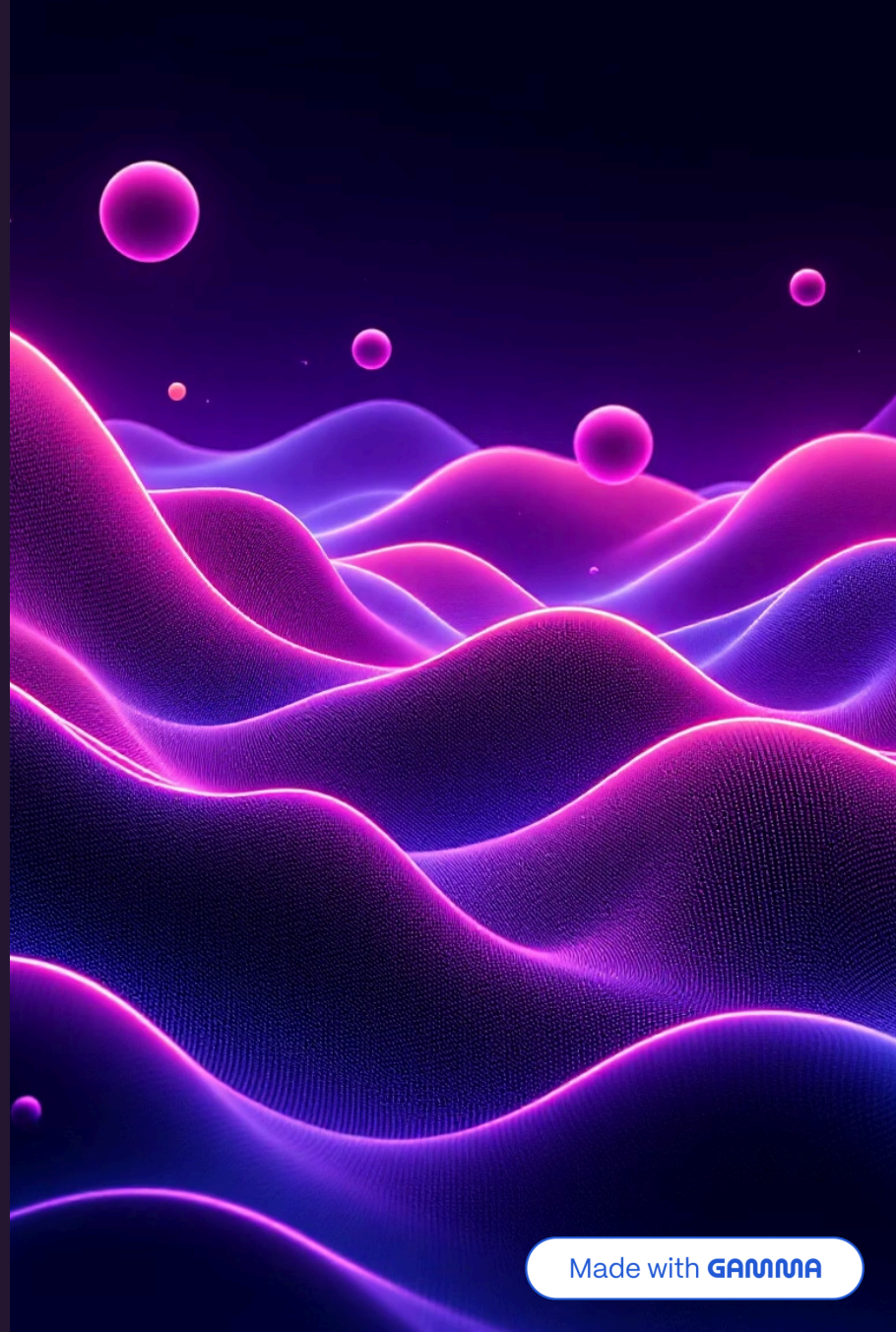
Credit Risk Prediction

Understanding, Predicting, and Mitigating Default Risk

Your Name: Andrei Bitca

Course / Program: Data Analytics with Artificial Intelligence



Credit Risk: Understanding the Challenge

What is Credit Risk?

- Credit risk refers to the possibility of a borrower failing to meet their debt obligations (like credit card payments).
- For financial institutions, this translates directly into potential financial losses.

Why is Predicting Credit Risk Important?

- **Mitigating Financial Losses:** Accurately identifying clients at high risk of default allows institutions to take proactive measures, minimizing financial write-offs.
- **Informing Lending Decisions:** Provides data-driven insights to make smarter choices about offering credit, setting limits, and determining interest rates.
- **Optimizing Customer Strategies:** Enables tailored approaches for different risk segments, from offering support to high-risk individuals to nurturing loyal, low-risk clients.

Our Project's Goal:

- To leverage historical data to analyse and identify key factors that drive credit card default.
- To build and evaluate predictive models that can forecast a client's likelihood of defaulting on future payments.

Business Questions & Goals

1 **Which features most influence credit default?**

- *Why this matters:* Understanding the most impactful factors allows financial institutions to pinpoint core risks and focus their efforts.

2 **What customer groups are most at risk?**

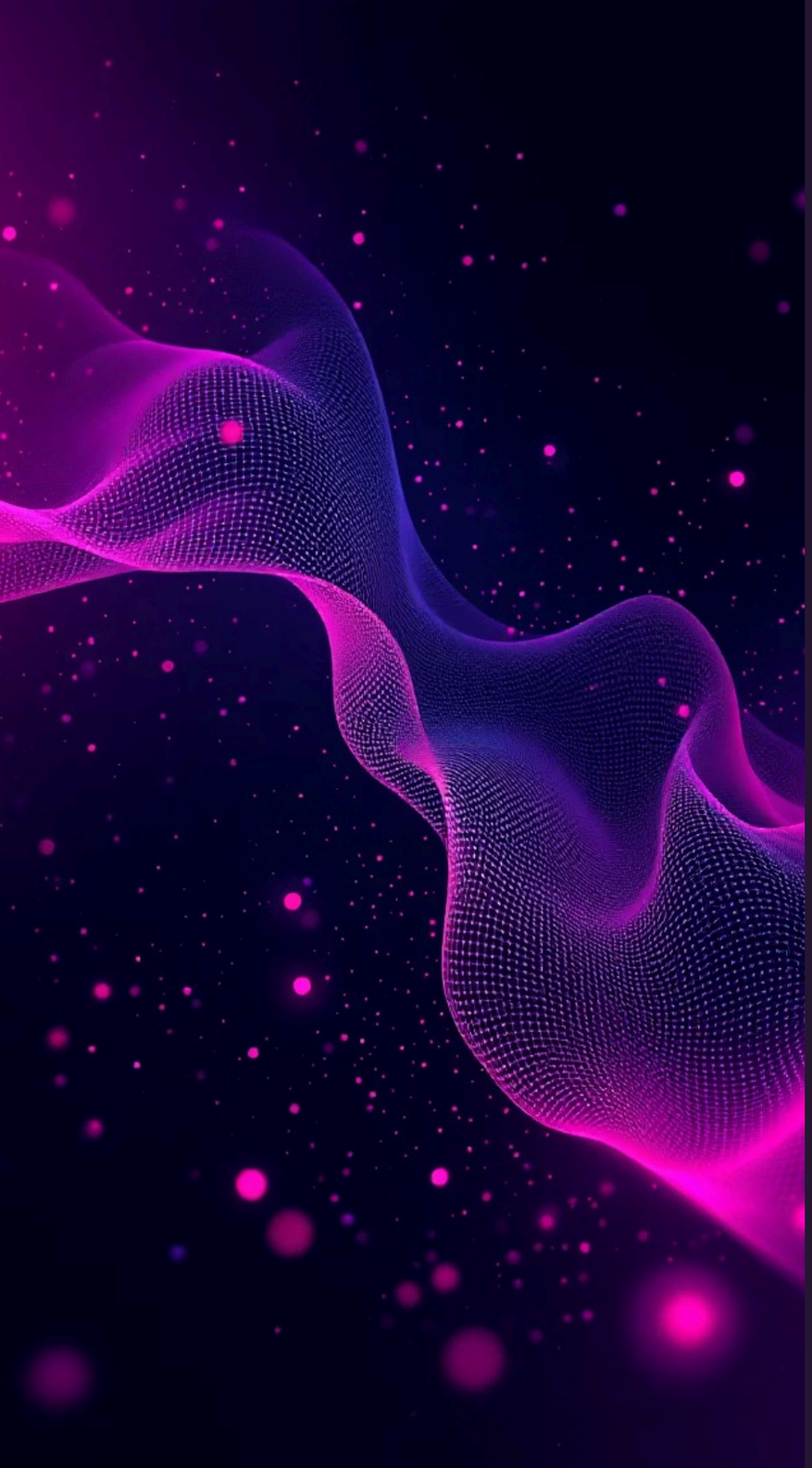
- *Why this matters:* Identifying specific segments (e.g., by demographics or financial behaviour) enables targeted risk management and tailored customer strategies.

3 **Can we predict default risk with a model?**

- *Why this matters:* Building reliable predictive models provides a powerful tool for proactive risk assessment, helping to minimize future losses.

4 **Overall Business Goals:**

- To equip financial institutions with **data-driven insights** to minimize financial losses associated with credit card defaults.
- To enable **more informed and precise lending decisions**, fostering healthier financial portfolios.
- To support the development of **proactive customer management strategies**, potentially reducing defaults before they occur.



Dataset Overview

Source:

We utilized a publicly available dataset titled "Default of Credit Card Clients Dataset", sourced from the **UCI Machine Learning Repository**.

Dataset Size:

This robust dataset comprises detailed information on **30,000 credit card clients**.

What the Data Contains:

For each client, the dataset provides a comprehensive view:

- **Demographic Information:** Gender, Education Level, Marital Status, and Age.
- **Financial Information:** Their Credit Limit, detailed Bill Amounts for the past six months, and Payment Amounts made over the same period.
- **Payment History:** Records indicating any Payment Delays for each of the past six months.
- **Key Outcome (Target Variable):** Whether the client ultimately **defaulted** on their next payment.

Approach & Tools

Data Preparation

This initial phase involved cleaning, transforming, and organizing the raw dataset to ensure its quality and readiness for analysis. We handled missing values and prepared features.

Exploratory Data Analysis (EDA)

We delved into the data to uncover initial patterns, distributions, and relationships between various factors and credit default. This helped us understand the data's story.

Feature Engineering

We created new, more insightful variables from existing data (e.g., `Total Bill Amount`, `Payment Delay Count`) to improve the predictive power of our models.

Statistical Analysis

We performed formal statistical tests to validate our hypotheses about the relationships between different data points and credit default.

Machine Learning Modelling

We built and evaluated various predictive models to forecast the likelihood of a client defaulting.

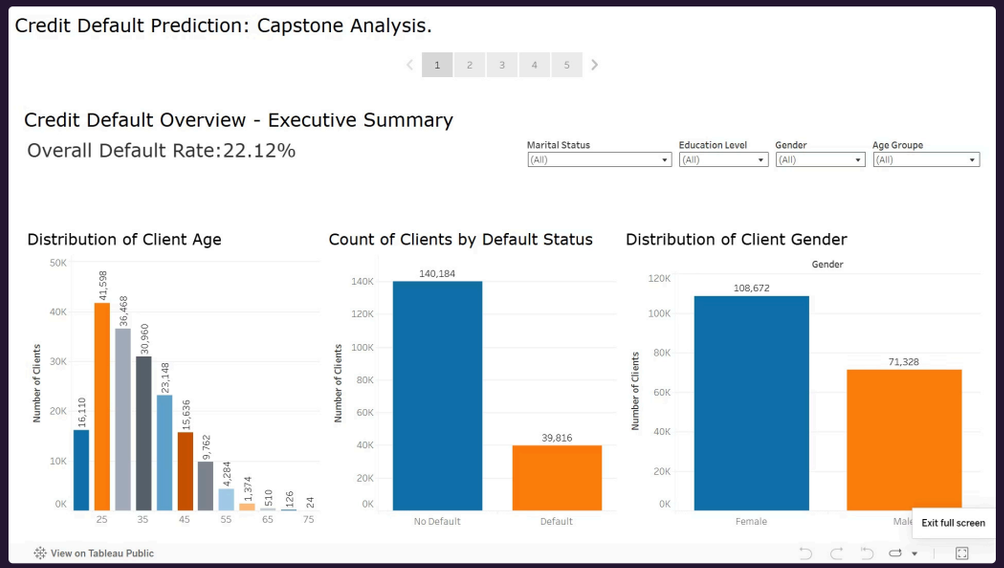
Dashboarding & Communication

The final step was to create interactive dashboards to visually present our findings and make the insights easily accessible.

Key Tools Utilized:

- **Python:** Our primary tool for all data manipulation, in-depth analysis, statistical testing, and the development of machine learning models. We specifically used libraries like `pandas`, `numpy`, `scikit-learn`, `matplotlib`, and `seaborn`.
- **Tableau:** Used for designing and building interactive, user-friendly dashboards that effectively communicate our key findings and model performance.

Key Findings: The Overall Picture of Default



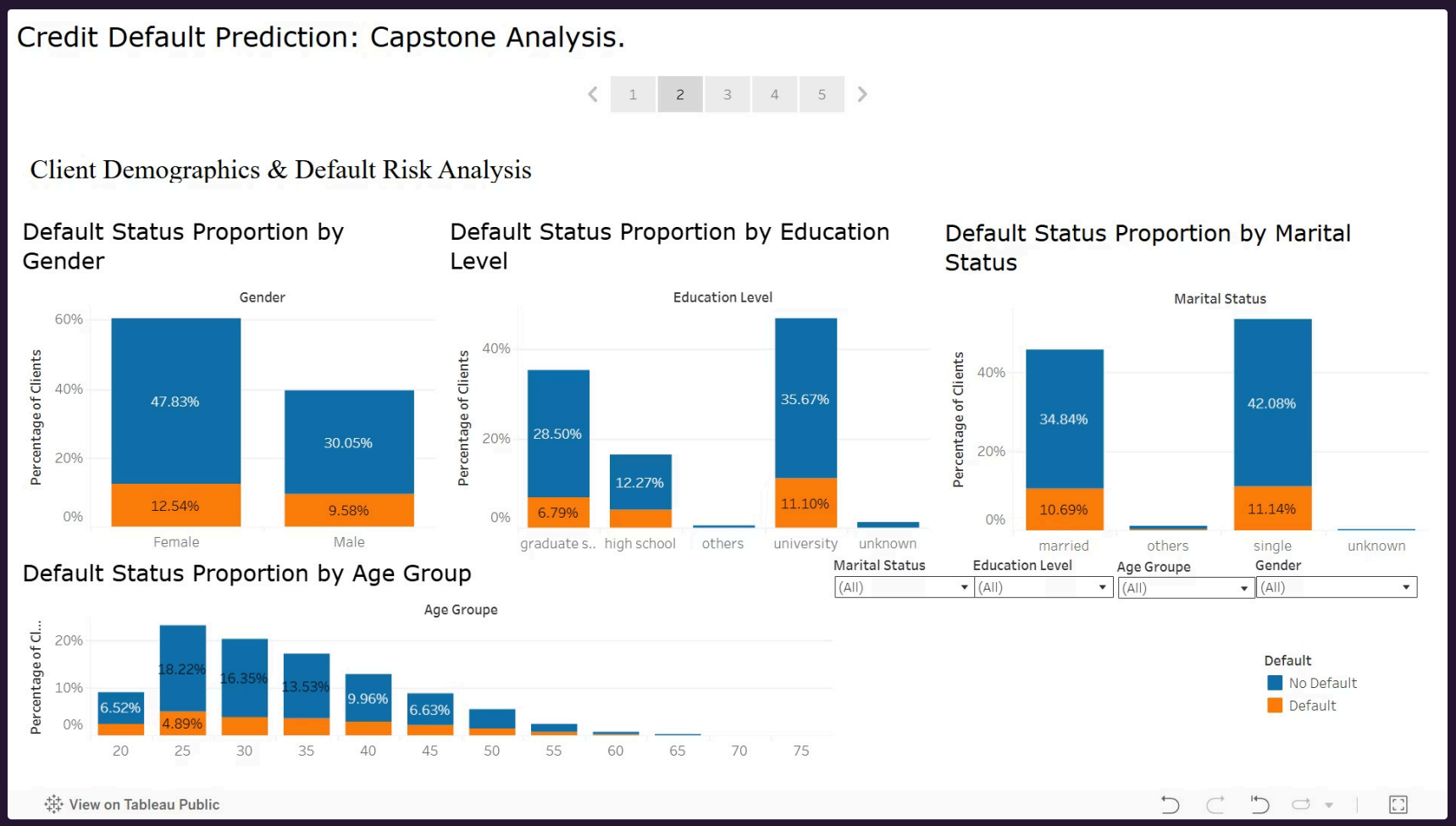
The Overall Default Rate:

- Our analysis revealed an overall credit default rate of **22.12%** among the clients in this dataset.
- This indicates that approximately **1 in 5 clients** did not meet their payment obligations, highlighting a significant area of risk.
- Specifically, out of 179,996 clients, **39,816 defaulted** while **140,184 did not**.

General Client Demographics:

- **Gender Distribution:** The dataset shows a distribution of **108,672 Female** clients and **71,328 Male** clients.
- **Age Distribution:** The largest groups of clients fall within the **25-35 year range**, with a significant number in their **20s**. For instance, there are 41,598 clients aged 25 and 36,468 clients aged 30.
- *Note:* While these overall distributions are important for context, we will delve into how default *rates* vary by these demographics on the next slide.

Key Findings: Demographics and Default Propensity



Gender:

- Our analysis indicates that **Male** clients have a slightly **higher default rate (24.17%)** compared to **Female** clients (20.77%).
- *(While females represent a larger percentage of overall clients, males show a higher propensity to default within their group.)*

Education Level:

- Education appears to be a notable factor. Clients with **High School** education show the highest default rate at **24.39%**.
- This is closely followed by clients with **Graduate School** education at **23.73%**.
- Clients with **University** education had a lower default rate of **19.24%**.

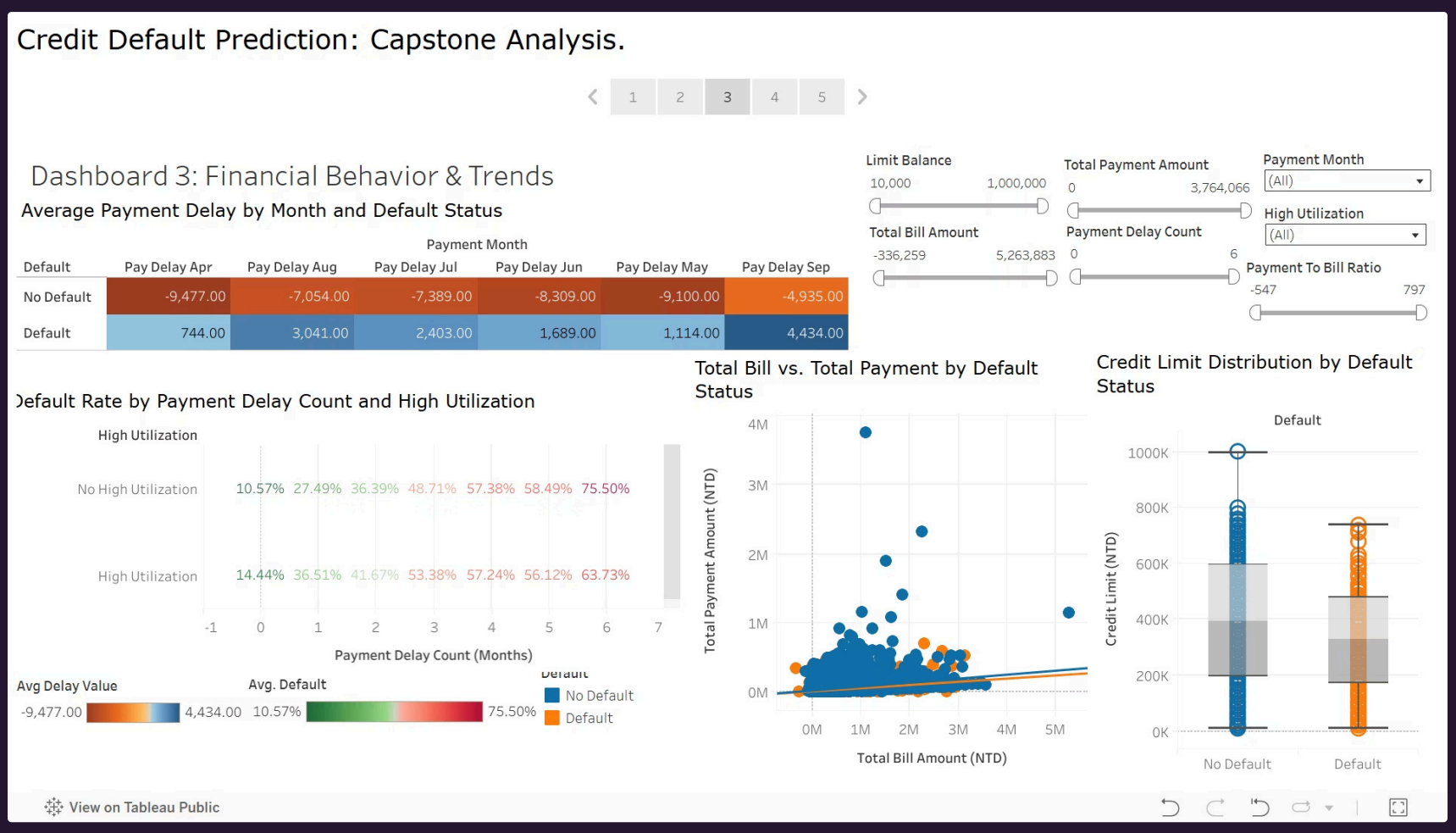
Marital Status:

Married clients exhibit a slightly **higher default rate of 23.48%** compared to **single** clients at **20.93%**.

Age Groups:

- The **youngest age group (20 years old)** shows the highest default rate at **26.35%**.
- Generally, the default rate tends to **decrease as age increases**, with the lowest rates observed in the older age groups (e.g., 55 years old at 14.91%).

Key Findings: Financial Behaviour - The Strongest Predictors



Payment Delays are a Major Indicator:	Impact of High Utilization:	Bill vs. Payment Patterns:	Credit Limit Distribution:
• Our analysis clearly shows a stark contrast: **non-defaulting clients** typically have **negative average payment delays** (meaning they pay on or before the due date, e.g., an average of -9,477 for April payments). • In contrast, **defaulting clients** consistently show **positive average payment delays** (meaning they pay after the due date, e.g., an average of 744 for April payments), indicating a pattern of late payments. • **The default rate drastically escalates with the number of payment delays.** For instance, clients with **5 payment delays** show a staggering default rate of **75.50%** (for those with no high utilization). Even with `High Utilization`, 5 payment delays still result in a 63.73% default rate.	When combined with payment delays, **High Utilization** (meaning a high percentage of their credit limit is being used) consistently **increases the default rate** across most payment delay counts. For example, at 0 payment delays, the default rate jumps from 27.49% (no high utilization) to 36.51% (high utilization).	• The scatter plot visually confirms that **defaulters (orange dots)** tend to have **higher total bill amounts relative to their total payments**, suggesting they are struggling to keep up with their balances. • **Non-defaulters (blue dots)**, on the other hand, typically show a healthier balance between bills and payments, or lower overall bill amounts.	Clients who defaulted generally possess **lower credit limits** compared to non-defaulting clients. This could indicate that those with lower initial creditworthiness are more susceptible to default, or that they are granted lower limits due to existing risk factors.

Building Predictors: Our Machine Learning Approach

What is Predictive Modelling?	Our Model Selection Strategy
• Predictive modelling uses algorithms to analyse historical data and identify patterns. • These patterns are then used to forecast future outcomes – in our case, whether a credit card client is likely to default.	• We focused on establishing a strong baseline using commonly recognized and interpretable classification algorithms. • These models help classify clients into "Default" or "No Default" categories.

Models We Implemented and Evaluated:

Logistic Regression:	Decision Tree:	K-Nearest Neighbors (KNN):
• A statistical model used for predicting binary outcomes (like default/no default). • It's known for its interpretability, showing how each factor contributes to the likelihood of default.	• A tree-like model that makes decisions by splitting data based on various features. • It mimics human decision-making and can easily show the rules leading to a prediction.	• A non-parametric model that classifies a data point based on how its "neighbors" (similar data points) are classified. • It's effective for complex decision boundaries but can be computationally intensive.

Predictive Power: Model Performance & Insights

Understanding Our Key Metrics:

1 Accuracy:

Overall correctness of the model's predictions (both defaulters and non-defaulters).

2 Precision (Default):

Of all clients predicted to default, how many actually did? (Minimizes false positives - predicting default when they don't)

3 Recall (Default):

Of all clients who *actually* defaulted, how many did the model correctly identify? (Minimizes false negatives - missing actual defaulters). This is often the **most critical metric for credit risk** as it directly impacts loss prevention.

4 F1-Score (Default):

A balance between Precision and Recall.

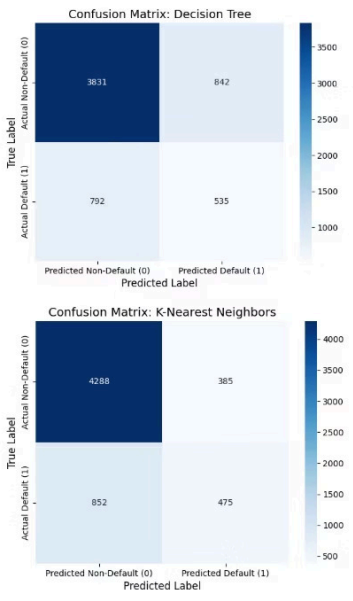
Credit Default Prediction: Capstone Analysis.

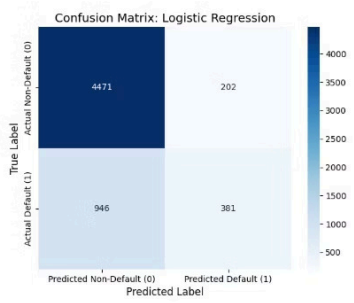
1 2 3 4 5

Model Performance & Key Insights

Baseline Model Performance Metrics

Model Name				
Decision Tree	72.77%	39.57%	38.85%	40.32%
K-Nearest Neighbors	79.38%	43.44%	55.23%	35.80%
Logistic Regression	80.87%	39.90%	65.35%	28.71%





View on Tableau Public

Overall Accuracy:

The **Logistic Regression** model achieved the highest overall accuracy at **80.87%**.

Identifying Defaulters (Recall):

For credit risk, our primary goal is to **minimize missed defaulters**. The **Decision Tree** model demonstrated the strongest ability to correctly identify actual defaulting clients, with a **Recall of 40.32%**. This means it was more successful at catching potential losses, even if its overall precision was lower.

Precision vs. Recall Trade-off:

While Logistic Regression had higher precision (fewer false alarms), its lower recall suggests it missed more true defaulters. The choice of model depends on the business's tolerance for false positives versus false negatives.

Translating Insights into Action: Recommendations

Proactive Intervention for High-Risk Accounts:

- **Focus on Payment Behaviour:** Actively monitor for clients showing **consistent payment delays** (e.g., 2 or more months past due) and those with **high credit utilization**. Our analysis shows default rates significantly escalate in these groups (e.g., up to 75.50% for clients with 5 delays).
- **Early Outreach:** Implement automated alerts or human outreach programs (e.g., financial counselling, flexible payment plans) for clients showing these early warning signs, *before* they default.

Refined Credit Assessment & Limit Setting:

- **Integrate Behavioural Data:** Incorporate historical payment behaviour and utilization patterns more heavily into credit scoring models for new and existing clients.
- **Adjust Credit Limits:** Based on predictive model scores, adjust credit limits to align with a client's risk profile, potentially reducing exposure for high-risk individuals.
- **Consider Demographic Nuances:** While financial behaviour is primary, be mindful of demographic insights, such as the slightly higher default rate among **males (24.17%)** and clients with **High School education (24.39%)**.

Optimized Resource Allocation based on Model Strengths:

- **Leverage Recall for Loss Prevention:** Utilize models with higher **Recall for Default** (e.g., **Decision Tree at 40.32%**) to prioritize accounts for review. While this might lead to more false positives, it ensures fewer actual defaulters are missed, directly reducing potential losses.
- **Targeted Marketing/Support:** Use model predictions to segment customers for tailored offers (e.g., lower interest rates for low-risk clients) or specific financial support programs.

Continuous Monitoring & Model Improvement:

- Regularly monitor the performance of deployed models with new data to ensure their accuracy remains high.
- Periodically retrain models and explore more advanced algorithms to capture evolving risk patterns.

Navigating Challenges & Paving Future Paths

Challenges Encountered

Data Imbalance: Dealing with the imbalance between defaulting and non-defaulting clients was a key challenge. A significantly smaller proportion of clients defaulted, which can make it harder for models to learn effectively.

Feature Engineering Complexity: Identifying and creating truly impactful new features from the raw transactional data required careful thought and iterative experimentation.

Model Interpretability vs. Performance: Balancing the desire for highly accurate models with the need for clear interpretability (especially for business stakeholders) was a consideration.

Future Enhancements & Roadmap

1. **Explore Advanced Models:** Investigate more sophisticated machine learning algorithms like Gradient Boosting (e.g., XGBoost, LightGBM) or Neural Networks, which often yield higher predictive performance.

2. **Deepen Feature Engineering:** Analyse payment history data at a more granular, time-series level to capture dynamic behavioural shifts.

   Incorporate external data sources, such as economic indicators (e.g., unemployment rates, inflation), which could influence credit risk.

3. **Advanced Sampling:** Apply more advanced techniques to handle data imbalance (e.g., SMOTE, ADASYN) to potentially improve model performance, especially Recall.

4. **Automated Monitoring:** Develop a system for continuous monitoring of model performance in a production environment and automated retraining with fresh data to adapt to changing economic conditions and client behaviours.

5. **Interactive Scenarios:** Enhance the Tableau dashboards to allow users to explore "what-if" scenarios, helping them understand how changes in client behaviour might impact their risk profile.

Conclusion & Questions

Our journey through credit risk prediction has revealed powerful insights and practical applications. Let's recap the core value of this project.

Recap: The Value of This Project

- Our "Credit Risk Prediction" project successfully analysed key factors influencing credit card default, demonstrating the power of data analytics in financial risk management.
- We identified critical behavioural patterns, such as payment delays and credit utilization, as strong indicators of default risk.
- The predictive models developed provide a valuable tool for financial institutions to proactively identify and manage high-risk accounts, potentially preventing significant financial losses.

Key Takeaways

- Data-driven insights are indispensable for robust credit risk assessment.
- Focusing on specific behavioral indicators can significantly improve prediction accuracy.
- Machine learning models offer a powerful capability to enhance traditional risk management.

Thank You!