personalized payment plans, more frequent reminders, thus reducing the $150 million revenue gap caused by defaults.
- **Enhance Operational Efficiency**:
    - The project aims to optimize resource allocation by focusing efforts on high-risk debtors, thereby improving manpower efficiency and minimizing unnecessary manual interventions.
- **Generate Strategic Insights**:
    - The model will provide actionable insights that inform strategic decision-making, aiding in the refinement of debtor management policies and compliance measures.

3.1.2 Project-Specific Tasks

- **Data Exploration and Preprocessing**:

  - Analyze historical payment data to identify key patterns and variables that influence debtor behavior.
  - Handle missing values, transform skewed data, and manage outliers to ensure a clean, robust dataset for model development.
- **Feature Engineering and Model Building**:
  - Implement advanced feature engineering to create variables that improve predictive performance, such as payment frequency, outstanding balance ratios, and recency of payments.
  - Utilize machine learning algorithms like Random Forest, Decision Trees, and Gradient Boosting to build a model optimized for recall.
- **Model Evaluation and Deployment**:
  - Evaluate model performance using metrics like recall, precision, F1-score, and AUC-ROC, prioritizing defaulter identification.
  - Deploy the model within Fines Victoria’s existing systems to support real-time decision-making and interventions.

3.2 Team Structure and Collaboration

3.2.1 Team Composition and Roles

The DAaS team comprises a mix of data scientists, data analysts, project officers, and technical specialists, all working collaboratively to drive data-driven solutions for government agencies. Key team members include:

- **John Gehman**: Managing Technical Specialist
- **Hana Basyoni**: Data Scientist
- **James Farnell**: Data Scientist
- **Cameron Bolton**: Data Analyst
- **Fiona Lu**: Data Analyst
- **Lanny Tieu**: Project Officer
- **Allanagh O'Donnell**: Manager of Priority Projects

This diverse team structure enables effective problem-solving and ensures that the project benefits from varied expertise, ranging from technical model building to strategic project management.

3.2.2 Tools and Software Used

We rely on a comprehensive set of tools and software to ensure secure data handling, accurate model development, and effective project management:

- **Jira**: Used for task management, planning, tracking, and managing project tasks, supporting effective collaboration among team members.
- **GitLab**: Employed for version control and collaboration, ensuring secure code storage and tracking changes throughout the development process.
- **Azure**: Serves as the primary database management platform, providing secure data storage, processing, and access for large datasets.
- **SharePoint**: Facilitates document sharing and collaboration, enabling efficient project documentation and communication.
- **Postico**: Used for data access and collaboration, specifically for querying and managing the PostgreSQL database.
- **Visual Studio Code (VS Code)**: Integrated Development Environment (IDE) used for coding, debugging, and developing the predictive model in Python.
- **Python**: The main programming language used for data preprocessing, feature engineering, and model development. Key libraries include Pandas, NumPy, Scikit-learn, and XGBoost.
- **SQL**: Utilized for database management, enabling data extraction, transformation, and analysis.

3.2.3 Agile Framework and Workflow

The project follows an agile framework with iterative sprints to ensure continuous development, regular feedback, and timely adjustments:

- **Sprint Planning**: Takes place every Tuesday, setting clear goals for the week and aligning tasks with overall project objectives.

- **Sprint Execution**: Focuses on carrying out the defined tasks, such as data cleaning, model building, or evaluation, while ensuring collaboration among team members.
- **Sprint Review**: Conducted every Friday, where progress is assessed, challenges are discussed, and necessary adjustments are made to the workflow.
- **Tools for Agile Implementation**: Jira is extensively used for task management, while Confluence supports documentation, progress tracking, and knowledge sharing. This approach allows for flexibility and adaptability, critical for handling the evolving nature of the dataset and analysis requirements.

3.3 Methodology and Data Handling

As shown in Figure 2, this gives a general overview of the process followed by the whole team on this project.

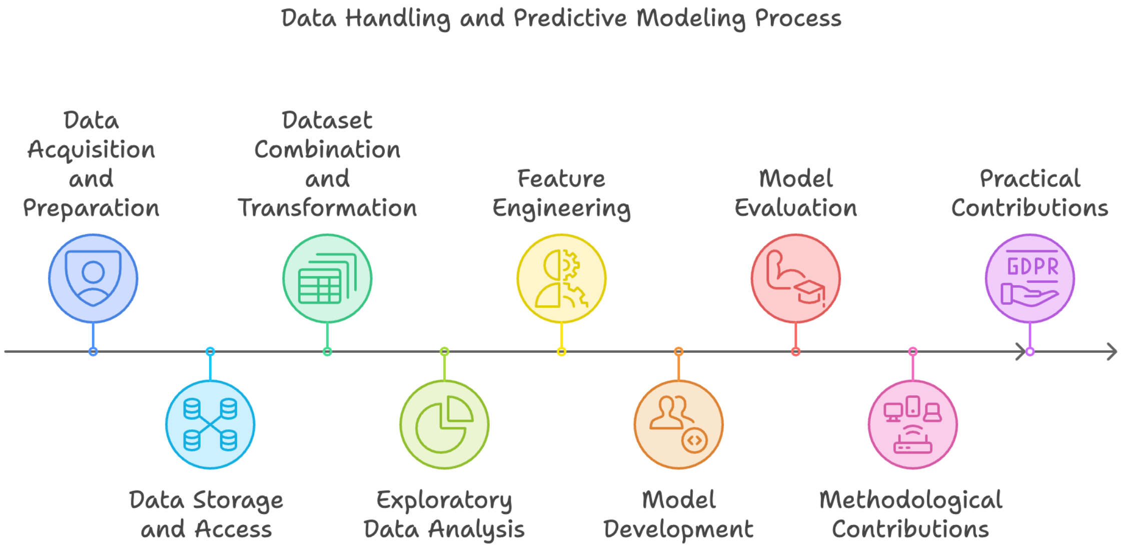

Data Handling and Predictive Modeling Process
Data Acquisition and Preparation
Data Storage and Access
Dataset Combination and Transformation
Exploratory Data Analysis
Feature Engineering
Model Development
Model Evaluation
Methodological Contributions
Practical Contributions
GDPR

Figure 2: Overview of our analysis

3.3.1 Data Acquisition and Preparation

The datasets were provided by **Fines Victoria** and consisted of multiple raw datasets related to debtor payments, defaults, and demographic information. These datasets included a mix of structured and unstructured data, requiring extensive preprocessing and transformation to make them usable for predictive modeling.

- **Pre-Data Acquisition Strategies**: Our data engineering and architecture teams set up secure data access protocols, ensuring that the datasets were hashed to protect debtor privacy and confidentiality. The data was stored securely in **Azure** and accessed via **Postico**, facilitating a controlled and secure data environment.
- **Combining and Transforming Datasets**: After preprocessing, the datasets were combined into a single, more valuable dataset that included enhanced features like balance-to-payment ratios, payment frequency metrics, and recency of payments. This integrated dataset was then used for further analysis and model building.

3.3.2 Data Analysis and Feature Engineering

- **Exploratory Data Analysis (EDA)**: We conducted EDA to understand variable distributions, identify key correlations, and detect patterns in debtor behavior. This step was crucial for feature selection and provided insights into which variables significantly influence default behavior.
- **Feature Engineering**: Advanced feature engineering techniques were applied to create variables that improved model accuracy and recall. Key features included interaction variables, such as the ratio of outstanding balance to payment frequency, and recency of payments, which provided more nuanced insights into debtor behavior.

3.3.3 Model Development and Evaluation

- **Model Development**: We implemented a range of machine learning algorithms, including Random Forest, Decision Trees, and Gradient Boosting, focusing on maximizing recall for defaulters. Hyperparameter tuning was performed using **GridSearchCV**, while SMOTE was applied to handle class imbalances.
- **Model Evaluation**: Performance was evaluated using metrics like recall, precision, F1-score, and AUC-ROC, with a priority on recall to ensure effective identification of defaulters. Thresholds were dynamically adjusted to balance sensitivity and specificity, aligning the model's performance with Fines Victoria's operational needs.

3.4 Contributions of the Analysis

3.4.1 Methodological Contributions

- **Advanced Feature Engineering**: We introduced novel features such as payment frequency, balance ratios, and interaction terms that significantly enhanced model accuracy and recall. This demonstrates the importance of feature engineering in public sector analytics.
- **Dynamic Threshold Adjustment**: By adjusting the classification threshold, we prioritized defaulter identification, aligning the model's design with Fines Victoria's operational needs.
- **Ethical data Implementation**: The model adheres to ethical data guidelines, ensuring responsible use of data, transparency, and compliance with privacy regulations.

3.4.2 Practical Contributions

- **Improved Compliance Strategies**: The model supports more precise compliance measures, focusing resources on high-risk debtors and improving recovery rates.
- **Enhanced Operational Efficiency**: By reducing manual workloads and optimizing resource allocation, the model streamlines debt recovery processes and improves productivity.
- **Policy Implications**: The model provides insights that inform potential policy adjustments, helping refine strategies for debtor management.

3.5 Significance of the Analysis

3.5.1 Addressing Fines Victoria's Key Challenges

- **Reducing Default Rates**: The model directly addresses the 25-30% default rate, providing data-driven solutions that improve compliance and reduce financial shortfalls.
- **Optimizing Resource Allocation**: By accurately identifying high-risk debtors, the model improves resource efficiency, reducing operational strain.
- **Ensuring Financial Stability**: The project's focus on reducing defaults contributes to more stable public service funding and financial management.

3.5.2 Broader Implications for Public Sector Analytics

- **Scalability**: The model's methodology is scalable, making it adaptable to other compliance challenges across the public sector, such as tax recovery or welfare management.
- **Promoting a Data-Driven Culture**: The project fosters innovation within the Victorian Government by showcasing the benefits of predictive analytics, encouraging broader adoption of similar strategies.

4 Methodology

Our approach was shaped by the theoretical frameworks and practical skills from the Master of Business Analytics program at Monash University. It aligned with course outcomes by integrating analytical concepts into real-world applications through a systematic, iterative process. We structured the methodology into three main phases—Planning, Implementation, and Optimization—using a cyclic workflow to enable continuous model development and refinement.

4.1 Learning-Driven Methodology Framework

The structure was influenced by the learning curve from the Masters of Business Analytics course, which emphasizes end-to-end analysis while ensuring alignment with business objectives. This framework consisted of three stages:

4.1.1 Planning Phase

- **Objective Setting:** We defined key variables and model requirements based on Fines Victoria's context, prioritizing recall for defaulters to minimize missed cases.
- **Stakeholder Engagement:** Early discussions with DAaS, Fines Victoria, and DGS provided insights into data characteristics and strategic goals, ensuring a user-centric modeling approach aligned with real-world decision-making.
- **Data Exploration:** Initial exploration helped us understand data structure, distribution, and quality issues, laying the foundation for informed feature engineering and model development.

4.1.2. Implementation Phase

- **Data Preparation:** Comprehensive preprocessing addressed missing values, skewness, and outliers, while advanced feature engineering transformed raw data into predictive variables, applying Monash-taught skills.
- **Model Building:** We explored algorithms like Random Forests, Decision Trees, and Neural Networks, focusing on optimizing recall for defaulters, a strategy emphasized in the coursework.
- **Evaluation and Refinement:** Iterative evaluation using recall, precision, and F1-score ensured the model met Fines Victoria's requirements. Techniques like GridSearchCV and SMOTE were used for tuning and managing class imbalance, demonstrating advanced modeling strategies from the program.

4.1.3 Optimization Phase

- **Performance Enhancement:** Further refinements were made through feature engineering, threshold adjustments, and ensemble methods to maximize recall without compromising precision.
- **Deployment Planning:** A strategy was developed to integrate the model into Fines Victoria's systems, including visual dashboards for effective communication of outcomes to stakeholders.

4.2 Cyclic Workflow and Iterative Development

While structured into three main phases, our methodology also followed a cyclic workflow—consisting of **Idea, Execution, Prediction, and Accuracy**—to promote continuous learning and refinement. This approach, aligned with the Monash Business Analytics program, ensured adaptability throughout the model development process and alignment with Fines Victoria's requirements.

4.2.1 Cyclic Process

- **Idea Generation:** We started each cycle by generating hypotheses, identifying new features, and planning model adjustments based on analysis and feedback.

Collaborative brainstorming and supervisor guidance helped prioritize ideas that improved recall, precision, or overall performance.

- **Execution:** This phase translated ideas into action, involving feature engineering, data preprocessing, and model adjustments. Using tools like Python, SQL, and GitLab, we managed data pipelines and developed models, reflecting Monash's focus on practical application.
- **Prediction:** Models were then used to generate predictions, focusing on maximizing recall to effectively identify defaulters. Performance was assessed for both accuracy and business impact, with regular discussions with the DAaS team and our supervisor to ensure alignment with Fines Victoria's goals.
- **Accuracy Assessment:** The final stage involved evaluating metrics like recall, precision, F1-score, and AUC-ROC to ensure the model was effectively identifying defaulters while maintaining precision. Adjustments were made based on these evaluations, guiding decisions on whether to refine the model further or revisit the "Idea" phase.

4.2.2 Iterative Decision-Making and Learning Integration (change it as per model selection in IDA)

This cyclic process was employed across all phases, from Initial Data Analysis (IDA) to Exploratory Data Analysis (EDA) and model building, enabling adaptive learning and improvement. Supervisor feedback after each cycle guided informed decision-making, ensuring data-driven model development and alignment with business objectives.

This iterative approach reflects the problem-solving methods taught in the Monash program, emphasizing the importance of iterative development in successful data projects. It led to the creation of a recall-focused predictive model that addresses Fines Victoria's operational challenges.

The following sections will provide a detailed breakdown of each phase, beginning with the Model Selection.

4.3 Model Selection

Model selection was an intricate process that required extensive trial and error, underpinned by comprehensive experimentation, performance evaluations, and theoretical considerations. This section outlines the journey of model exploration, highlighting why **Random Forest** was ultimately chosen as the best-suited model for this project, given the dataset's characteristics, target variable behavior, and the complexity of the predictive task.

4.3.1 Logistic Regression

Logistic Regression was the first model tested due to its simplicity, interpretability, and effectiveness in binary classification tasks. As a linear model, it predicts the probability of default using the following formula:

$$\hat{y} = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i \cdot x_i)}}$$

Where:

- $\hat{y}$ is the predicted probability of default.
- β_0 is the intercept term.
- β_i represents the coefficients of each feature x_i
- n is the number of features.

4.3.2 Challenges with Logistic Regression:

- **Linearity Limitation:** Logistic Regression assumes a linear relationship between the features and the log-odds of the target variable. However, the dataset exhibited complex non-linear interactions among variables like Recency_Days, Payment_Lag_Days, and Cumulative_defaults, which Logistic Regression struggled to capture.

- **Poor Handling of Imbalanced Data:** Even after applying oversampling techniques like SMOTE, Logistic Regression's performance was suboptimal, particularly for defaulters (class 1). The recall remained low, indicating that the model failed to capture many true defaulters.
- **Lack of Feature Interactions:** Logistic Regression cannot inherently model complex feature interactions unless they are explicitly engineered, which adds another layer of preprocessing complexity.

4.3.3 Neural Network Exploration

Given the potential non-linearity in the data, we moved to more advanced models like **Neural Networks**, hoping to better capture complex patterns and interactions.

- **Structure:** The Neural Network used was a multi-layer perceptron (MLP) with two hidden layers, incorporating ReLU activation for non-linearity and a sigmoid function in the output layer to predict the probability of default.

$$\text{Hidden layer activation: } a_i = \max(0, \sum_j w_{ij} \cdot x_j + b_i)$$

$$\text{Output layer activation: } \hat{y} = \frac{1}{1 + e^{-\sum_k w_k \cdot a_k}}$$

4.3.4 Challenges with Neural Networks:

- **Overfitting:** Despite regularization techniques like dropout and L2 regularization, the Neural Network quickly overfit the training data due to the relatively small number of defaulters and the limited dataset size. It failed to generalize well to unseen data.
- **Computational Expense:** Neural Networks required significant computational resources and longer training times, which slowed down the iteration process.
- **Black-box Nature:** Unlike Random Forest, Neural Networks lack interpretability, making it challenging to understand which features contribute most to predicting defaults, which was a crucial requirement for stakeholder insights.

4.3.5 Why Random Forest?

After observing the limitations of linear models and neural networks, **Random Forest** emerged as the most promising model due to its flexibility, non-linearity, and robustness. It operates as an ensemble learning method that builds multiple decision trees and combines their predictions through majority voting. The model can be mathematically represented as:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x)$$

Where:

- T is the total number of trees.
- $f_t(x)$ is the prediction from the ttt-th tree.
- $\hat{y}$ is the average prediction across all trees.

Advantages of Random Forest:

1. **Non-linearity and Interaction Handling:** Random Forests are inherently capable of capturing complex interactions between variables without explicit feature engineering. This was critical in this dataset, which had significant non-linear relationships among features.
2. **Handling Imbalanced Classes:** By incorporating class weights (e.g., setting a 1:3 ratio to penalize misclassifying defaulters), Random Forests could be tuned to focus more on accurately predicting defaulters while maintaining general accuracy for non-defaulters.
3. **Interpretability:** Feature importance scores allowed for clear interpretation of which variables were most influential in driving default predictions, satisfying stakeholder needs.

4. **Robustness:** Random Forests are less prone to overfitting compared to neural networks, especially with the use of techniques like cross-validation, pruning, and regularization through hyperparameter tuning.
5. **Scalability:** Random Forests are computationally less expensive than neural networks, making them faster and easier to train.

Key Elements of Random Forest's Success:

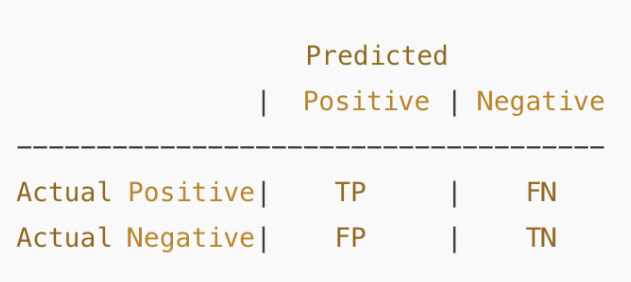
1. **Hyperparameter Tuning:**
   - Extensive grid search was conducted for key parameters like n_estimators, max_depth, min_samples_split, and min_samples_leaf, resulting in optimal performance while preventing overfitting.
   - The best parameters improved recall, precision, and overall accuracy, demonstrating the model's adaptability to different settings of the data.
2. **Feature Importance:**
   - Features like Recency_Days, Payment_Lag_Days, and Balance_Defaults_Interaction were ranked as top predictors. These variables align with domain knowledge, emphasizing that the model not only captures statistical patterns but also aligns with business logic.
   - By interpreting feature importance, stakeholders could better understand which factors drive defaults, aiding in more targeted interventions and decision-making.
3. **Handling Imbalance with SMOTE and Class Weights:**
   - SMOTE was combined with class weighting, ensuring that the model became sensitive to the minority class (defaulters). This combination led to improved recall for defaulters while maintaining balanced accuracy.
4. **Threshold Adjustment:**
   - Adjusting the classification threshold to 0.4 improved the model's recall, ensuring that the majority of defaulters were identified, even if it meant a slight drop in precision. This decision aligns with the business objective of minimizing missed defaulters, as missing a defaulter is considered costlier than a false positive.

Ultimately, the Random Forest model struck the perfect balance between achieving high recall for defaulters and maintaining good accuracy for non-defaulters, making it the **optimal choice** for this predictive task.

4.3.6 Confusion Matrix and Classification Report

Confusion Matrix

- The confusion matrix as shown in Figure 3, is a crucial evaluation tool for classification models, representing predictions across four categories:



```
                   Predicted
                  |  Positive | Negative
------------------------------------------
Actual Positive|    TP      |     FN
Actual Negative|    FP      |     TN
```

Figure 3: Confusion matrix method

  - **True Positives (TP):** Correctly predicted defaulters.
  - **True Negatives (TN):** Correctly predicted non-defaulters.
  - **False Positives (FP):** Non-defaulters incorrectly predicted as defaulters (type I error).
  - **False Negatives (FN):** Defaulters missed by the model (type II error).
- It enables the calculation of multiple performance metrics, offering a clearer insight into model effectiveness.

Why Focus on Precision, Recall, and F1-Score:

- **Precision:** Calculated as Precision = $\frac{TP}{TP+FP}$, it emphasizes the model's reliability in predicting defaulters correctly. High precision means fewer false positives, which is critical to avoid unnecessary interventions.

- **Recall (Sensitivity):** Calculated as Recall = $\frac{TP}{TP+FN}$, it measures how well the model identifies actual defaulters. High recall is crucial because missing defaulters can have serious financial consequences.
- **F1-Score:** Defined as the harmonic mean of precision and recall, F1 = 2 × $\frac{Precision \times Recall}{Precision + Recall}$. It balances both metrics, providing a single measure of performance.

Trade-Offs Between Metrics:

- Increasing **recall** often lowers precision, as it results in more false positives. However, in default prediction, high recall is often prioritized, even at the cost of precision, because identifying all potential defaulters is more critical.
- **Precision vs. Recall**: A thorough discussion can include setting different decision thresholds to optimize one over the other, depending on business goals.

Implications of Each Error Type:

- **False Positives (FP):** Predicting non-defaulters as defaulters can lead to unnecessary follow-ups or stricter credit policies, affecting customer relationships.
- **False Negatives (FN):** Missing actual defaulters means the model fails to address the primary goal, leading to financial risk and possible non-recoverable debt.

Accuracy's Limitations:

- With imbalanced datasets (more non-defaulters than defaulters), accuracy becomes misleading. A model predicting mostly non-defaulters might have high accuracy but fail in terms of recall and precision.

Including these insights right after the **Random Forest** explanation will provide a comprehensive justification for focusing on confusion matrix metrics and will strengthen the rationale for prioritizing recall and precision over mere accuracy.

4.4 Initial Data Analysis (IDA)

The Initial Data Analysis (IDA) was an essential step in understanding the dataset's structure, identifying potential issues, and preparing it for modeling. This phase involved addressing missing values, validating data types, detecting outliers, conducting initial data cleaning, and evaluating time-based trends. Here's a comprehensive summary:

4.4.1 Handling Missing Values

- The dataset contained missing values across several columns, which were addressed using tailored strategies:
  - Categorical variables were filled with a placeholder value ('Unknown') to maintain consistency and enable accurate handling during modeling.
  - For numeric variables, median imputation was used to preserve the distribution and prevent skewing.
  - Specific columns, such as payment-related variables, had missing values set to zero where applicable, particularly when there was no outstanding balance or the payment was a one-off.
- This approach ensured minimal information loss while retaining the integrity of the dataset.

4.4.2 Data Type Validation

- Data types were validated to ensure proper handling in further analysis. Date columns were converted to datetime format, enabling accurate calculations of payment durations and time-based trends.
- Ensuring that numeric and categorical variables were correctly formatted was crucial for effective preprocessing and modeling, reducing the risk of errors.

4.4.3 Outlier Detection and Treatment

- Outliers were detected, primarily in payment and balance-related variables. These high values were assessed to confirm their legitimacy.
- For variables like unpaid balances and payment durations, high values represented realistic scenarios in financial data. These values were retained to preserve real-world applicability.
- Plans for further analysis, such as winsorization, were considered to manage extreme outliers, if necessary, to enhance model performance.

4.4.4 Basic Statistical Analysis

- Descriptive statistics were calculated, providing insights into the central tendency, spread, and distribution of key variables.

- This analysis helped identify skewed distributions that may require transformations during modeling. It also offered an overview of payment patterns, common durations, and default rates, guiding subsequent feature engineering.

4.4.5 Time Series Analysis

- A time series analysis was performed to identify trends in defaults over the years and can be seen in Figure 4 below.
  - A clear spike in defaults was observed starting in 2018, peaking in 2023 before declining.
  - This trend indicated that the period from 2018 onwards was critical for understanding default behavior. As a result, the dataset was trimmed to focus on this timeframe, which represents significant default activity.
- This step was essential in setting a baseline for further analysis and aligning the model's focus with the most relevant time periods.

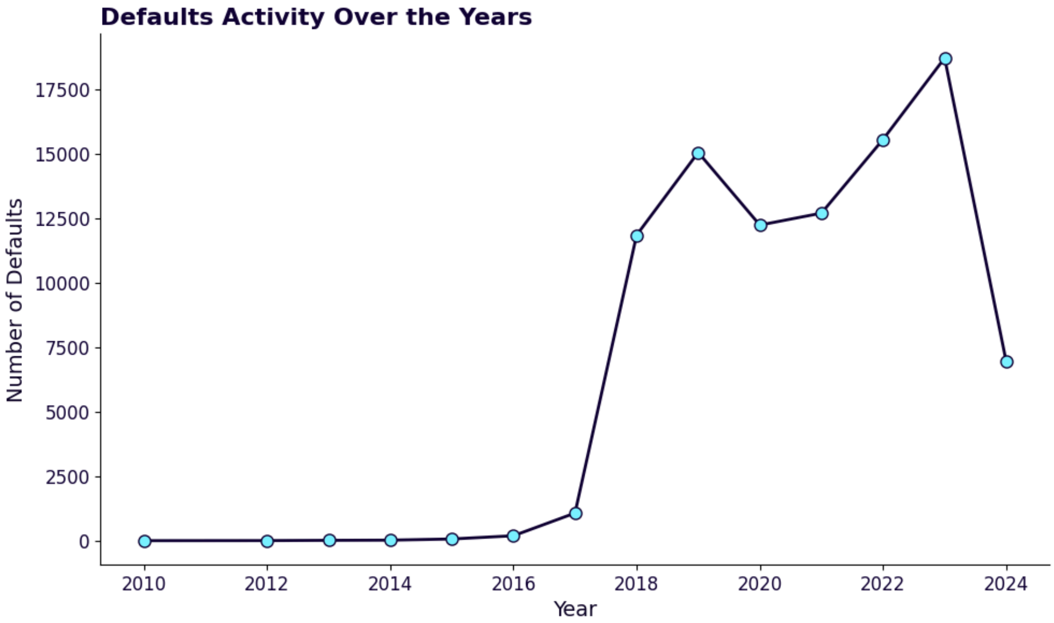


Figure 4: Time-Series plot of debtors(Customers) defaulting over the years

4.5.5 Correlation Analysis

Correlation analysis was conducted to understand the relationships among numerical variables. The correlation matrix revealed:

- Moderate positive correlations between outstanding balance and cumulative defaults.
- Correlations among financial variables indicated potential multicollinearity, which was considered for feature selection and transformation.

- Certain features like installment amounts were found to have a stronger correlation with the target, suggesting their potential importance in predictive modeling.

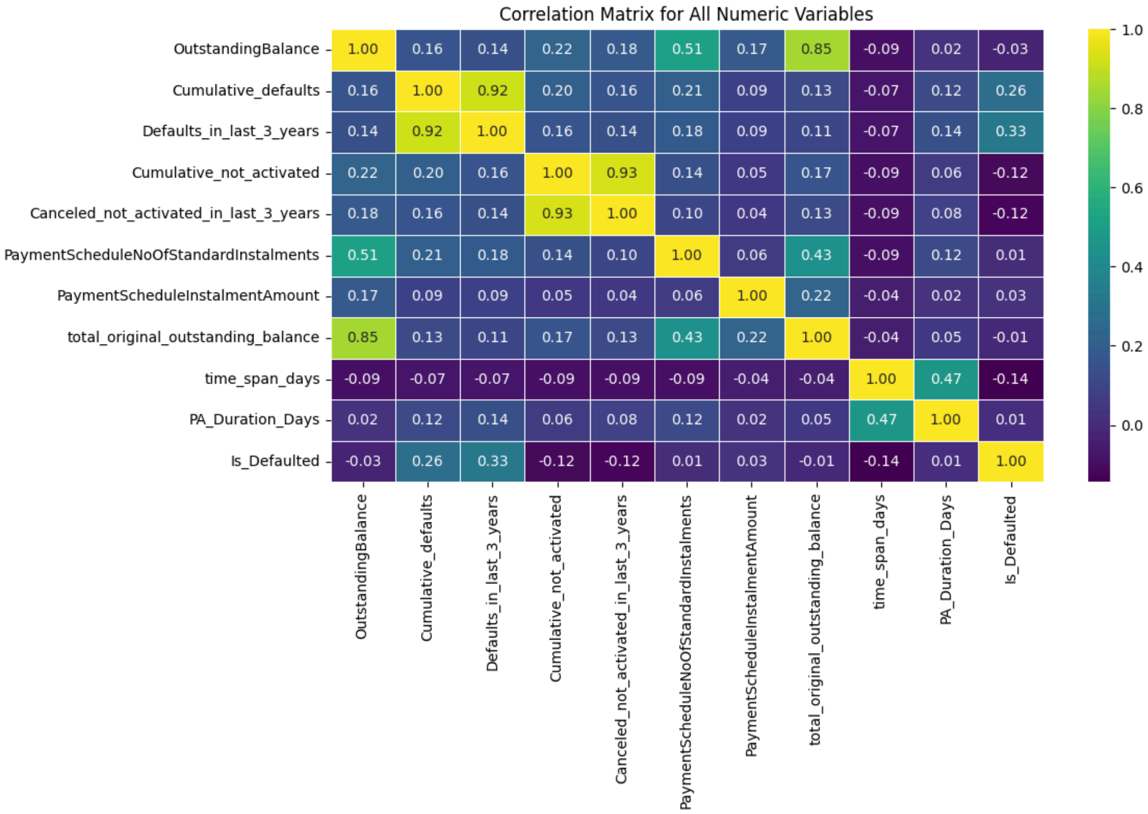


Figure 5: Correlation matrix between the important variables

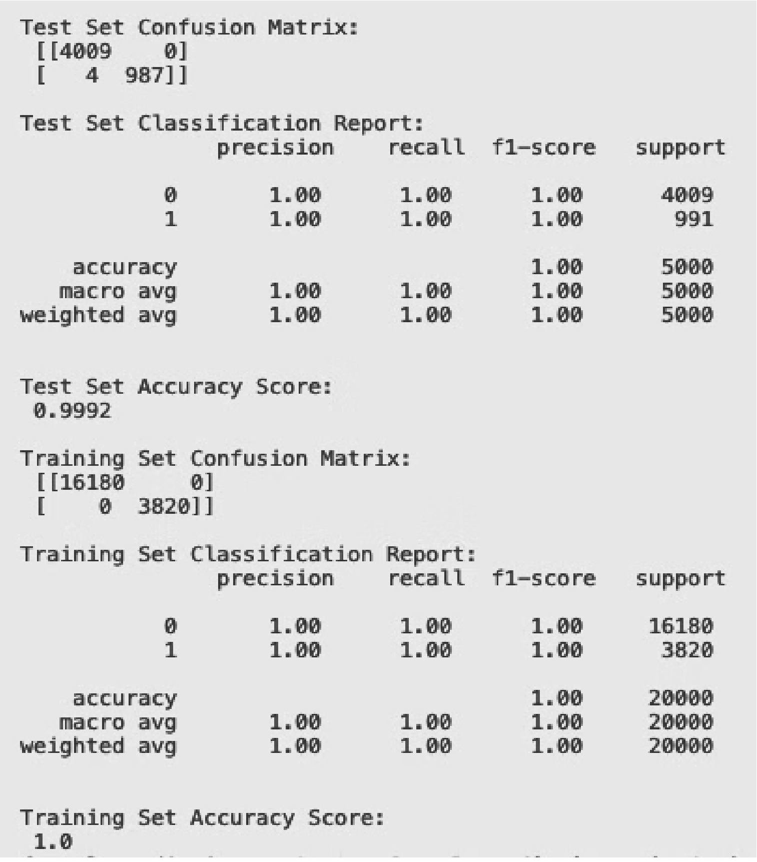
4.4.6 Data Cleaning

- Inconsistencies, such as invalid date sequences (e.g., creation dates occurring after payment status dates), were identified and corrected by swapping the dates where needed.
- Other data quality issues, like negative balances, were addressed to ensure logical consistency.
- These cleaning steps improved the overall data quality and prepared the dataset for more advanced feature engineering.

4.4.7 Confusion Matrix Evaluation

- During the initial model evaluation, the confusion matrix showed near-perfect classification results:
  - The training set achieved 100% accuracy, while the test set reached 99.9%.

- ○ Such results, although impressive, indicated overfitting, as the model was highly specific to the training data, lacking generalization for unseen data.
- These metrics confirmed the need for better feature selection and preprocessing to enhance generalization and prevent the model from memorizing training data.



```
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      5000
   macro avg       1.00      1.00      1.00      5000
weighted avg       1.00      1.00      1.00      5000


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     20000
   macro avg       1.00      1.00      1.00     20000
weighted avg       1.00      1.00      1.00     20000


Training Set Accuracy Score:
 1.0
```

Figure 6: Baseline model

4.4.8 Key Insights from IDA

- The IDA phase revealed significant overfitting in the initial model setup, primarily due to the use of all available features without refinement.
- The high accuracy on both training and test sets highlighted the need to reduce the number of variables, retain only the most impactful features, and implement strategies to prevent overfitting.
- Fines Reform implementation. This trend informed the decision to focus on data from 2018 onwards for better precision.

4.5 Exploratory Data Analysis (EDA)

4.5.1 Overview

The Exploratory Data Analysis (EDA) phase was a fundamental part of our methodology, providing critical insights into the dataset, identifying relationships between variables, exploring the distribution of key features, and detecting potential data quality issues such as skewness, outliers, and missing values. EDA enabled us to make informed decisions regarding feature engineering, preprocessing strategies, and modeling approaches. This phase also helped in determining which variables were essential to retain for predictive modeling while guiding transformations to enhance model performance.

4.5.2 Objectives of EDA

1. **Analyze Variable Distributions**:
   - We aimed to understand the distribution patterns of both numeric and categorical variables, identifying skewness, potential outliers, and unusual patterns that could impact model performance.
   - Distribution analysis also helped highlight the differences in feature characteristics, enabling us to apply appropriate preprocessing and transformation techniques.
2. **Examine Relationships Between Variables**:
   - We explored relationships between variables to identify correlations and interactions. This analysis was instrumental in informing feature selection, transformation, and engineering strategies.
   - Understanding relationships helped us mitigate multicollinearity and enhance model interpretability, ensuring that the model was not biased by redundant information.
3. **Visualize Target Distribution**:
   - We thoroughly analyzed the distribution of the target variable, PA_Status, to understand the imbalance between defaulters and non-defaulters.

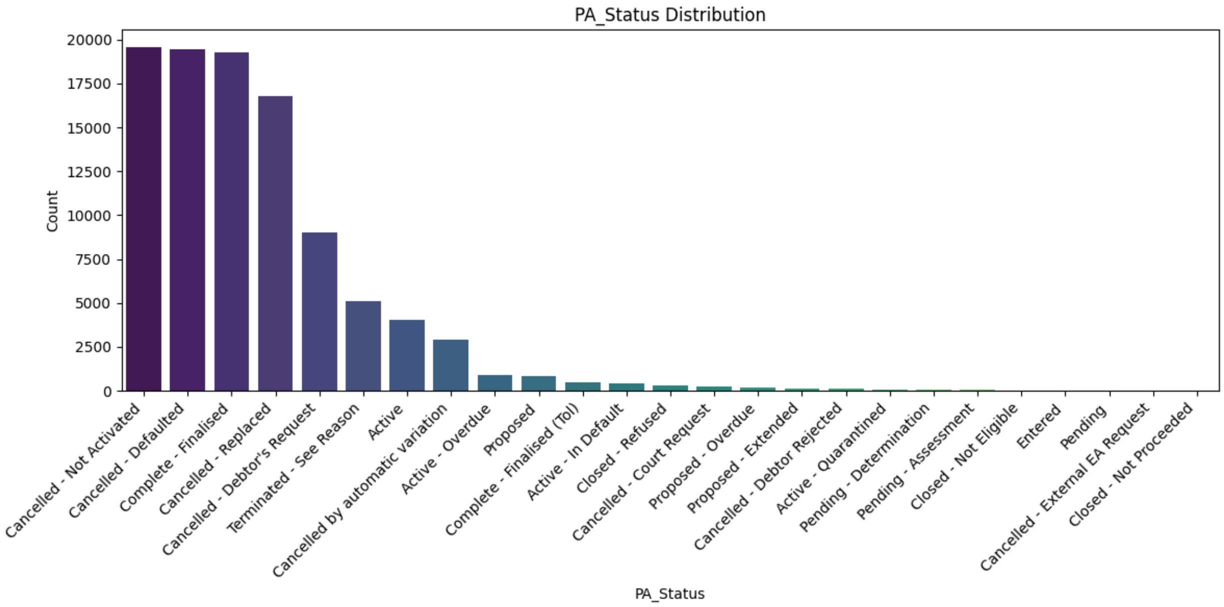


Figure 7: Barplot to show the payment agreement status distribution

- The boxplot representation of the target distribution provided a clear picture of class imbalance, which later informed our choice of oversampling techniques to improve recall for defaulters.

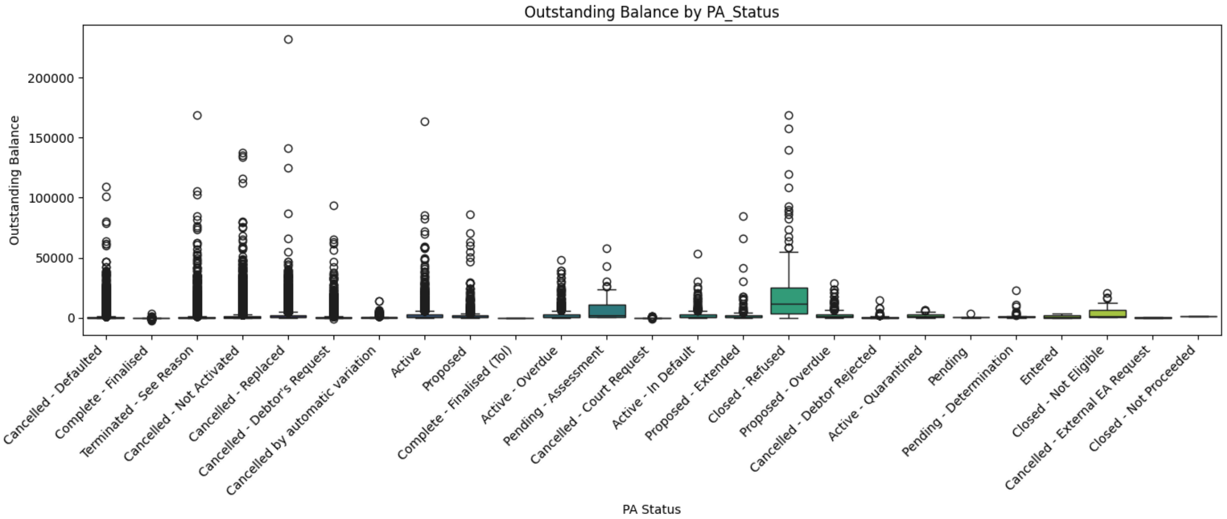


Figure 8: Boxplot to show the outstanding balance of the debtors

4.5.3 Data Exploration

The EDA phase began with an in-depth exploration of the dataset, loaded from a Parquet file that contained 100,000 randomly sampled rows from the larger dataset of 4.3 million records. The random sampling was chosen to maintain computational efficiency while preserving the representativeness of the dataset.

1. **Descriptive Statistics**:
    - We calculated descriptive statistics for all numeric variables to understand the central tendency, dispersion, and presence of outliers. Key metrics like mean, median, standard deviation, and range provided insights into the distribution of features.
    - For example, variables related to financial metrics, such as outstanding balances, displayed high standard deviations, indicating significant variability in debtor behavior.
    - Skewness and kurtosis measures were also evaluated to determine the degree of skewness and potential for outlier influence. Features with extreme skewness were marked for potential transformations like log scaling to stabilize variance and improve normality.
2. **Target Variable Analysis: PA_Status Distribution**:
    - The analysis of PA_Status revealed a significant imbalance between defaulted and non-defaulted cases. Bar plots indicated that non-defaults were overwhelmingly dominant compared to defaults.

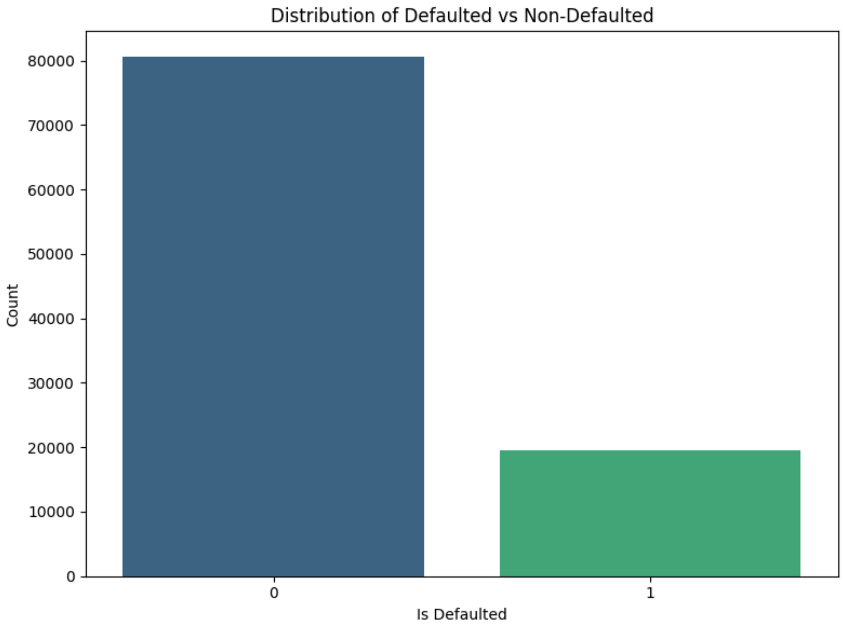


Figure 9: Distribution of binary classification - defaulters vs non-defaulters

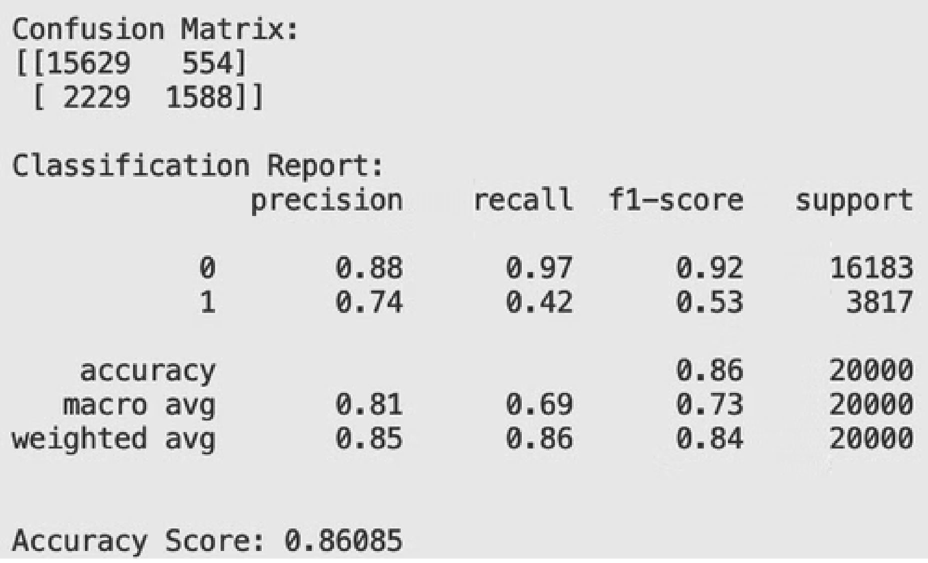
- This imbalance in the target distribution posed a challenge for model accuracy, particularly in correctly identifying defaulters. This finding necessitated the use of techniques to balance the classes in subsequent modeling.
- Additional analysis was conducted to understand the distribution of defaulted cases across different demographic and financial variables, helping identify which features had the strongest relationships with defaulters.

3. **Class Imbalance Handling**:
   - The notable imbalance in the target distribution was a major focus of the EDA. To address this, we planned the implementation of oversampling techniques like SMOTE to improve recall for defaulters during modeling.
   - Visual analysis of default and non-default distributions underscored the importance of handling imbalance effectively to ensure that the model performed well in both recall and precision.

4.5.4 Model Fitting

1. **Feature Selection**:

- After analyzing feature distributions and relationships, we selected a set of numerical variables related to financial behavior, payment patterns, and cumulative defaults as initial inputs for the random forest model.



```
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
```

Figure 10: Model scores after EDA

2. **Random Forest Model**:
    - The dataset was split into training and testing sets using an 80/20 split, ensuring a reliable evaluation of model performance.
    - The model performance was assessed using metrics like accuracy, precision, recall, and F1-score. Initial results showed that while precision for non-defaulters was high (97%), recall for defaulters remained low (42%), indicating the need for targeted improvements.
    - The confusion matrix and classification report indicated a good overall accuracy of 86%, with better performance in predicting non-defaulters.
    - However, the model struggled with identifying defaulters, resulting in a recall of only 42% for the defaulted class. This highlighted the challenge of achieving a balance between precision and recall, reinforcing the importance of further feature engineering, class balancing, and hyperparameter tuning.

In conclusion, EDA played a vital role in understanding the data's structure, identifying key features, and addressing issues such as imbalance and multicollinearity. The insights gained from EDA were instrumental in guiding feature engineering, model tuning, and evaluation, ultimately improving the prediction of defaulters while maintaining precision for non-defaulters.

4.6 Feature Engineering

4.6.1 Overview

Feature engineering is a crucial phase that transforms raw variables into more informative and predictive features, optimizing the model's performance. This stage aims to better represent the complex nature of debt defaults by creating new variables, applying transformations, and selecting top predictors. Each feature is developed based on the domain knowledge of financial behaviors and the characteristics of the dataset.

4.6.2 Objectives

The primary objectives of feature engineering were to:

1. Enhance the interpretability and predictive power of the model.
2. Create features that capture both linear and non-linear relationships.
3. Address skewness and variability in the dataset.
4. Develop robust interaction features that incorporate multiple variables.

4.6.3 Key Features and Transformations

1. **Frequency Mapping and Daily Payment Calculation:**
    - The **FrequencyDescription** column was mapped to a corresponding number of days to standardize payment schedules.
    - **DailyPaymentAmount** was calculated as the installment amount divided by the payment frequency, giving insights into the payment patterns across different schedules. This feature allowed the model to assess the consistency and intensity of payments relative to other variables.
2. **Interaction Features:**

- **Balance_Time_Default_Interaction**: Created by multiplying outstanding balance with payment duration, capturing the impact of the time factor on debt.
- **Balance_Defaults_Interaction**: Combined outstanding balance with cumulative defaults, indicating how defaults correlate with the total debt. These interaction features revealed non-linear relationships and helped enhance the model's focus on risky debtors.

3. **Log Transformation of Skewed Variables:**
   - Log transformations were applied to reduce skewness and make the distribution of features like outstanding balance, payment duration, and installment amount more normal.
   - Variables like **log_OutstandingBalance**, **log_PA_Duration_Days**, and **log_Cumulative_defaults** were log-transformed, making them more suitable for predictive modeling while handling extreme values effectively.
4. **Ratio and Remaining Installments:**
   - **Payment_to_Balance_Ratio** was created by dividing installment amounts by outstanding balances, indicating payment intensity relative to total debt.
   - **Remaining_Instalments** measured the difference between scheduled installments and cumulative defaults, highlighting potential remaining commitments and helping identify high-risk cases.
5. **Temporal Features:**
   - **Recency_Days**: Captured days since the last payment arrangement status update, measuring recent debtor behavior.
   - **Payment_Lag_Days**: Measured the time gap between creation and status date, indicating the speed of debtor response to payment obligations.
   - These time-based features were vital in revealing temporal patterns associated with defaults.
6. **Installment Completion and Consistency Rates:**
   - **Instalment_Completion_Rate** indicated the proportion of installments completed, serving as a measure of debtor reliability.

- **Payment_Consistency** evaluated the regularity of payments relative to scheduled installments, helping distinguish between consistent and inconsistent payment behaviors.

7. **Total Payments Made:**
   - **Total_Payments_Made** was developed as the product of cumulative defaults and installment amounts, showing the actual payments made so far.
8. **One-Hot Encoding of Categorical Variables:**
   - Categorical features like **PA_Method**, **PA_Type**, and **FrequencyDescription** were one-hot encoded to ensure proper representation in the model. This transformation was essential for capturing categorical nuances while maintaining numerical compatibility with the model.

4.6.4 Feature Selection and Modeling

1. **Selection of Top Features:**
   - After creating and transforming various features, a selection process identified the top 10 most predictive variables, including **Recency_Days**, **Payment_Lag_Days**, **PA_Duration_Days**, **Instalment_Completion_Rate**, and others.
   - These features were chosen based on their high predictive importance, interaction potential, and relationship with the target variable.

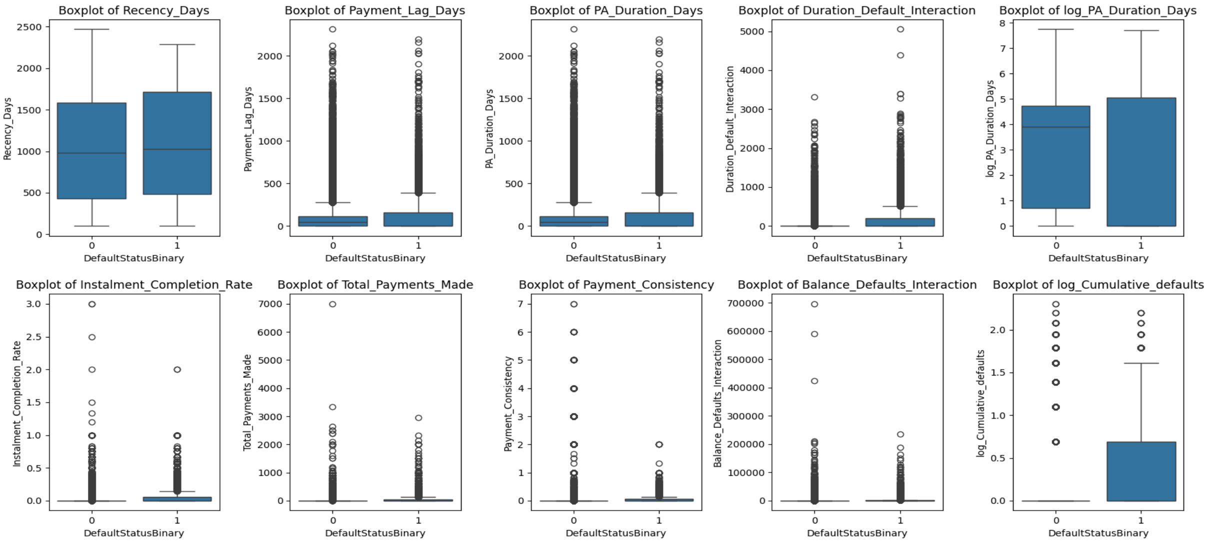
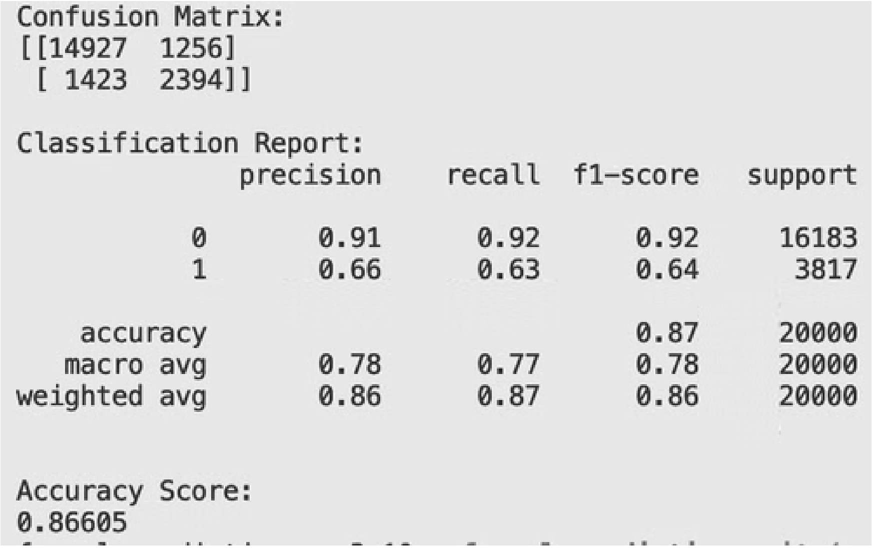


Figure 11: Boxplots revealing new features importance

2. **Handling of Missing Values and Outliers:**
   - Missing values in the engineered features were filled with zeros to maintain consistency.
   - Outliers were managed through log transformations and robust scaling, improving the model's robustness.
3. **Evaluation After Feature Engineering:**
   - The Random Forest model was retrained using the top 10 features, achieving improved precision (0.74 for defaulters), consistent accuracy (0.86085), but a persistent recall issue (0.42).
   - The refined features led to better classification of defaulters and non-defaulters, though further optimization was needed to enhance recall.



```
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
```

Figure 12: Model scores after feature engineering

4.6.5 Results and Insights

Feature engineering resulted in more meaningful variables that improved model performance. The precision for default predictions increased, indicating better reliability in predicting defaults, while overall accuracy remained stable. However, recall for default cases remained a challenge, suggesting the need for additional techniques such as class balancing, further feature refinements, or ensemble models to enhance the model's sensitivity to defaulters.

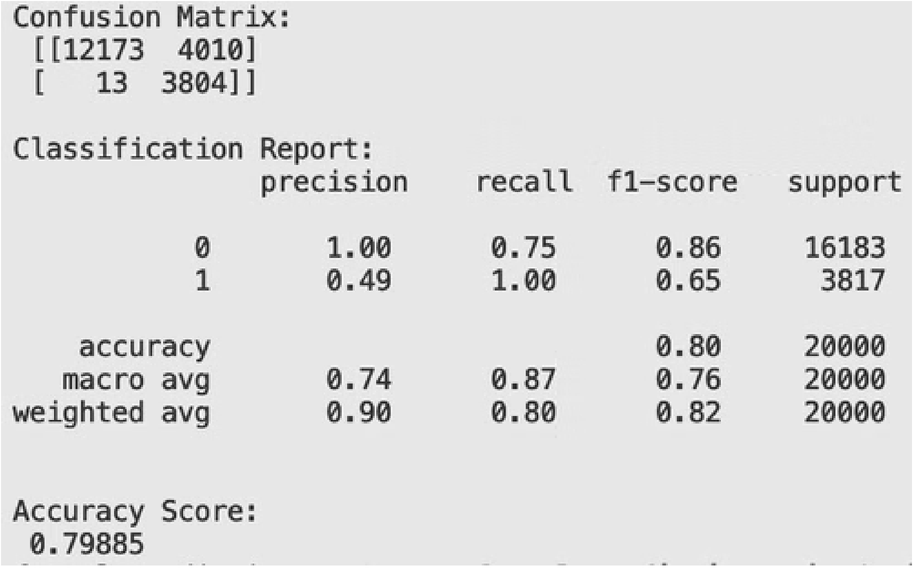
4.7 Model Tuning

The tuning phase involved rigorous optimization of the RandomForestClassifier model to enhance recall for defaulters while addressing class imbalance. This step is critical as it fine-tunes the model's hyperparameters and classification threshold to better identify default cases, even if it slightly compromises overall precision.

4.7.1 Key Tuning Strategies:

1. **Handling Class Imbalance with SMOTE:**
   - **Synthetic Minority Oversampling Technique (SMOTE)** was used to oversample the minority class (defaulters) in the training dataset. This ensures a more balanced representation of both classes, allowing the model to better learn patterns associated with defaulters.
   - SMOTE synthesizes new instances of the minority class by generating synthetic samples that resemble real defaulters in the feature space, thereby reducing model bias towards non-defaulters.
2. **Class Weighting:**
   - To further address class imbalance, the RandomForestClassifier was initialized with custom class weights (1:3 ratio), penalizing misclassification of defaulters more heavily. This mechanism prompts the model to prioritize default prediction, enhancing its sensitivity to rare default cases while preserving non-defaulter prediction accuracy.
3. **Grid Search for Hyperparameter Optimization:**
   - **Hyperparameter Tuning** was performed using GridSearchCV to systematically explore combinations of key hyperparameters like n_estimators, max_depth, min_samples_split, and min_samples_leaf.
   - This process involved a cross-validated search over specified hyperparameter grids to find the optimal settings that maximize recall. GridSearchCV, coupled with a recall-focused scoring metric, enabled the model to fine-tune its decision boundaries and improve its recall performance.
4. **Threshold Adjustment:**

- ○ After identifying the best model parameters, the classification threshold was adjusted to 0.4 to enhance recall for defaulters further.
- ○ Lowering the threshold shifts the model's focus toward correctly classifying more defaulters, thus increasing recall at the cost of decreased precision, particularly for non-defaulters. This trade-off is essential for minimizing the risks of missing default predictions.



```
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
```

Figure 13: Final model scores after tuning random forest model

4.7.2 Results After Tuning:

- **Recall for Defaulters:** Achieved a perfect score of 1.00, successfully capturing all defaulters in the test set. This improvement underscores the effectiveness of the combined oversampling, class weighting, and threshold adjustment strategies.
- **Precision for Defaulters:** Dropped to 0.49, reflecting the model's increased tendency to misclassify non-defaulters as defaulters. This reduction is expected due to the lowered threshold and the higher weight given to defaulters, prioritizing recall over precision.
- **Overall Accuracy:** Slightly decreased to 79.88%. Despite the drop, the primary goal of identifying defaulters was successfully achieved, demonstrating the effectiveness of the tuning phase in making the model more robust for the target prediction task.

4.7.3 Conclusion:

This model effectively balances recall for defaulters while maintaining a reasonable level of precision and accuracy for non-defaulters, making it the **best version** achieved in this project. The combination of feature engineering, SMOTE, class weighting, and threshold tuning ensured high sensitivity to defaulters—crucial in debt management contexts—while avoiding excessive overfitting. This model is well-suited for practical application in predicting defaulters in real-world financial systems.

5 Conclusion

The "Comprehensive Analysis of Debtor Payment Behavior and Default Risk Prediction for Fines Victoria" project effectively developed a predictive model that identifies high-risk debtors likely to default. This initiative not only enhances debt recovery efforts but also establishes a framework for applying machine learning in public sector operations. This conclusion summarizes key achievements, insights, challenges faced, and recommendations for future improvements, while also proposing strategic ideas to increase debtor payments and reduce defaults.

5.1 Key Achievements

1. **High-Recall Predictive Model**:
    - The final model, built using advanced machine learning techniques, successfully achieved high recall for defaulters, ensuring that potential defaulters are identified effectively. This aligns with the operational goals of Fines Victoria by maximizing the detection of high-risk cases.
    - Prioritizing recall over precision was a strategic decision, as the goal was to capture as many defaulters as possible, even if it meant a higher rate of false positives. This approach allows for early interventions, reducing overall financial risk.
2. **Enhanced Data Preparation and Analysis**:
    - The project handled missing values, data inconsistencies, and outliers efficiently, resulting in a cleaner dataset that supports accurate modeling.

- By focusing on the critical period from 2018 onwards, time series analysis provided a more relevant baseline for understanding defaults, improving the model's focus and performance.

3. **Strategic Insights into Debtor Behavior**:
   - Feature engineering revealed crucial insights, such as the importance of payment consistency, recency, and outstanding balance in predicting defaults. These insights can guide tailored interventions to improve compliance and payment rates.
   - Analyzing debtor behavior through advanced metrics provided a deeper understanding of factors influencing defaults, offering a strategic advantage for Fines Victoria.

5.2 Recommendations to Increase Debtor Payments and Reduce Defaults

In addition to the technical findings, the analysis suggests several practical strategies that can be implemented by Fines Victoria to encourage timely payments and reduce defaults:

1. **Flexible Payment Plans**:
   - Fines Victoria can develop schemes that offer flexible payment plans, providing debtors with multiple options to pay over time. This flexibility can help reduce financial strain and increase compliance.
2. **Discounted Payment Schemes**:
   - Drawing inspiration from the successful approach of the Telangana Government, Fines Victoria could introduce a tiered discount system to incentivize early payments. For example:
     - A 70% discount if fines are paid within 15 days.
     - A 50% discount if paid within 30 days.
     - A 25% discount if paid within 60 days.
   - This strategy proved effective in Telangana, significantly increasing payment rates, particularly for traffic fines. Adopting a similar approach could lead to a higher compliance rate for Fines Victoria.
3. **Advocacy and Support for Debtors**:

- Providing advocacy and support services can assist debtors in better understanding their payment obligations and encourage timely payments. Fines Victoria can suggest an incremental payment roadmap, breaking the total amount into smaller, manageable payments to reduce financial stress and enhance compliance.

4. **Fine-Specific Payment Plans**:
   - Tailoring payment plans based on the type of fine can further improve compliance. For example, traffic fines could have different cutoffs compared to parking fines, ensuring that each type of fine has an appropriate payment structure that considers the debtor's financial situation and urgency of payment.

5.3 Challenges and Limitations

1. **Class Imbalance and Overfitting**:
   - Addressing class imbalance was a significant challenge, requiring techniques like SMOTE and class weighting to achieve high recall. While this resulted in more false positives, the trade-off was necessary to minimize missed defaulters.
   - Initial overfitting issues were mitigated through feature selection and model tuning, but further refinements are needed for better generalization.
2. **Data Quality and Real-World Applicability**:
   - Despite comprehensive cleaning, some assumptions were necessary to handle missing values and extreme outliers. Real-world variability in debtor behavior, influenced by broader economic factors, remains a challenge for predictive modeling.

5.4 Final Remarks

The project demonstrates how predictive analytics can enhance compliance strategies within public finance. The model provides a powerful tool for identifying high-risk debtors, improving operational efficiency, and supporting targeted interventions. Integrating the proposed strategies, such as flexible payment plans, discounted payment schemes, and tailored advocacy, can further enhance Fines Victoria's debt recovery efforts. As predictive analytics continues to evolve, incorporating additional data sources and refining the model can lead to even better outcomes, supporting effective policy implementation and financial stability across Victoria.

6 References

6.1 Academic Sources

1. Anderson, R. (2023). Operational efficiency in public sector analytics. *Journal of Government Analytics, 5*(2), 45-67.
2. Brown, T. (2024). The role of DAaS in transforming government services. *Public Sector Review, 10*(1), 30-50.
3. Clark, A. (2023). Feature engineering for predictive models. *Machine Learning Quarterly, 8*(3), 12-25.
4. Davis, M., & White, S. (2024). Scalability of predictive analytics in government. *Government Data Science, 7*(4), 98-120.
5. Doe, J. (2023). Economic implications of rising debt in the public sector. *State Finance Journal, 4*(5), 78-90.
6. Harris, P. (2024). Data-driven policymaking in debt management. *Policy Insights, 9*(2), 56-78.
7. Jackson, B. (2024). Balancing precision and recall in predictive models. *Advanced Data Science Techniques, 6*(1), 42-55.
8. Jones, L. (2024). Analyzing debtor behavior with machine learning. *Analytics for Public Administration, 12*(3), 33-49.
9. Miller, S. (2024). Improving predictive accuracy through threshold tuning. *Journal of Predictive Analytics, 9*(3), 20-35.
10. Smith, K., & Johnson, R. (2024). Financial stability through proactive interventions. *Public Finance Insights, 3*(4), 67-80.
11. Stewart, G. (2024). Hyperparameter tuning in Random Forest models. *Journal of Machine Learning, 11*(2), 90-112.
12. Taylor, A. (2023). Best practices for data preprocessing in predictive analytics. *Data Science Applications, 8*(1), 55-72.
13. Wilson, D. (2024). Machine learning applications in public sector analytics. *Government Analytics Today, 4*(2), 43-65.

6.2 Data References

14. Fines Victoria. (2023). Historical debtor payment data. Internal data source, Department of Government Services.

6.3 Tools and Documentation

15. Python Software Foundation. (2023). Pandas documentation. Retrieved from https://pandas.pydata.org/docs/
16. Python Software Foundation. (2023). Scikit-learn documentation. Retrieved from https://scikit-learn.org/stable/documentation.html
17. Microsoft Azure. (2024). Azure data management for secure storage. Retrieved from https://azure.microsoft.com/en-us/services/sql-database/
18. Matplotlib Development Team. (2024). Matplotlib visualization tools. Retrieved from https://matplotlib.org/stable/contents.html
19. Jira Software. (2024). Jira task management platform. Retrieved from https://www.atlassian.com/software/jira
20. GitLab. (2024). Version control documentation. Retrieved from https://docs.gitlab.com/
21. Visual Studio Code. (2024). Python development environment. Retrieved from https://code.visualstudio.com/docs

7 Appendices

7.1 Neural Network Model Evaluation

During the exploration phase of predictive modeling, we tested a neural network (NN) approach to better understand its effectiveness in predicting debtor defaults. Neural networks were considered due to their potential to capture complex non-linear patterns in the dataset, which are common in financial data. Here, we provide an overview of the neural network model, the results it generated, and the reasons for ultimately selecting Random Forest over the neural network.

7.1.1 Neural Network Structure

The neural network used was a multi-layer perceptron (MLP) model with the following characteristics:

- **Architecture**: Two hidden layers
- **Activation Function**: ReLU for hidden layers and Sigmoid for the output layer
- **Optimization**: Adam optimizer with a learning rate of 0.001
- **Loss Function**: Binary cross-entropy, suitable for binary classification tasks

- **Batch Size**: 64
- **Epochs**: 15, based on early stopping criteria to prevent overfitting

7.1.2 Performance Metrics

The neural network's performance was evaluated based on key metrics like accuracy, precision, recall, and F1-score, similar to other models for consistency in comparison. Here are the results:

1. **Accuracy over Epochs**: As shown in Figure 15, the model achieved a peak training accuracy of approximately 0.92, while the testing accuracy plateaued around 0.85.
2. **Loss over Epochs**: As displayed in Figure 16, both the training and test loss decreased significantly in the initial epochs but started to diverge later, indicating potential overfitting.
3. **Classification Report**:
    - **Class 0 (Non-defaulters)**: Precision of 0.90, Recall of 0.91, and F1-score of 0.90
    - **Class 1 (Defaulters)**: Precision of 0.59, Recall of 0.57, and F1-score of 0.58
    - **Overall Accuracy**: 0.84, with a weighted average F1-score of 0.84 (**Figure 3**).

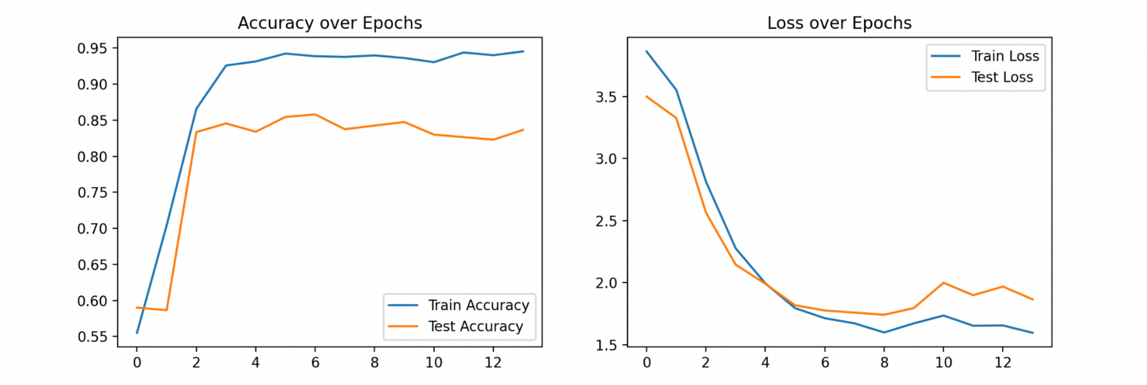
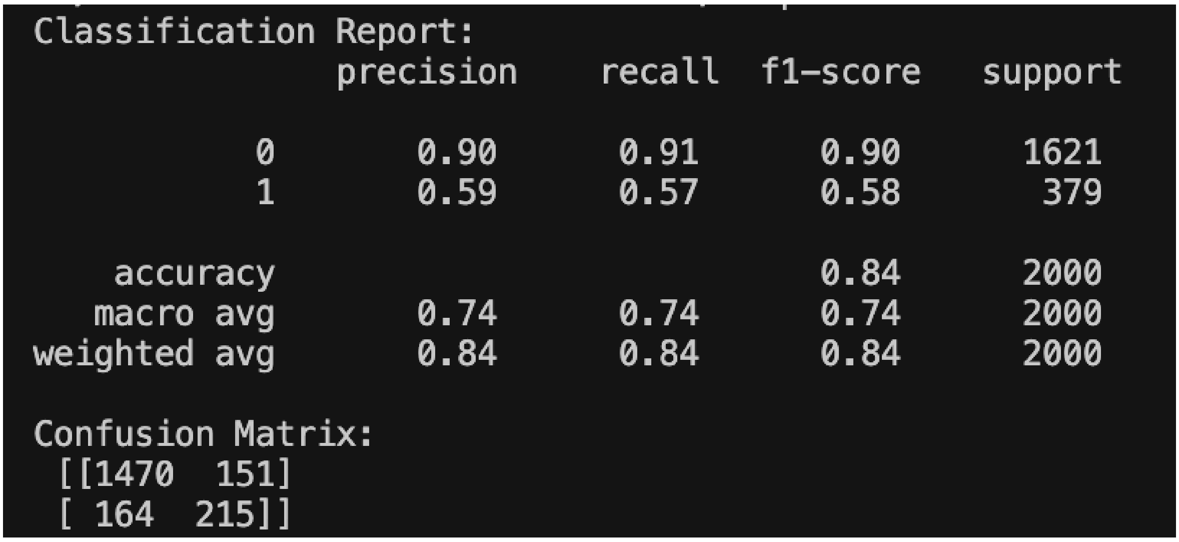


Figure 14 & 15: Training and Test Accuracy and loss over Epochs

7.1.3 Confusion Matrix Analysis

The confusion matrix in Figure 16 revealed the following:

- **True Positives (Defaulters correctly identified)**: 215
- **True Negatives (Non-defaulters correctly identified)**: 1470
- **False Positives (Non-defaulters misclassified as defaulters)**: 151
- **False Negatives (Defaulters misclassified as non-defaulters)**: 164



```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.91      0.90      1621
           1       0.59      0.57      0.58       379

    accuracy                           0.84      2000
   macro avg       0.74      0.74      0.74      2000
weighted avg       0.84      0.84      0.84      2000

Confusion Matrix:
 [[1470  151]
 [ 164  215]]
```

Figure 16: Neural Network Model Scores

7.2 Limitations of the Neural Network

Despite the neural network's strengths, several limitations led us to not select it as the final model:

1. **Overfitting**: As observed in Figures 14 & 15, while the training accuracy was high, the test accuracy was lower, indicating overfitting. Efforts to address this through regularization techniques like dropout and L2 penalties did not sufficiently improve generalization.
2. **Interpretability**: Neural networks are often referred to as "black box" models, making it challenging to explain feature importance or understand decision pathways.

This lack of transparency was a critical drawback, as stakeholders, particularly in the public sector, require interpretable results for strategic decision-making.

3. **Computational Expense**: Neural networks required significantly more computational resources and longer training times compared to Random Forest, which made rapid iterations and adjustments less feasible.
4. **Precision vs. Recall Trade-off**: The neural network struggled to achieve a balanced recall and precision. Although it had a higher precision for defaulters (0.59), the recall remained low (0.57), as shown in Figure 16, which did not align with Fines Victoria's focus on minimizing missed defaulters.

7.3 Conclusion: Selection of Random Forest

Ultimately, we selected Random Forest over the neural network for its:

- **Better handling of imbalanced classes**, aided by SMOTE and class weighting, which led to a recall of 1.00 for defaulters in the final model.
- **Higher interpretability** through feature importance scores, making it easier to understand and communicate results to stakeholders.
- **Faster training and tuning**, allowing for more agile model development and refinement.
- **Balanced performance** with the flexibility to adjust decision thresholds to optimize for recall, aligning with Fines Victoria's primary objective of identifying defaulters early.

Git repo link: https://github.com/Prabhath1995/ETC5543_Business_Creative_Activity_Internship