Table of Contents

1. Introduction

Financial institutions rely on loan approval processes to minimize risk and ensure credit is granted to qualified applicants. However, traditional approval models often produce errors, leading to incorrect approvals or denials. These inaccuracies can result in increased loan defaults due to the acceptance of high-risk applicants and lost revenue from rejecting creditworthy borrowers.

To address these challenges, a new loan approval model has been developed to enhance decision accuracy and reduce errors. This study aims to evaluate the effectiveness of the new model in assisting loan officers with application reviews. An A/B test was conducted, where loan officers were randomly assigned to either a control group, using the existing model, or a treatment group, utilising the new model.

The primary objective of this analysis is to compare the performance of the new computer prediction against the current system using key metrics and statistical evaluation. By establishing an Objective Evaluation Criterion (OEC), this study systematically assesses whether the new approach improves loan approval accuracy. The report details the data preparation, data analysis methods, findings and recommendations to determine whether the new model offers a significant improvement over the existing system.

2. Experiment Setup

2.1 Null Hypothesis and Alternative Hypothesis

- **Null Hypothesis (H0):** The new AI model does not significantly improve loan officer decision quality compared to the existing model.
- **Alternative Hypothesis (H1):** The new AI model significantly improves loan officer decision quality by reducing errors and increasing decision confidence.

2.2 Overall Evaluation Criteria (OEC) serves as the primary metric for success, evaluating decision quality through reductions in Type I and II errors, increased alignment with AI, and improved confidence scores. A significance level of p = 0.05 is used for statistical analysis.

3. Data Understanding

3.1 Data Exploration

The density plots illustrating Type I and II errors in Figure 1, along with other variables analysed in the coding in Appendix, indicate a non-normal distribution. However, if the sample size exceeds 30 per group, the t-test remains valid despite non-normality. For smaller sample sizes (≤30) with a non-normal distribution, non-parametric tests are recommended, with the Mann-Whitney U test suitable for two-group comparisons and the Kruskal-Wallis test for three or more groups. Given that our study includes more than 30 samples per group, the t-test is an appropriate statistical method for analysis.

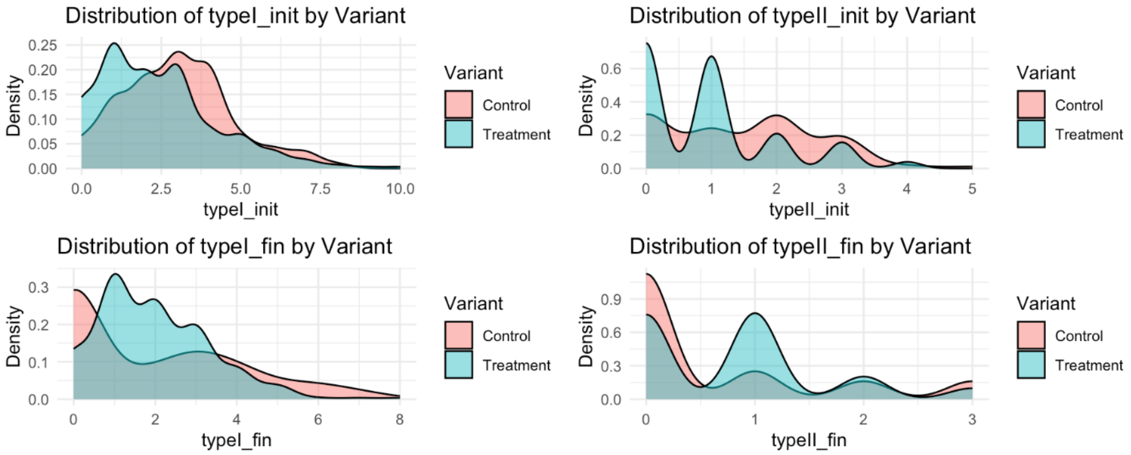


Figure 1: Example of Data Distribution (typeI and typeII)

3.2 Hypothesis Testing

For the hypothesis testing, the Treatment group performed better, with significantly lower Type I and Type II error rates than the Control group both before and after AI predictions as shown in Table 1. This indicates that the new computer model effectively reduces incorrect rejections of good loans and incorrect approvals of bad loans, leading to improved loan decision accuracy.

	TypeI_init	TypeI_fin	TypeII_init	TypeII_fin
Control	3.51	3.67	1.22	1.18
Treatment	2.41	2.02	1.02	0.77
P-value	< 0.05	< 0.05	0.17	< 0.05

Table 1: Hypothesis Testing of Type I and II errors

3.3 Data Preparation

To maintain dataset integrity, cases where fully_complt was not equal to 10 were excluded, as they indicated that loan officers did not strictly adhere to the step 1 and step 2 approval processes, representing a deviation from compliance.

Regarding the decision on whether to consolidate data by loan officers, we opted against this approach for several reasons. First, while variations in individual loan officers' decision-making abilities could influence the results, this effect primarily arises from the unequal distribution of data between the Variant groups. The primary objective of this study is to evaluate the effects of the old and new decision-support models on loan officer performance. If the models' effects were evenly distributed across all data points, variations in loan officers' decision-making capabilities would be significantly mitigated.

Second, consolidating the data would lead to the loss of critical information, potentially resulting in inaccurate conclusions. Additionally, after aggregation, only 38 observations would remain — an insufficient sample size to ensure the validity and generalizability of the findings. Thus, preserving the independence of individual data points is crucial for maintaining the comprehensiveness and reliability of the analysis.

Finally, it is important to acknowledge the imbalance in the number of loan officers between the Control and Treatment groups, as illustrated in Figure 2. To address this, Welch’s t-test was employed to account for unequal variances and sample sizes, ensuring the robustness and validity of the statistical results.

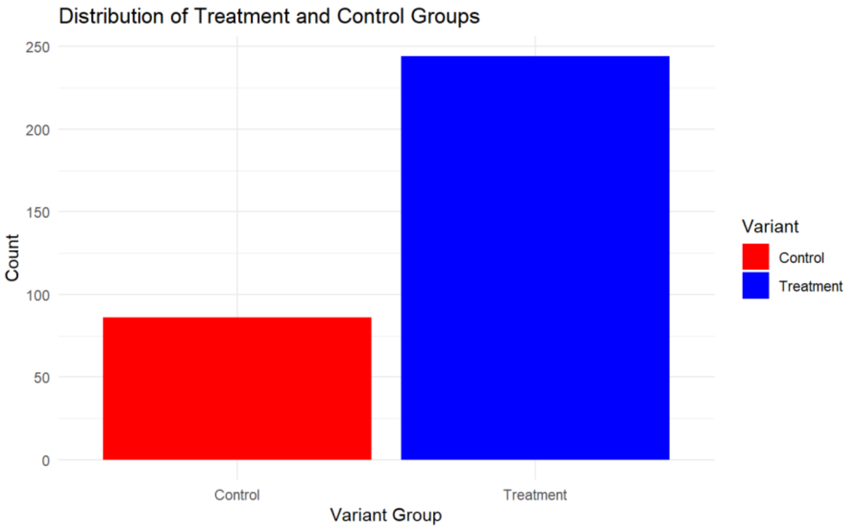


Figure 2: Imbalance Observations of Treatment and Control group

3.4 Feature Engineerining

Category	Metric	Definition	Reason
Recall	recall_init, recall_fin, recall_imp	Measures the ability to identify bad loans before and after AI	Evaluates AI's effectiveness in detecting risky loans
Precision	precision_init, precision_fin, precision_imp	Measures the ability to correctly approve good loans	Assesses AI's impact on reducing false approvals.
Conflict Rate	conflict_init_rate, conflict_fin_rate, conflict_dec	Tracks loan officer disagreement with AI before and after use	Measures AI's role in improving decision consistency
Confidence	confidence_imp	Change in loan officer confidence after AI assistance	Evaluates trust-building in AI-driven decisions

Table 2: Feature Engineering

From Table 2, Recall measures how well the model correctly identifies true positives from all positive samples in the dataset. In the current experiment, Bad loans is the type of error, which is more costly than good loans, therefore in the computation part, the number of bad loans is defined as a "positive" label.

The recall improvement represents the change in recall before and after the loan officer sees the model prediction. Precision measures how well the model avoids false positives when predicting true positives (bad loans). In this experiment, precision improvement shows whether the model reduces Type I errors after its predictions are seen by the loan officers.

Conflict decline measures the difference in the rate of conflict decreases after loan officers trust the model. A higher conflict decline means loan officers increase their trust in the model's predictions. It is a metric to describe how the model can help officers to reduce hesitations during their decision process, hence improving productivity and completion of the workflow.

Confidence improvement measures how much more certain loan officers feel about their final decisions after using the AI's recommendations. A higher confidence improvement suggests that the model helps officers make decisions more decisively.

The "Confidence" column has a different scale compared to other columns; therefore, data transformation and standardization will be applied as shown in Table 3.

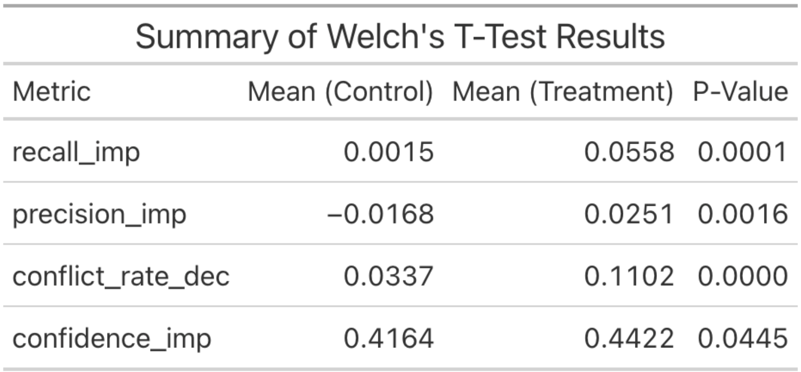
Category	Metric	Process	Reason
Data Transformation	log_confidence_imp	Log transformation	Reduces skewness from large values (max ~1000), ensuring a more balanced distribution
Standardization	scle_log_confidence_imp	Min-Max normalization	Scales values to [0,1] for better comparability across variables

Table 3: Data Transformation and Standardization

4. Data Analysis

4.1 Welch's Two-Sample t-tests

To compare the key metrics between the control and treatment groups, Welch's Two-Sample t-tests were conducted. The results indicated significant differences in recall, precision, conflict rate, and confidence as shown in Table 4.



Summary of Welch's T-Test Results

Metric	Mean (Control)	Mean (Treatment)	P-Value
recall_imp	0.0015	0.0558	0.0001
precision_imp	−0.0168	0.0251	0.0016
conflict_rate_dec	0.0337	0.1102	0.0000
confidence_imp	0.4164	0.4422	0.0445

Table 4: T-Test Results

The treatment group exhibited a significant increase in recall, meaning that loan officers using the AI model were better at identifying actual defaulters. Precision also saw a notable improvement, indicating that the new computer prediction of officers made fewer incorrect loan approvals. The conflict rate, which measures disagreement between loan officers and AI, significantly decreased, suggesting that officers in the treatment group were more aligned with AI recommendations over time. Confidence levels in the treatment group showed a statistically significant but smaller improvement, reflecting a gradual increase in trust in AI recommendations.

4.2 Building Overall Evaluation Criterion (OEC)

The OEC was constructed to quantify the effectiveness of the AI model, with Principal Component Analysis (PCA) used to determine the importance of each variable in OEC weighting. The key variables considered were recall improvement, precision improvement, conflict rate decrease, and log-transformed confidence improvement. PCA results showed that recall improvement accounted for 34.77% of the variance, making it the most critical factor, followed by precision improvement at 28.89%. The Conflict rate decrease contributed 23.49%, taking the total explained variance to 87.16%, while confidence improvement had the least impact at 12.84%. Using these results, the final OEC weights were assigned. This weighting ensures that OEC prioritises improvements in recall and precision while still considering conflict reduction and confidence enhancement.

4.3 Computing the Overall Evaluation Criterion (OEC)

The Welch's t-test showed a significant improvement in the Overall Evaluation Criterion (OEC) for the Treatment group compared to the Control group ($p < 0.05$). The mean OEC score for the Treatment group (0.1107) was significantly higher than that of the Control group (0.0570), indicating that the AI model improves decision-making performance. The 95% confidence interval [-0.0692, -0.0380] supports this finding, indicating that the observed improvement is unlikely to be due to random variation. This demonstrates a significant positive impact of AI intervention, confirming its efficacy in optimising loan approval decisions. The statistical significance of these findings provides strong evidence that new computer model is a viable addition to the current decision-making framework.

4.4 Computing Mean Differences

An analysis of mean differences between the control and treatment groups further reinforced these findings. The OEC showed a significant performance improvement, with mean decision quality increasing from 0.057 in the control group to 0.111 in the treatment group, which is an improvement of 93.96%. Recall saw a remarkable increase of 3596.31%, precision improved by 249.02%, conflict rates decreased by 226.93%, and confidence levels experienced a 6.17% rise. These results shown in Table 5 highlight the potential of the new computer model in enhancing loan officer decision-making efficiency and effectiveness.

Metric	Difference (Actual)	Difference (%)
OEC	0.0536	93.96
Recall	0.0543	3596.31
Precision	0.0419	249.02
Conflict	0.0765	226.93
Confidence	0.0257	6.17

Table 5: Mean Differences

4.5 Cohen’s d Effect Size Analysis

Cohen’s d was used to measure the practical significance of the difference in OEC between the Treatment and Control groups. The analysis resulted in Cohen’s d = 0.65, with a 95% confidence interval of [0.39, 0.90]. This corresponds to a medium effect size, indicating that the treatment group showed meaningful improvement in overall decision quality compared to the control group.

5. Recommendations and Conclusions

The findings show that the new computer model improves loan officer accuracy by reducing Type I errors (rejecting good loans) and Type II errors (approving bad loans), allowing more qualified applicants to receive credit while lowering default risks. The significant decrease in conflict rate suggests better alignment with AI recommendations, which simplifies the approval process. Although confidence levels have risen modestly, loan officers are gradually gaining trust in the AI model.

Given these findings, Analytics Manager and the Executives should move forward with deploying the AI model while ensuring optimal utilisation. Enhancing loan officer confidence is critical, as some may be hesitant to completely rely on AI-driven decisions. Structured training on how the AI generates recommendations, combined with clear and understandable explanations, will increase trust and adoption, allowing officers to seamlessly integrate the system.

A major takeaway from this analysis is that while the current sample size was sufficient to detect moderate effect sizes, however, for a more robust and reliable analysis, data aggregation is necessary. Aggregation reduced the sample size to 38, which is not enough statistical power. The power analysis indicated that a much larger sample of around 64 samples in each group would be needed to capture small but potentially meaningful effects with high confidence. This means that, while the results support implementing the model now, continued experimentation and data collection should be prioritised. Extending the experiment to include a larger dataset over a longer period will provide deeper insights into the model's impact, long-term trends in officer behavior and loan performance.

Continuous monitoring of the AI model's performance is critical, as effectiveness can change over time due to economic changes, applicant behaviour, or data biases. To ensure long-term reliability, model outputs should be reviewed on a regular basis and decision-making parameters refined. The AI model has demonstrated a significant improvement in loan approval accuracy and efficiency. By integrating the system while addressing user confidence, continuous monitoring, and additional data collection, the company can maximise the benefits of AI-driven decision-making, resulting in improved financial outcomes for both the organisation and its customers.