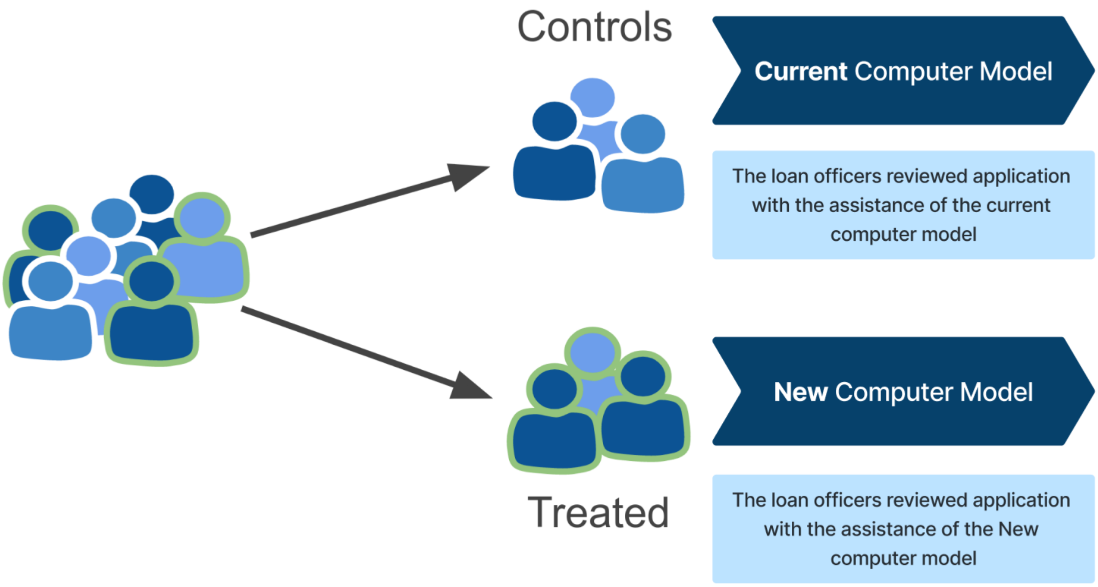
TABLE OF CONTENTS

1. Introduction

The company's loan approval process currently relies on an old AI model, prompting the Analytics Department to evaluate a new alternative. The goal is to determine whether the new model can optimize financial performance compared to the existing system.

An A/B test was conducted by randomly assigning loan applications to either a control group (current model) or a treatment group (new AI model). The evaluation focused on two Overall Evaluation Criteria (OECs): overall process efficiency and the influence of AI on loan officers. Statistical analyses, including Welch's t-test and effect size calculations (Cohen's d), were employed to assess the outcomes. The subsequent sections provide detailed insights into the experimental design, data analysis procedures, and recommendations for optimizing AI integration in the loan approval process.



The loan officers reviewed application with the assistance of the current computer model

The loan officers reviewed application with the assistance of the New computer model

2. Exploratory Data Analysis (EDA)

Through the EDA, the aim is to understand the dataset, identify patterns, inconsistencies, and potential biases that may affect the results.

2.1. AI Model Performance

The first important finding is that the AI consistently performs better in the treatment group, as measured by Type I and Type II error rates. The average error rates remain consistent across all offices, as shown in Figure 1.

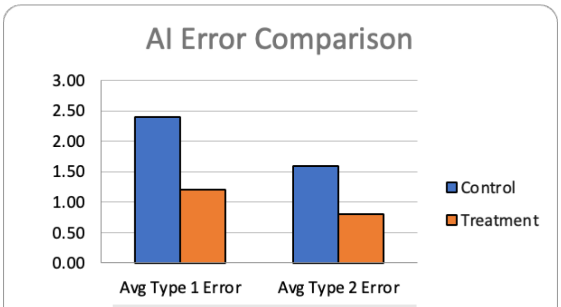


Figure 1

2.2. Treatment and Control Imbalance

Also, it's noticed that each office is assigned exclusively to either the control or treatment group, with no office participating in either. Furthermore, it revealed that the variant groups are imbalanced, as shown in Figure 2.

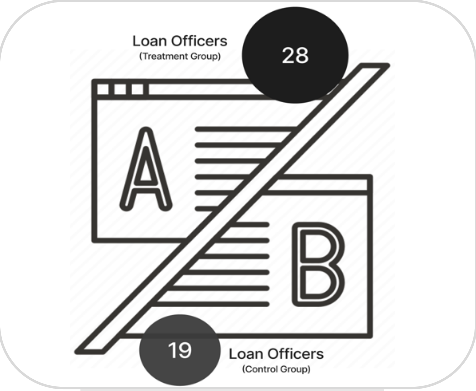


Figure 2

2.3. Officer Individual Performance

Additionally, human decisions before AI intervention are consistently better in the treatment group, understanding that it makes better decisions even before considering the model.

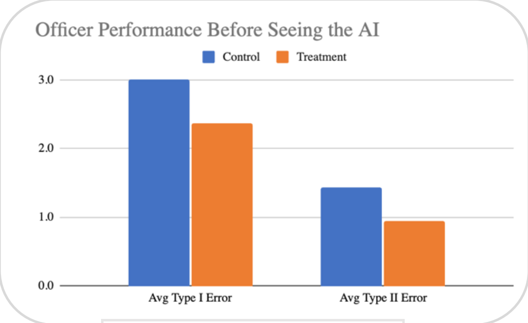


Figure 3

Data inconsistencies were identified with some missing values due to data entry or recording errors. Partial totals in some rows exist which suggest potential issues such as unprocessed decisions or misclassifications. These inconsistencies raise concerns about data integrity, but to maintain statistical power, all data is retained. However, this may lead to skewed or inconsistent findings. For a detailed examination, please refer to Appendix 2.

Additional EDA can be found in Appendix 8.

3. Data Preparation

The problem encompasses both the overall process and its three independent domains, each representing a key aspect of decision-making. The analysis of the system is done through three main factors: human decisions, model performance, and AI's influence on human decisions. The first two were addressed through the EDA, while the remaining two are evaluated through the defined OECs. It is important to note that the evaluation was conducted at an officer level, with each metric averaged per loan officer.

To assess overall process performance, the Profit Impact Error Index (PIEI) is introduced, a metric that balances profit and cost by calculating a weighted average of Type I and Type II error rates. To measure AI's influence on human decisions, the AI Influence Metric (AII) is defined, quantifying the extent to which loan officers align with or override AI recommendations.

3.1. PIEI

It measures the error rate in loan approval decisions by weighing each error according to its associated financial impact. Instead of merely counting errors, it incorporates the cost of each error, offering a view of how mistakes affect the company financially.

The goal of the process is to minimize financial risk by reducing costly errors. The PIEI metric measures how decision errors affect financial outcomes in the loan approval process. A lower PIEI signifies fewer errors and reduced financial impact, while a higher PIEI indicates substantial costs from errors, highlighting areas needing improvement. Thus, PIEI provides actionable insights into the model's effectiveness in minimizing both the occurrence and financial severity of mistakes.

Formula:

$$PIEI = \alpha * Avg\ \Delta\ Type\ I\ Error\ Rate + \beta * Avg\ \Delta\ Type\ II\ Error\ Rate$$

where:

$Avg\ \Delta\ Type\ I\ Error\ Rate = Average\ Difference\ of\ Type\ I\ Errors$

$Avg\ \Delta\ Type\ II\ Error\ Rate = Average\ Difference\ of\ Type\ II\ Errors$

$\alpha =\ Weight\ assigned\ to\ Type\ I\ Error,\ representing\ the\ loss\ due\ to\ approving\ bad\ loans$

$\beta =\ Weight\ assigned\ to\ Type\ II\ Error,\ representing\ the\ profit\ due\ to\ rejecting\ good\ loans$

Financial assumptions regarding the average interest rate (21.44%) (Wangman, 2024), average loan amount ($27,500), and recovery rate (70.90%) (spglobal.com, n.d) are made in order to calculate “loss per bad loan” and “profit per good loan”. These values are then used to determine Loss Weight and Profit Weight. Further information about the calculation can be found in the Appendix 3.

Since a bad loan results in a higher absolute loss ($8,730) than the profit from a good loan ($6,432), the loss weight is greater than the profit weight. As a result, Type II errors (false negatives—approving bad loans) contribute more heavily to the weighted OEC than Type I errors (false positives—rejecting good loans). This aligns with the business objective of minimizing financial risk, as approving bad loans is costlier than rejecting good ones.

3.2. AII

The AII metric measures the extent to which loan officers follow or override AI recommendations. It quantifies the alignment between AI predictions and human decisions, capturing how much influence the AI model has on the final loan approval process. It helps to assess whether the AI guides the decision-making or if officers tend to disregard its recommendations.

Formula:

$$AII\ = Avg\left(\frac{Human\ Decisions\ Aligned\ with\ AI - Human\ Decisions\ Overriding\ AI}{Total\ Human\ Decisions}\right)$$

where:

$$\mathit{Human\ Decisions\ Aligned\ with\ AI} = \mathit{Cases\ where\ the\ loan\ officer\ followed\ the\ AI\ recommendation}$$
$$\mathit{Human\ Decisions\ Overriding\ AI} = \mathit{Cases\ where\ the\ loan\ officer\ disagreed\ with\ the\ AI,\ made\ a\ different\ decision}$$
$$\mathit{Total\ Human\ Decisions} = \mathit{The\ sum\ of\ all\ cases\ where\ the\ loan\ officer\ made\ a\ final\ decision}$$

This metric ranges from −1 to +1. Values closer to +1 indicate that loan officers rely heavily on AI recommendations, consistently aligning their decisions with the model’s suggestions. Conversely, values closer to −1 suggest that loan officers frequently override AI recommendations, showing a preference for independent judgment. A value near 0 implies that AI has little to no influence on human decisions. Further information about the calculation of this metric can be found in Appendix 4.

4. Statistical Analysis & Interpretation

4.1. Profit Impact Error Index (PIEI)

A Welch’s Two-Sample t-test was conducted to assess the effectiveness of the new loan approval model in (Treatment) compared to the existing model (Control) based on the PIEI metric to see the overall reduction in error rate based on profit over good loans and loss over bad loans.

4.1.1. Hypothesis:

- H_0: The new loan approval model does not cause a significant improvement in the profit impact error index of loan officers’ decisions compared to the old model.
- H_1: The new loan approval model leads to a significant improvement in the profit impact error index of loan officers’ decisions compared to the old model.

4.1.2. Results & Interpretation

An independent samples t-test revealed a significant difference between the Treatment (mean: −0.0638) and Control (mean: −0.2167), $p < 0.01$. Thus, the null hypothesis is rejected in favor of the alternative. This proves that the overall performance in the treatment group is significantly better than control. The complete analysis of the test can be found in Appendix 5.

Additionally, Cohen’s d (−0.9024) suggests a large effect size, confirming the significant impact of the new model on loan officers' decision-making. Given the calculation, the reduction in the error rate suggests that the new model is more effective in minimizing financial risk. These findings provide strong evidence that the new loan approval model improves decision quality and aligns with the business objectives.

4.2. AI Influence on Human Decision-Making Score

In order to see the influence of AI on Human Decision Making in the current model (Control) and the new model (Treatment), we conducted Welch's t-test.

4.2.1. Hypothesis:

- **Null Hypothesis (H_0):** The new AI model does not significantly influence loan officers' decisions compared to the old model.
- **Alternative Hypothesis (H_1):** The new AI model significantly influences loan officers' decisions; those in the Treatment group align more with AI recommendations than those in the Control group, indicating a measurable impact.

4.2.2. Results & Interpretation

The results indicate a statistically significant difference between the Control and Treatment (p= 0.0068), with the Treatment group showing a higher mean (0.1114) than Control (0.0425). The 95% confidence interval confirms the difference is unlikely due to chance, leading to the rejection of the null hypothesis. This confirms the new AI model significantly increases its influence on loan officers' decisions. The complete analysis of the test can be found in Appendix 6.

However, further analysis is needed to assess whether this influence improves loan approvals or fosters over-reliance on AI. The small effect size ($Cohens'\ d$ = −0.4616) further supports the AI model's influence, with the negative value indicating greater reliance on AI recommendations in the Treatment group.

5. Recommendations & Conclusion

The analysis reveals that the current design suffers from treatment-control imbalances and data inconsistencies, making it impossible to draw reliable conclusions. To address these issues, a new design should focus on intra-office randomization and mandatory data consistency. These, along with additional recommendations, are detailed in the following sections.

5.1. For the Analytics Manager

To improve the experiment's design and data integrity, we recommend rerunning the experiment with an enhanced design. Specifically, the following actions should be taken:

5.1.1. Data Quality and Consistency

Implement stricter data collection and validation protocols to address inconsistencies observed in the initial agreement and conflict checks, where in 20.4% of cases, the initial numbers did not match properly. Similarly, completion consistency checks reveal that 30% of the data compared different loan counts between the initial and final stages, compromising the validity of comparisons. To resolve this, we recommend making both the initial and final stages of the process mandatory, ensuring consistent data tracking throughout the whole process.

5.1.2. Enhanced Randomization

Adopt intra-office randomization to mitigate biases introduced by assigning entire offices to either the Control or Treatment group, ensuring that each loan officer experiences both conditions to eliminate officer-specific biases and achieve a balanced sample within each office. Since a new intra-office randomization process is proposed to eliminate biases from independent human decision-making, certain precautions must be taken to ensure its effectiveness. This includes proper loan-level randomization, ensuring that loans are evenly distributed across conditions, and maintaining a sufficiently large sample size to achieve reliable and statistically valid comparisons.

5.1.3. Inclusion of Financial Variables

Incorporate critical financial information, such as loan type, loan amount, and interest rate, to enhance the AI model's decision-making capabilities and improve financial impact assessments

5.2 For the Executive Team

To ensure a comprehensive and unbiased evaluation of the AI model, we propose the following strategic adjustments:

5.2.1. Experiment Design

To properly assess the new model's performance and reduce biases related to officer characteristics, we propose rerunning the experiment with an improved design. While the current experiment yields statistically significant results, its structure does not fully address the model's improvement, limiting the validity of the conclusions. The new experiment trial, incorporating the enhanced design, should be conducted over the same 10-day period but must include at least 49 loan officers to achieve statistically significant results, as determined by the power analysis detailed in Appendix 7.

5.2.2. Establish Experimental Objectives and Success Thresholds

The current experiment lacks well-defined, quantifiable targets making it difficult to determine whether the new AI truly delivers business value or merely shows technical improvement. By setting clear performance benchmarks, for example requiring a measurable drop in default risk or a confirmed uptick in officer confidence, organizations can tie the experiment's outcomes more closely to overall corporate goals. Once these are established, any benefits uncovered will map directly to operational efficiencies and profitability, providing a stronger justification.

5.2.3. Demand a Targeted Follow-Up Experiment Before Any Larger Implementation

Moving forward with partial or ambiguous findings could waste resources and risk reputational damage if the model underperforms or introduces biases. Approve a second, more rigorous experiment that addresses the design flaws and tracks the newly defined success metrics. Request interim updates on the model's progress, ensuring executive oversight remains aligned.

By implementing these recommendations, its aim is to achieve a more accurate and reliable assessment of the AI model's performance, ensuring that any conclusions drawn are well-founded and beneficial for the organization's decision-making processes.

6. References

- S&P Global Ratings. (2023). Default, Transition, And Recovery: U.S. Recovery Study: Loan Recoveries Persist Below Their Trend. S&P Global. (Accessed: 8 February 2025)

- Business Insider. (n.d.). Average personal loan interest rates. Business Insider. (Accessed: 8 February 2025)

7.2. Appendix 2: Data Integrity Checks

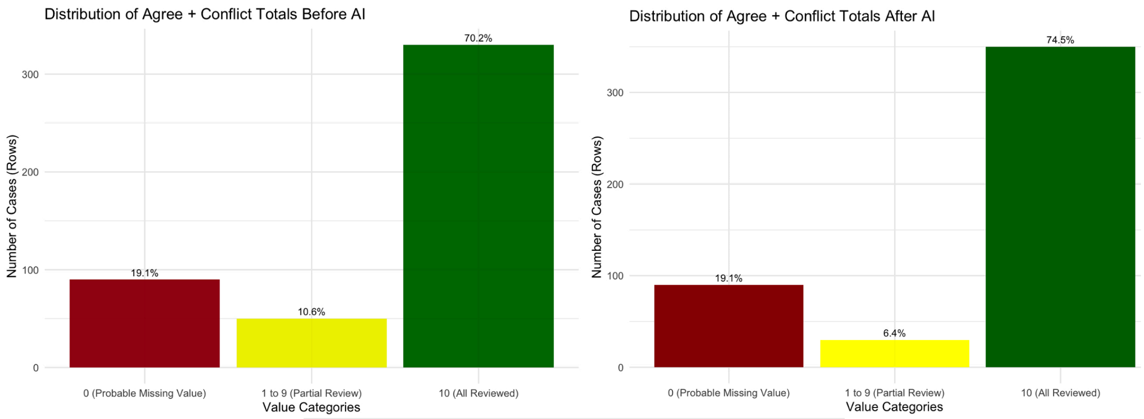


Figure 4, Distribution of Agree and conflict columns

These 2 charts highlight some critical data inconsistencies. About 100 rows, constituting 19.1% of the dataset, can reasonably be assumed to represent missing values. This is evident as their total for agree and conflict cases is zero, suggesting that these rows likely suffer from data entry or recording errors. These inconsistencies may stem from manual data entry mistakes or system issues in capturing officer decisions before and after using AI predictions.

For rows where the total of agree and conflict cases falls between 1 and 9, several plausible scenarios could explain these partial totals. It's possible that some officers left decisions unprocessed due to time constraints, leading to incomplete data. Alternatively, these could result from misclassification, where some decisions might have been marked as neutral or categorized incorrectly. Another potential cause is data entry errors, which might have omitted or miscounted certain decisions. Given the presence of these inconsistencies, there is also a concern about double-counting. Specifically, we cannot ascertain whether the agree and conflict cases recorded before AI and those after AI refer to the same set of loans.

Despite these challenges, we chose not to delete any rows to preserve the statistical power of the analysis and approach this as a real-world scenario. However, it must be acknowledged that the findings derived from this dataset might be skewed or inconsistent due to the inherent issues in the data. We recommend improving the data collection process in our recommendations section to ensure better accuracy in future analyses.

7.3. Appendix 3: PIEI Calculation

The calculation of this metric relies on assumptions based on referenced financial factors to accurately capture the impact of each error rate in terms of potential profit or loss. To achieve this, key financial parameters must be considered, including the interest rate, which is essential for estimating potential profit, and the recovery rate, which determines the average loss per loan. With these values, further calculations can be performed to assess the financial implications of decision-making errors effectively.

First, it is needed to calculate the error rates.

$$Type\ I\ Error\ Rate\ Initial = \frac{\#\ of\ \ Type\ I\ Error\ Initial}{\#\ of\ \ Good\ Loans}$$

$$Type\ II\ Error\ Rate\ Initial = \frac{\#\ of\ \ Type\ II\ Error\ Initial}{\#\ of\ \ Bad\ Loans}$$

Both calculations capture the potential harm caused by misclassifying loans into the wrong category. The Type I Error Rate measures the proportion of good loans incorrectly rejected relative to the total number of good loans, while the Type II Error Rate quantifies the proportion of bad loans incorrectly approved relative to the total number of bad loans. The same approach is applied at the final stage to assess the overall impact of decision-making errors throughout the process.

$$Type\ I\ Error\ Rate\ Final = \frac{\#\ of\ \ Type\ I\ Error\ Final}{\#\ of\ \ Good\ Loans}$$

$$Type\ II\ Error\ Rate\ Final = \frac{\#\ of\ \ Type\ II\ Error\ Final}{\#\ of\ \ Bad\ Loans}$$

The four previously mentioned calculations are performed at the officer-day level, meaning that for each day and for each loan officer, these four rates are individually computed, providing a granular assessment of performance over time.

Next, the difference between these rates is calculated, maintaining the officer-day level without any aggregation at this stage. This ensures that the analysis captures daily variations in performance for each loan officer before any further summarization.

$$\Delta\,Type\ I\ Error\ Rate = Type\ I\ Error\ Rate\ Final - Type\ I\ Error\ Rate\ Initial$$
$$\Delta\,Type\ II\ Error\ Rate = Type\ II\ Error\ Rate\ Final - Type\ II\ Error\ Rate\ Initial$$

After, the results need to be aggregated at the office level to enable a comparison of performance per officer and per variant. This is done by computing the average error rates for each loan officer, allowing for a clearer assessment of differences between the Control and Treatment groups.

$$\Delta Avg\ Type\ I\ Error\ Rate = \sum \frac{\Delta\,Type\ I\ Error\ Rate}{\#\,of\ Days\ per\ Officer}$$
$$\Delta Avg\ Type\ II\ Error\ Rate = \sum \frac{\Delta\,Type\ II\ Error\ Rate}{\#\,of\ Days\ per\ Officer}$$

To determine the weights, we first calculate the loss per bad loan and the profit per good loan.

$$Loss\ per\ bad\ loan\ =\ average\ loan\ amount\ *\ (1 -\ recovery\ rate)$$

$$Profit\ per\ good\ loan\ =\ average\ loan\ amount\ *\ average\ interest\ rate$$

Loss weight: $$\alpha\ =\ \frac{Loss\ per\ Loan\ per\ Officer}{Profit\ per\ Loan\ per\ Officer\ +\ Loss\ per\ Loan\ per\ Officer}$$

Profit weight: $$\beta\ =\ \frac{Profit\ \ per\ Loan\ per\ Officer}{Profit\ per\ Loan\ per\ Officer\ +\ Loss\ per\ Loan\ per\ Officer}$$

Finally, all essential components must be integrated into the final formula to ensure a comprehensive calculation that accurately captures the impact of decision-making errors.

$$PIEI = \alpha\ *\ \ \Delta\ Avg\ Type\ I\ Error\ Rate + \ \beta *\ \ \Delta\ Avg\ Type\ II\ Error\ Rate$$

Once the data is aggregated and integrated into the final formula, a statistical comparison will be conducted at the office level to determine whether there is a statistically significant difference between variants.

7.4. Appendix 4: All Calculation

The calculation of this metric is based on the final-stage decisions made by humans after viewing the AI recommendations. This allows for a comprehensive analysis of how AI influences human decision-making, helping assess whether loan officers are aligning their choices with the model's suggestions.

The metric is straightforward: first, we compute the difference between the number of decisions made in agreement with the AI model and those made against it. This difference is then divided by the total number of decisions, allowing us to measure the overall influence of AI on human decision-making. This step is crucial because the number of decisions made per day varies, and standardizing the results ensures a fair comparison across different days and officers.

Based on the column names provided in the dataset and without any aggregation, the formula would be:

$$AII\ unit = \frac{\#\ of\ decision\ revised\ per\ AI - \#\ of\ decisions\ revised\ against\ AI}{\#\ of\ total\ decisions}$$

Afterward, we compute the aggregation at the officer level by taking the average of the AI Influence Index (AII) unit, resulting in:

$$AII = \sum \frac{AII\ unit}{\#\ of\ Days\ per\ Officer}$$

Finally, this measure allows us to compare AI influence at the office level and conduct a statistical test to determine whether there is a statistically significant difference between the two variants.

7.5. Appendix 5: Welch's t-test for PIEI

The t-test for PIEI comparing the Control and Treatment groups yielded a t (209.09) = -8.2952. The 95% confidence interval for the mean difference, ranging from [-0.1893 to -0.1166], does not include zero, confirming that the difference is statistically significant. Specifically, the average value for the Control group (-0.2167) is significantly lower than that for the Treatment group (-0.0638), demonstrating a difference between the two groups.

7.6. Appendix 6: Welch’s t-test for AI

The Welch Two Sample t-test for AI influence by comparing the Control and Treatment groups yielded a t (275.63) = -4.865. The 95% confidence interval for the mean difference, ranging from -0.0968 to -0.0410, does not include zero, confirming that the difference is statistically significant. Specifically, the average value for the Control group (0.0425) is significantly lower than that for the Treatment group (0.1114), indicating that the Treatment group exhibits a higher average influence difference.

7.7. Appendix 7: Power Analysis

This power analysis call uses the pwr package’s t-test function to determine the sample size. The details are as followings:

- **Power = 0.8 (80% power):** This means we want an 80% chance of detecting a true effect if it exists.

- **Cohen’s d = .7:** A d value of 0.7 is considered a medium-to-large effect. By specifying this effect size, we are basing our sample size calculation on the assumption that the true difference between the groups is substantial enough to be detected if it exists.

- **Significance level = 0.01:** The alpha level is set to 0.01, which is more stringent than the more commonly used 0.05. A stricter significance level reduces the risk of Type I errors (falsely rejecting the null hypothesis). However, lowering the alpha increases the required sample size to maintain the same power.

- **Type = "two.sample":** This indicates that we are planning a two-sample t-test, comparing two independent groups (for example, a Control group versus a Treatment group).

7.8 Appendix 8: Additional EDA

7.8.1. Confidence Shift

The confidence scores in the two groups show opposite trends. The control group's confidence decreases after seeing AI predictions, while the treatment group's confidence increases, suggesting that the AI predictions may align more with officer judgment in the new model as shown in Figure 2.

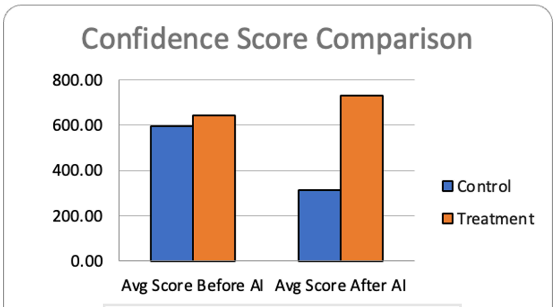


Figure 5

7.8.2. Decision Revision

The treatment group has a higher rate of decisions revised to match AI, suggesting officers trust AI more under the new model. If error rates decrease, this would confirm that the new model improves officer decisions.

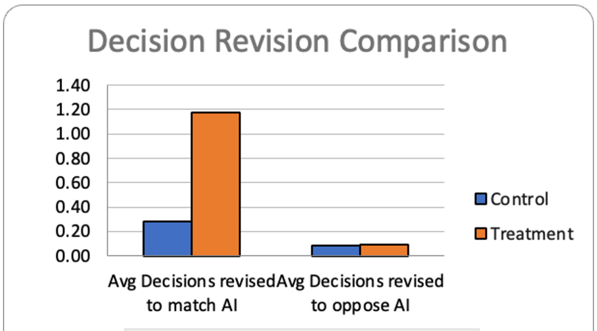


Figure 6