WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

Group Name **Students**	Group 17 5610758 5667293 5672753 5629770 5646570 5609131
Module Code	IB98D0
Module Title	Advanced Data Analysis
Submission Deadline	13-Feb-2025 12:00:00 PM
Date Submitted	12-Feb-2025 23:01:22 PM
Word Count	1988
Number of Pages	20
Question Attempted	Attempted all questions.
Have you used Artificial Intelligence (AI) in any part of this assignment?	Yes

If you have ticked "Yes" above, please briefly outline below which AI tool you have used, and what you have used it for. Please note, you must also reference the use of generative AI correctly within your assessment, in line with the guidance provided in your student handbook.

We use ChatGpt to check grammar and explain coding errors.

TABLE OF CONTENTS

Introduction

The company faces significant challenges due to high error rates in loan approvals, leading to many bad loans being approved and later defaulting, causing substantial financial losses. To address this issue, the Analytics Department developed a new AI-powered loan review model and conducted an A/B testing to evaluate its effectiveness. This report analyzes the preliminary results of the experiment to determine whether the new model reduces Type II errors (approving bad loans that later default) and aligns with the company's overall financial objectives. Additionally, the analysis evaluates improvements in secondary performance metrics, such as Type I error reduction, agreement rates, and confidence levels, to assess whether the new model enhances loan officer decision accuracy.

Overall Evaluation Criteria (OEC)

Primary OECs (Type two error improvement (typeII_fin - typeII_init)

The primary objective is to minimize financial losses from defaulted loans while maximizing profitability from successful repayments. However, the high error rates in the loan approval process cause many bad loans to be approved, significantly impacting financial stability. Type II error (false negatives)—has been chosen as the primary OEC to address this. High Type II errors lead to direct financial losses. Reducing these errors is critical as it reflects the AI's ability to minimize risky loan approvals, thereby improving decision accuracy and aligning with the company's goal of enhancing long-term profitability and stability.

Secondary OECs (Supporting performance metrics)

- **Agreement Rate**: Measures how often loan officers align with AI recommendations, indicating trust and AI reliability.

- **Revision Rate**: Tracks how frequently officers adjust decisions to follow AI, showing its influence on improving judgment.

- **Confidence Level**: Reflects how certain loan officers feel about their decisions after AI assistance, assessing AI's impact on decision clarity.

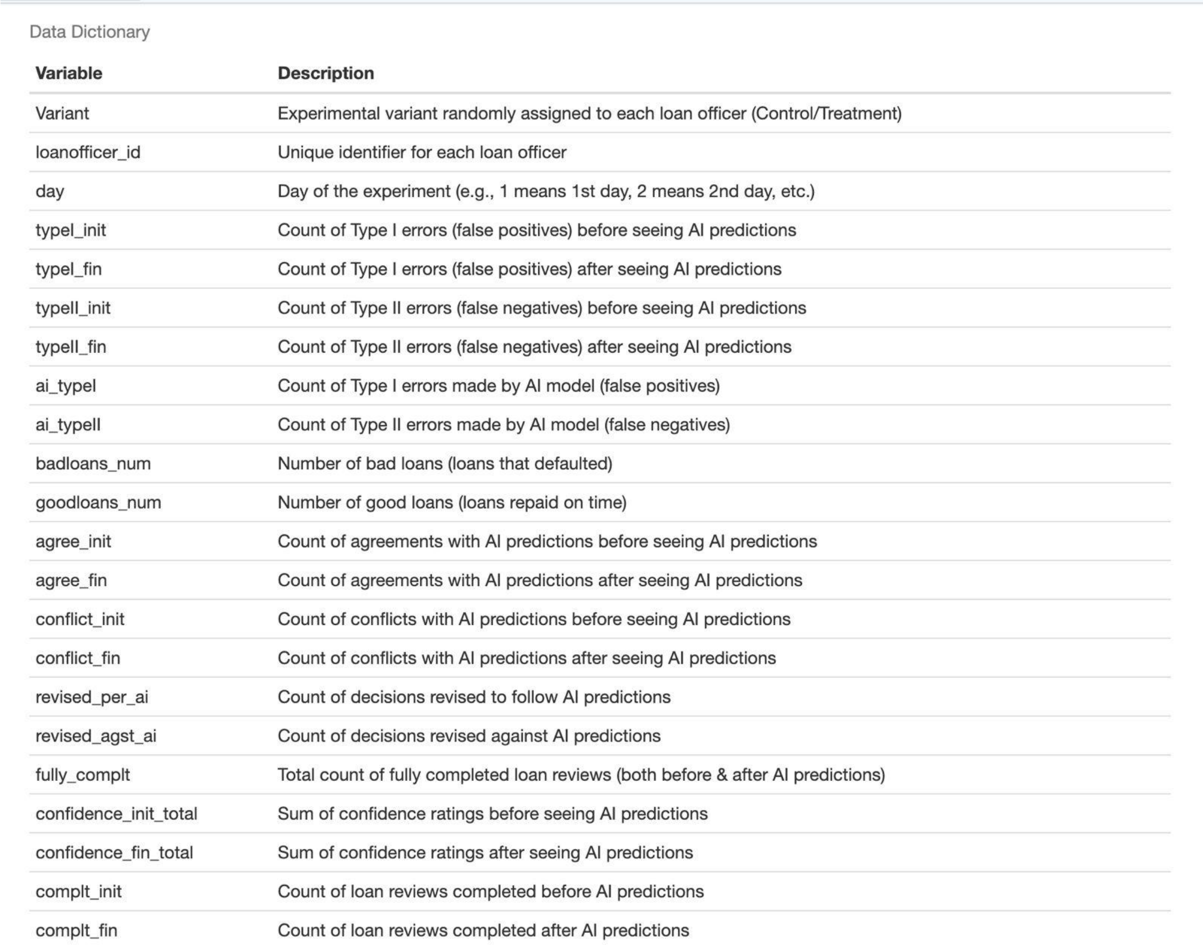
- **Type I Error Improvement**: Evaluates AI's role in reducing false positives (rejecting good loans), ensuring profitability isn't compromised.

To assess how well the AI model distinguishes between good and bad loans

- **Precision:** Ensures that when AI identifies a loan as bad, it is truly a bad loan, minimizing unnecessary rejections of good applicants.
- **Recall Rate:** Measures AI's ability to detect actual bad loans, ensuring risky approvals are minimized.
- **F1 Score:** Balances precision and recall, comprehensively measuring AI's classification accuracy in reducing financial risks.

Data Preparation

Data Dictionary



Data Dictionary

Variable	Description
Variant	Experimental variant randomly assigned to each loan officer (Control/Treatment)
loanofficer_id	Unique identifier for each loan officer
day	Day of the experiment (e.g., 1 means 1st day, 2 means 2nd day, etc.)
typeI_init	Count of Type I errors (false positives) before seeing AI predictions
typeI_fin	Count of Type I errors (false positives) after seeing AI predictions
typeII_init	Count of Type II errors (false negatives) before seeing AI predictions
typeII_fin	Count of Type II errors (false negatives) after seeing AI predictions
ai_typeI	Count of Type I errors made by AI model (false positives)
ai_typeII	Count of Type II errors made by AI model (false negatives)
badloans_num	Number of bad loans (loans that defaulted)
goodloans_num	Number of good loans (loans repaid on time)
agree_init	Count of agreements with AI predictions before seeing AI predictions
agree_fin	Count of agreements with AI predictions after seeing AI predictions
conflict_init	Count of conflicts with AI predictions before seeing AI predictions
conflict_fin	Count of conflicts with AI predictions after seeing AI predictions
revised_per_ai	Count of decisions revised to follow AI predictions
revised_agst_ai	Count of decisions revised against AI predictions
fully_complt	Total count of fully completed loan reviews (both before & after AI predictions)
confidence_init_total	Sum of confidence ratings before seeing AI predictions
confidence_fin_total	Sum of confidence ratings after seeing AI predictions
complt_init	Count of loan reviews completed before AI predictions
complt_fin	Count of loan reviews completed after AI predictions

Data Cleaning

The given dataset was cleaned to ensure data integrity and logical consistency:

1. **Missing values check**

A check was performed to confirm no missing values remain in the dataset.

2. **Data Filtering**

We filtered the dataset to include only rows where fully_complt == 10, reducing entries from 470 to 330. This ensures all loan officers had the same number of completed reviews, enabling fair comparison and improving analysis reliability.

Feature Engineering

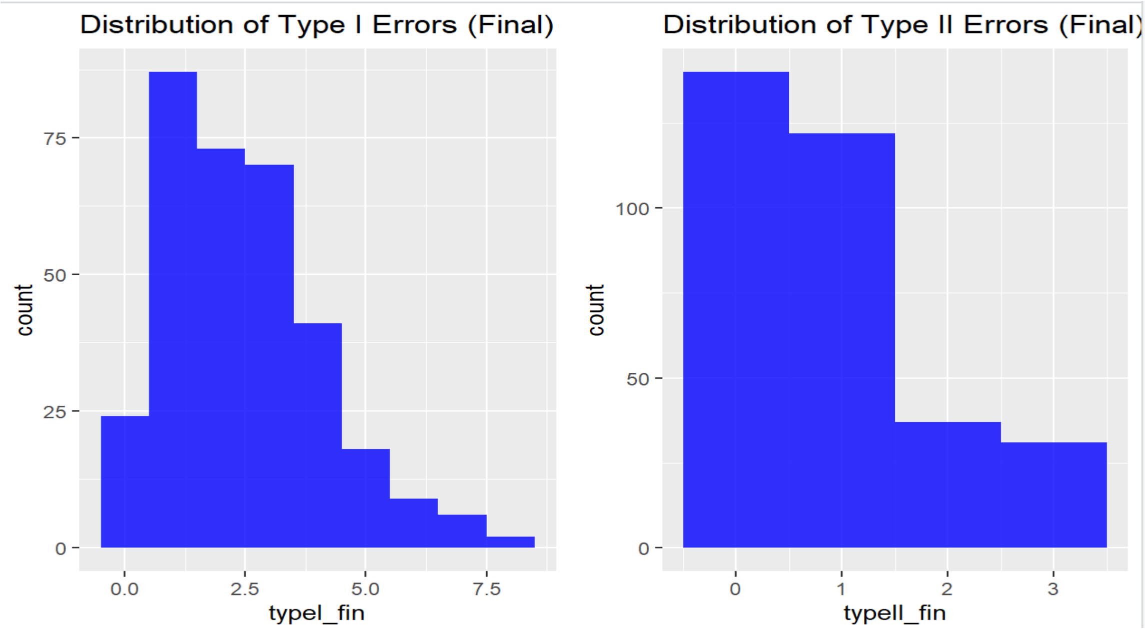
New variables were created to quantify performance:

- TypeII Error Improvement= typeII_fin - typeII_init
- Agreement Rate = agree_fin / (agree_fin + conflict_fin)
- Revision Rate = revised_per_ai / (revised_per_ai + revised_agst_ai)
- Confidence Level = confidence_fin_total / complt_fin
- TypeI Error Improvement = (typeII_fin - typeII_init)
- F1 Score = 2 * (precision * recall) / (precision + recall)
- precision = badloans_num / (badloans_num + typeI_fin)
- Recall Rate = badloans_num / (badloans_num + typeII_fin)

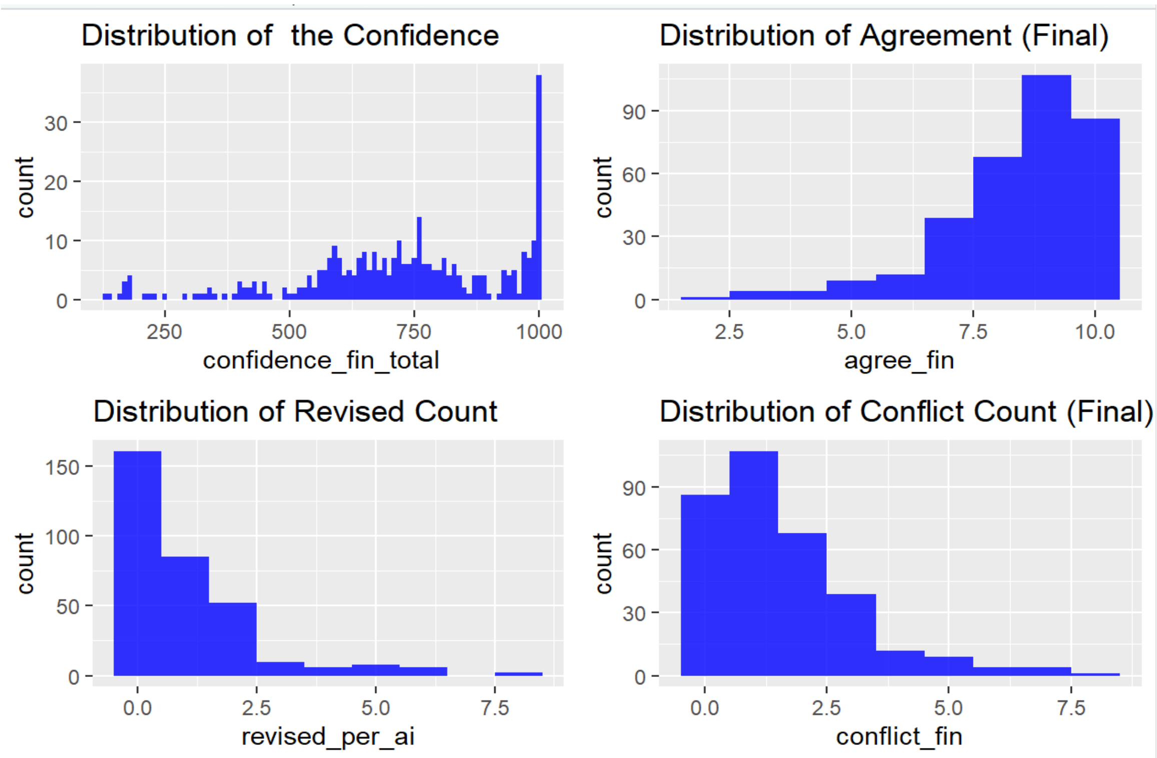
Exploratory data analysis

Key metrics were summarized to assess data distribution and trends.

Distribution Analysis:

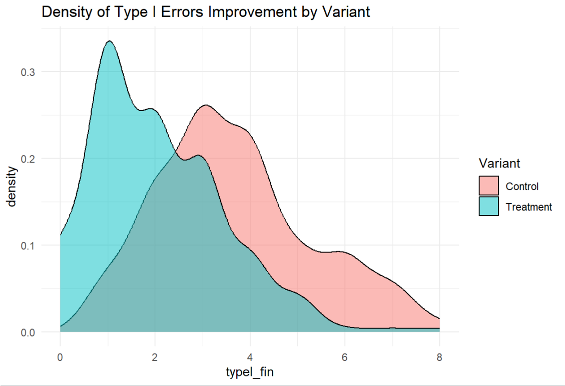


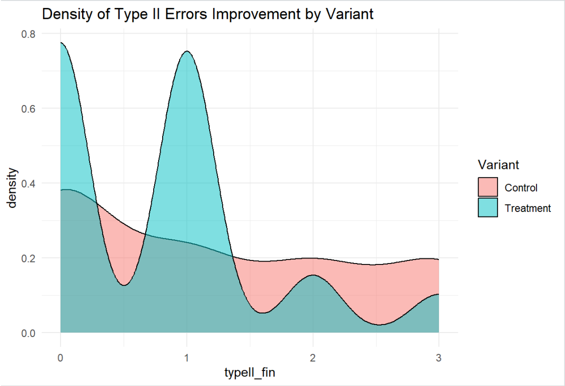
The histogram shows that most values are concentrated on the left, primarily around 0 and 1, while fewer cases have higher Type II Errors around 2 or 3. The histogram of Type I errors (Final) indicates that the majority of loan officers made between 0 and 3 false positive errors, with only a few officers making excessive Type I errors.



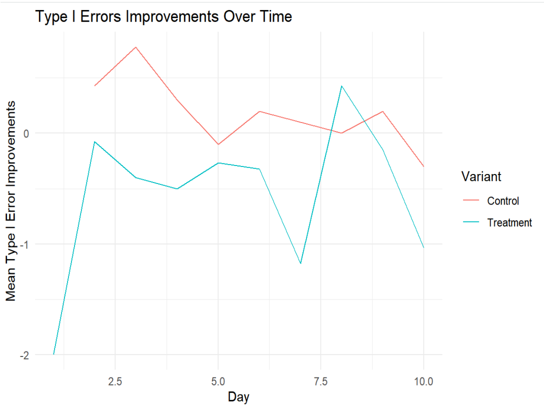
This Confidence distribution shows that most loan officers reported moderate to high confidence, with a sharp peak at the maximum value, indicating some reported very high confidence in their decisions after seeing AI predictions. The final agreement indicates that the majority of loan officers aligned with the computer predictions, experiencing minimal conflict. The histogram of the Revised Count reveals that most loan officers made between 0 and 2 revisions in favor of AI, with only a small number making a significantly higher number of revisions. Lastly, The final conflict count demonstrates that most officers experienced little or no conflicts with AI predictions.

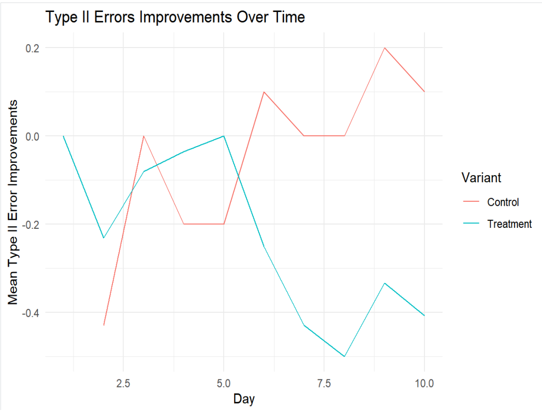
Comparing Control vs. Treatment Groups:





The density analysis suggests that AI improved loan decisions, with the Treatment group showing a higher density of lower Type I errors (reducing the rejection of good loans) and a higher concentration of lower Type II errors (better rejecting bad loans).





The results of the two graphs show that the treatment group appears to be more responsive to change but less consistent, while the control group has a more consistent but less significant improvement in pattern. This may indicate that while the treatment initially reduces error, its impact decreases over time.

Statistical Analysis

Hypothesis Testing

Null Hypothesis (H_0): There is no significant difference between the Control and Treatment groups for the given metric.

Alternative Hypothesis (H_1): There is a significant difference between the Control and Treatment groups for the given metric.

* T-tests were performed to compare key metrics between the Control and Treatment groups.

* p-value < 0.05 was considered statistically significant.

Table 1: T-test results with interpretation

Metric	p-value	Mean (Control Group)	Mean (Treatment Group)	Mean Difference	Interpretation
Type I Errors Improvement	9.799e-06	0.16	-0.40	-0.56	Treatment group had significantly lower Type I errors.
Type II Errors Improvement	0.002662	-0.04	-0.25	-0.21	Treatment group had significantly lower Type II errors.
Agreement Rate	1.196e-08	0.74	0.88	0.14	Treatment group had a significantly higher agreement rate.
Revision Rate	0.01047	0.71	0.91	0.20	Treatment group revised significantly more decisions to align with AI
Confidence Final Rate	3.174e-10	61.17	76.47	15.30	Treatment group had significantly higher confidence in final decisions
F1 Score	0.001695	0.57	0.63	0.06	Treatment group had a significantly higher F1 Score
Recall	0.118	0.79	0.75	-0.04	Not significant

Interpretation and Business Impact

Primary OEC

- Type II Errors Improvement

The treatment group also reduced Type II errors by -0.21, meaning fewer bad loans were mistakenly approved than the control group. From a financial standpoint, reducing Type II errors is crucial because it minimizes the likelihood of loan defaults, thereby lowering financial losses.

Secondary OEC

- Type I Errors Improvement

The treatment group significantly reduced Type I errors by -0.56, meaning fewer good loans were mistakenly rejected than the control group. This indicates that the new AI model helps loan officers identify and approve creditworthy applicants.

- Agreement Rate

Loan officers in the treatment group showed higher agreement with AI recommendations (+0.14), indicating increased trust in its predictions. If AI predictions are accurate, this improvement can help reduce human decision errors to ensure more consistent decision-making, reducing human variability and improving efficiency in the loan approval process.

- Revision Rate

Loan officers revised their decisions to align with AI more frequently (+0.20), demonstrating that the model provides valuable insights to correct human errors. This improvement potentially reduces manual decision biases, leads to more accurate loan approvals, and promotes data-driven decision-making.

- Confidence Final Rate

The treatment group reported significantly higher confidence in their final decisions (+15.30), suggesting that the new model provides more reliable decision support, improving decision clarity and reducing uncertainty. Increased confidence may speed up loan approval processes and improve operational efficiency.

- Recall

The difference in recall (-0.04) indicates that while the AI model improves other aspects of decision-making, its ability to detect bad loans has not significantly improved, which means there is still a risk of approving loans that may default.

- F1 Score

The treatment group achieved a higher F1 score (+0.06), indicating improved classification accuracy by minimizing false positives and false negatives, thereby enhancing the overall quality of loan approvals. This can lead to better loan management, reduced financial risks, and increased operational efficiency.

Recommendations

The experiment should continue, as the Primary OEC showed a significant reduction in false negatives, demonstrating measurable benefits in reducing risky loan approvals. However, the reduction remains moderate, suggesting that while the new AI model helps lower default risks, further refinement is needed to enhance its ability to identify bad loans more effectively.

To address this, we examined the effect size (Cohen’s d) to assess the practical impact of the AI model beyond statistical significance, providing a clearer understanding of how meaningful these improvements are for business performance. Additionally, we calculated the required sample size to ensure the experiment achieves 80% statistical, allowing us to determine whether the current dataset is sufficient for reliable conclusions or if further data collection is necessary.

Table 2: Cohen’s d and effect size

OEC	Cohen’s d	Effect size	Required Sample Size	Current Control Group Size	Current Treatment Group Size	Reliability
Type II Error Improvement	0.43	Moderate	87	86	244	Relatively reliable, but the Control group slightly below the requirement
Type I Error Improvement	0.33	Small to Moderate	146	86	244	Limited by a small Control group sample, needs more data
F1 Score	-0.27	Slight Decline	214	86	244	Sample size far below requirement, results may be unstable
Recall Rate	0.16	Very Small	607	86	244	The far below-required sample size cannot reliably assess

To the Analytics Manager

1. **Continue the Experiment with an Extended Sample Size**
   - ○ Based on the analysis of the results, we recommend extending the test to a larger population of loan officers. This will help validate the model’s effectiveness across different environments and reduce the risk of overfitting to specific experimental conditions. Expanding the Control group sample size and increasing Recall Rate data are essential to strengthen the analysis and support data-driven decision-making.

2. **Adjustments By Monitoring Long-Term Trends**
   - ○ Since AI-based decision-making can be affected by external factors like economic conditions or applicant trends, it needs to be reviewed quarterly to identify errors and make timely adjustments, ensuring data quality and proper model maintenance.

3. **Analyze Decision-Making Patterns**
   - ○ While the new model improves agreement rates, further investigation into decision revision behavior is needed. Specifically, determining why some loan

officers continue to revise decisions against AI recommendations will provide insights into model trust and usability issues.

4. **Segment Loan Officers by Performance**
    - Loan officers with consistently high error rates should be identified and given targeted training or interventions. Further analysis can reveal if certain officers' benefit more from AI assistance than others.

5. **Analyzing cases where AI still misclassifies bad loans:**
    - By studying instances where the AI failed to flag risky loans, improvements can be made to its risk assessment criteria.

To the Executive Team

1. **Phased Deployment**
    - The new AI model should first be introduced to 25% of loan officers in high-volume branches, allowing for controlled monitoring and officer adaptation. If the model continues to show improvement, a full-scale deployment can follow within 30 days to maximize impact.

2. **AI adoption and Optimization**
    - Implement training and feedback systems to enhance AI adoption and usability. E-learning, case studies, and real-time support will help loan officers apply AI insights confidently. A feedback loop will ensure continuous model improvements based on user insights and real-world challenges, enhancing decision alignment and trust.

3. **Cost-benefit analysis for Full Deployment**
    - Perform a financial impact assessment to compare the cost savings from reduced loan defaults with the expenses for implementation and training. Evaluate whether the AI system leads to a positive return on investment (ROI) before expanding its use company-wide.

4. **Monitoring Long-Term Trends**
    - Continuing the experiment longer (e.g., another 30~60 days) could help validate if these improvements hold consistently over time and strengthen the company’s overall risk management strategy. Additionally, gathering more data on specific cases where Type II errors persist would help refine the model further.

Conclusion

In conclusion, the new AI model demonstrates strong potential to improve loan review accuracy and efficiency, but additional refinements and long-term validation are required before full implementation. With continued experimentation and structured deployment, the AI system can enhance risk management, optimize loan approvals, and drive sustainable financial growth for the company.

.

Appendices

~~~
```{r 1, include=FALSE}
# Load necessary libraries
library(dplyr)
library(tidyr)
library(ggplot2)
library(Hmisc)
library(ggcorrplot)
library(gridExtra)
library(effectsize)
library(pwr)

# Load the dataset
data <- read.csv("ADA_project.csv")

# Check for missing values
missing_values <- colSums(is.na(data))
print(missing_values)
```


```{r 2, include=FALSE}
# Remove rows with missing values
data_cleaned <- data %>% filter(fully_complt >= 10)

# Verify that no missing values remain
print(colSums(is.na(data_cleaned)))

```


```{r 3, include=FALSE}
# Summary statistics
summary(data_cleaned)

#  Distribution of key metrics

p1 <- ggplot(data_cleaned, aes(x = typeI_fin)) +
~~~

```
  geom_histogram(binwidth = 1, fill = "blue", alpha=0.8) +
  ggtitle("Distribution of Type I Errors (Final)")

p2 <- ggplot(data_cleaned, aes(x = typeII_fin)) +
  geom_histogram(binwidth = 1, fill = "blue", alpha=0.8) +
  ggtitle("Distribution of Type II Errors (Final)")

p3 <- ggplot(data_cleaned, aes(x = agree_fin)) +
  geom_histogram(binwidth = 1, fill = "blue", alpha=0.8) +
  ggtitle("Distribution of Agreement (Final)")

p4 <- ggplot(data_cleaned, aes(x = revised_per_ai)) +
  geom_histogram(binwidth = 1, fill = "blue", alpha=0.8) +
  ggtitle("Distribution of Revised Count")

p5 <- ggplot(data_cleaned, aes(x = conflict_fin)) +
  geom_histogram(binwidth = 1, fill = "blue", alpha=0.8) +
  ggtitle("Distribution of Conflict Count (Final)")

print(p2)
grid.arrange(p1, p3, p4, p5, ncol = 2)


# Compare control vs. treatment groups
ggplot(data_cleaned, aes(x = typeI_fin, fill = Variant)) +
  geom_density(alpha = 0.5) +
  ggtitle("Density of Type I Errors Improvement by Variant") +
  theme_minimal()

ggplot(data_cleaned, aes(x = typeII_fin, fill = Variant)) +
  geom_density(alpha = 0.5) +
  ggtitle("Density of Type II Errors Improvement by Variant") +
  theme_minimal()
```

~~~
```{r 4, include=FALSE}
# Compare Type I errors improvement between control and treatment groups
t_test_typeI_impro <- t.test(typeI_fin-typeI_init ~ Variant, data = data_cleaned)
print(t_test_typeI_impro)


# Compare Type II errors improvement between control and treatment groups
t_test_typeII_impro <- t.test(typeII_fin-typeII_init ~ Variant, data = data_cleaned)
print(t_test_typeII_impro)
```



```{r 5, include=FALSE}

data_cleaned <- data_cleaned %>%
  mutate(
    # Agreement rate
    agree_rate = agree_fin / (agree_fin + conflict_fin),

    # Revision rate
    revision_follow_ai_rate = revised_per_ai / (revised_per_ai + revised_agst_ai),

    # Confidence level
    confidence_final_rate = confidence_fin_total / complt_fin,
  )

t.test(agree_rate ~ Variant, data = data_cleaned, var.equal = FALSE)
t.test(revision_follow_ai_rate ~ Variant, data = data_cleaned, var.equal = FALSE)
t.test(confidence_final_rate ~ Variant, data = data_cleaned, var.equal = FALSE)

```
~~~

~~~
```{r 6, include=FALSE}
data_cleaned <- data_cleaned %>%
  mutate(
    recall = badloans_num / (badloans_num + typeII_fin),
    precision = badloans_num / (badloans_num + typeI_fin),
    F1_score = 2 * (precision * recall) / (precision + recall)
  )
data_cleaned <- data_cleaned %>%
  mutate(
    recall = ifelse(is.na(recall) | is.infinite(recall), 0, recall),
    F1_score = ifelse(is.na(F1_score) | is.infinite(F1_score), 0, F1_score)
  )
t.test(recall ~ Variant, data = data_cleaned, var.equal = FALSE)
t.test(F1_score ~ Variant, data = data_cleaned, var.equal = FALSE)

```
~~~

~~~
```{r 7, include=FALSE}
# Aggregate data by day and variant
daily_summary <- data_cleaned %>%
  group_by(day, Variant) %>%
  summarise(
    mean_typeI = mean(typeI_fin-typeI_init),
    mean_typeII = mean(typeII_fin-typeII_init),
    #mean_confidence = mean(confidence_fin_total)
  )

# Plot trends over time
ggplot(daily_summary, aes(x = day, y = mean_typeI, color = Variant)) +
  geom_line() +
  ggtitle("Type I Errors Improvements Over Time") +
  xlab("Day") +
  ylab("Mean Type I Error Improvements") +
  theme_minimal()
~~~

~~~
ggplot(daily_summary, aes(x = day, y = mean_typeII, color = Variant)) +
  geom_line() +
  ggtitle("Type II Errors Improvements Over Time") +
  xlab("Day") +
  ylab("Mean Type II Error Improvements") +
  theme_minimal()

```




```{r 8, include=FALSE}
# Data Analysis: Compute & Interpret Effect Size (Cohen's d)
# Type II Error Improvement
cohens_d_typeII <- cohens_d(data_cleaned$typeI_fin-data_cleaned$typeI_init ~
data_cleaned$Variant)

# Type I Error Improvement
cohens_d_typeI <- cohens_d(data_cleaned$typeII_fin-data_cleaned$typeII_init ~
data_cleaned$Variant)

# F1 Score
cohens_d_F1 <- cohens_d(data_cleaned$F1_score ~ data_cleaned$Variant)



# Recall Rate
cohens_d_recall <- cohens_d(data_cleaned$recall ~ data_cleaned$Variant)

list(
  "Type II Error Improvement" = cohens_d_typeII,
  "Type I Error Improvement" = cohens_d_typeI,
  "F1 Score" = cohens_d_F1,
  "Recall Rate" = cohens_d_recall
)

```
~~~

~~~
```{r 9, include=FALSE}
# Compute Required Sample Size for Desired *Power* Level & Effect Size
library(pwr)

sample_size_typeII <- pwr.t.test(
  power = 0.8,
  d = cohens_d_typeII$Cohens_d,  # Cohen’s d
  sig.level = 0.05,
  type = "two.sample"
)

sample_size_typeI <- pwr.t.test(
  power = 0.8,
  d = cohens_d_typeI$Cohens_d,
  sig.level = 0.05,
  type = "two.sample"
)



sample_size_F1 <- pwr.t.test(
  power = 0.8,
  d = cohens_d_F1$Cohens_d,
  sig.level = 0.05,
  type = "two.sample"
)

sample_size_recall <- pwr.t.test(
  power = 0.8,
  d = cohens_d_recall$Cohens_d,
  sig.level = 0.05,
  type = "two.sample"
)



list(
  "Type II Error Improvement" = sample_size_typeII$n,
~~~

```
  "Type I Error Improvement" = sample_size_typeI$n,
  "F1 Score" = sample_size_F1$n,
  "Recall Rate" = sample_size_recall$n

)
table(data_cleaned$Variant)


```