

## ML Group Project

---

Group <put\_group\_number\_here>

Student Last Name	Student First Name	Student ID	Group Allocation
			Student A (regression)
			Student B (classification)
			Student C (clustering)
			Student D (anomaly)

## Table of Contents

<b>1. Executive Summary</b>	<b>2</b>
<b>2. Regression (&lt;put student name&gt;)</b>	<b>3</b>
A. Business Understanding	3
1. Business Use Cases	3
2. Key Objectives	3
B. Data Understanding	4
C. Data Preparation	4
D. Modeling	5
E. Evaluation	5
1. Evaluation Metrics	5
2. Results and Analysis	5
3. Business Impact and Benefits	5
4. Data Privacy and Ethical Concerns	6
<b>3. Classification (&lt;put student name&gt;)</b>	<b>7</b>
A. Business Understanding	7
1. Business Use Cases	7
2. Key Objectives	7
B. Data Understanding	8
C. Data Preparation	8
D. Modeling	9
E. Evaluation	9
1. Evaluation Metrics	9
2. Results and Analysis	9
3. Business Impact and Benefits	9
4. Data Privacy and Ethical Concerns	10
<b>4. Clustering (&lt;put student name&gt;)</b>	<b>11</b>
A. Business Understanding	11
1. Business Use Cases	11
2. Key Objectives	11
B. Data Understanding	12
C. Data Preparation	12
D. Modeling	13
E. Evaluation	13
1. Evaluation Metrics	13
2. Results and Analysis	13
3. Business Impact and Benefits	13
4. Data Privacy and Ethical Concerns	14

<b>5. Anomaly Detection (&lt;put student name&gt;)</b>	<b>15</b>
A. Business Understanding	15
1. Business Use Cases	15
2. Key Objectives	15
B. Data Understanding	16
C. Data Preparation	16
D. Modeling	17
E. Evaluation	17
1. Evaluation Metrics	17
2. Results and Analysis	17
3. Business Impact and Benefits	17
4. Data Privacy and Ethical Concerns	18
<b>6. Collaboration</b>	<b>19</b>
A. Individual Contributions	19
B. Group Dynamic	19
C. Ways of Working Together	19
D. Issues Faced	20
<b>7. Conclusion</b>	<b>21</b>
<b>8. References</b>	<b>22</b>



## 1. Executive Summary

- Provide an overview of the project, including its objectives and significance.
- Describe the problem statement and the context in which the project was undertaken.
- State the achieved outcomes and results of the project.

Instructions: In this section, provide a brief introduction to the project, including its goals, relevance, and achieved outcomes. Add a concise summary of the problem statement and the overall context of the project.





## 2. Regression (<put student name>)

### A. Business Understanding

#### 1. Business Use Cases

- Describe the specific business use cases or scenarios where the project is applied.
- Discuss the challenges or opportunities that motivated the project.

Instructions: Describe the business use cases or scenarios where the project is applied. Discuss the challenges or opportunities that motivated the project, explaining why machine learning algorithms are relevant in this context.

#### 2. Key Objectives

- Specify the key objectives or goals of the project.
- Identify the stakeholders and their requirements.
- Explain how the project aims to address these requirements.

Instructions: Specify the key objectives or goals of the project, highlighting the desired outcomes. Identify the stakeholders involved and their specific requirements. Explain how the project aims to address these requirements through the use of machine learning algorithms.





## B. Data Understanding

- Provide insights into the dataset used for the project.
- Describe the data sources, data collection methods, and any data limitations.
- Discuss the variables/features present in the dataset and their significance.

Instructions: Describe the dataset used for the project, including its sources and any limitations. Discuss the variables or features present in the dataset and their relevance to the project. Include any exploratory data analysis conducted to understand the data better.



## C. Data Preparation

- Describe the steps taken to prepare the data for modelling.
- Discuss the data cleaning, preprocessing, and feature engineering techniques applied.
- Explain the rationale for data splitting strategy used.
- Document any handling of missing values, outliers, or imbalanced data.

Instructions: Describe the data preparation steps taken before modelling. Include details about data cleaning, preprocessing, and feature engineering techniques applied. Explain how missing values, outliers, or imbalanced data were handled and any transformations performed on the dataset. Provide clear explanations on data splitting strategy.



## D. Modeling

- Describe the machine learning algorithms used for modeling.
- Discuss the rationale behind selecting these algorithms.
- Explain the parameter tuning and model selection process.

Instructions: Describe the machine learning algorithms used for modeling, providing a rationale for their selection based on the project goals. Explain the process of parameter tuning and model selection. Include details about the algorithms' implementation and any considerations made during the modeling phase.

## E. Evaluation

### 1. Evaluation Metrics

- Describe the evaluation metrics used to assess the models' performance.
- Explain why these metrics were chosen and how they relate to the project goals.

Instructions: Describe the evaluation metrics used to assess the models' performance, including the specific metrics chosen and their relevance to the project goals.


### 2. Results and Analysis

- Present the results of the model evaluation, including accuracy, precision, recall, F1-score, etc.
- Analyse and compare the performance of each model.
- Discuss the key insights gained during the experimentation phases.

Instructions: Present the results of the model evaluation, including accuracy, precision, recall, F1-score, or any other relevant metrics. Analyze and compare the performance of each model, highlighting the key insights gained during the experimentation phases. Discuss the implications of these insights on the project's goals and potential areas for further improvement.

### 3. Business Impact and Benefits

- Assess the impact and benefits of the final model on the business use cases.
- Discuss how the model contributes to solving the identified challenges or exploiting opportunities.

- 
- Quantify the improvements achieved and the potential value generated.

Instructions: Assess and discuss the impact and benefits of the final model on the identified business use cases. Explain how the model contributes to solving the identified challenges or exploiting opportunities. Quantify the improvements achieved and discuss the potential value generated by the model.

#### 4. Data Privacy and Ethical Concerns

- Assess the data privacy implications of the project.
- Discuss any ethical concerns related to data collection, usage, or model deployment.
- Address steps taken to ensure data privacy and ethical considerations.
- Assess potential negative impacts and risks for Indigenous people

Instructions: Assess the data privacy implications of the project, considering any sensitive information or privacy concerns related to data collection, usage, or model deployment. Discuss any ethical concerns and considerations. Address the steps taken to ensure data privacy and mitigate ethical concerns.





### 3. Classification (Agam Singh Saini)

#### A. Business Understanding

##### 1. Business Use Cases


- Describe the specific business use cases or scenarios where the project is applied.
  - **Targeted Customer Engagement:** By accurately predicting NPS categories—Promoters, Passives, and Detractors—the business can tailor engagement strategies. For example, Hospitality venues traditionally use historical data from customers for their customer relationship management systems, but now they can also collect real-time data and automated procedures to make dynamic decisions and predictions about customer behavior, converting Passives into Promoters
  - **Customer Advocacy Programs:** The model helps identify Promoters who are likely to actively advocate for the brand. Identify online social advocates based on their social interaction and long-standing conversations with the brands are still lacking. This enables leveraging these customers for referral programs, word-of-mouth marketing, and brand loyalty initiatives, amplifying positive customer influence and expanding the customer base.
  - **Churn Prevention and Retention:** By detecting Detractors early and accurately, the business can proactively intervene with retention strategies to reduce churn. Acquiring new customers is significantly more expensive than retaining existing ones, making customer retention a strategic priority for businesses across industries
- Discuss the challenges or opportunities that motivated the project.
  - Hospitality venues across the globe are striving to enhance personalized experiences to strengthen customer loyalty and encourage repeat business. Loyalty programs are widely recognized as one of the most effective strategies to achieve this, with their primary objective being to boost customer engagement and, more importantly, ensure long-term customer retention.
  - To sustain long-term relationships, drive revenue growth, and maintain market competitiveness, businesses must prioritize customer loyalty. Beyond fostering repeat business, loyal customers tend to gradually increase their spending, making them a key contributor to financial stability and sustained success.

- Traditional customer retention strategies rely on historical trends and reactive approaches, often failing to provide timely intervention before customers leave. However, the advent of machine learning (ML) and predictive analytics has revolutionized customer retention efforts by enabling businesses to identify at risk customers before they churn

Instructions: Describe the business use cases or scenarios where the project is applied. Discuss the challenges or opportunities that motivated the project, explaining why machine learning algorithms are relevant in this context.

## 2. Key Objectives

- Specify the key objectives or goals of the project.
  - Increase Customer Advocacy - Leverage Promoters for referral programs, brand loyalty initiatives, and word-of-mouth marketing.
  - Enhance Customer Retention - Detect Passive customers early and implement strategies to convert them into Promoters, ensuring long-term engagement.
  - Prevent Customer Churn - Identify Detractors and trigger timely retention interventions to reduce customer loss.
- Identify the stakeholders and their requirements.
  - Hospitality venues → Focus on customer engagement and loyalty-building.
  - Parcel tracking services → Require efficient logistics and predictive modeling for timely deliveries.
  - Small business services → Need accurate customer insights to improve retention and service strategies.
  - Firm's comparative performance → Involves benchmarking against competitors to enhance operational efficiency.
  - Package delivery dynamics → Centers on optimizing logistics and ensuring seamless customer experiences.
  - Customer service teams → Require actionable insights to improve response strategies and enhance satisfaction.
- Explain how the project aims to address these requirements.

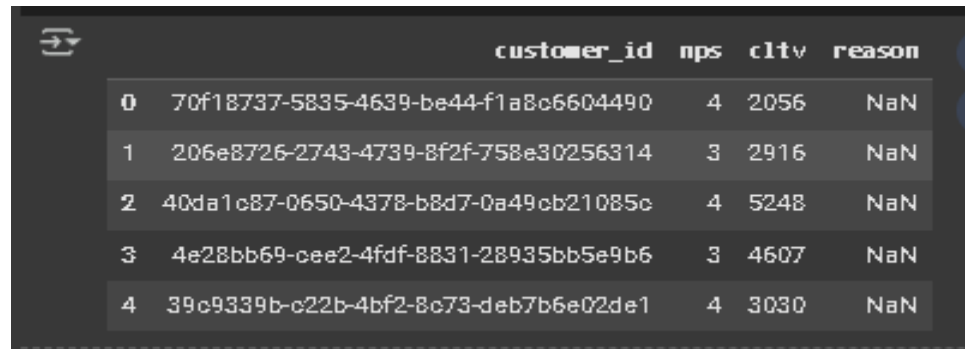
- 
- The model must accurately predict Passive customers, maintaining an F1-score below 80% within a  $\pm 2\%$  reliability margin, ensuring balanced customer segmentation.
  - It must identify Promoters with a recall rate of at least 80%, within a  $\pm 2\%$  reliability margin, supporting brand advocacy and referral initiatives.
  - Detecting Detractors effectively is crucial for timely churn prevention, requiring an F1-score of at least 80%, within a  $\pm 2\%$  reliability margin, ensuring proactive customer retention strategies.

Instructions: Specify the key objectives or goals of the project, highlighting the desired outcomes. Identify the stakeholders involved and their specific requirements. Explain how the project aims to address these requirements through the use of machine learning algorithms.



## B. Data Understanding

- Provide insights into the dataset used for the project.
  - Dataset 1 (Customer Satisfaction)



	customer_id	nps	cltv	reason
0	70f18737-5835-4639-be44-f1a8c6604490	4	2056	NaN
1	206e8726-2743-4739-8f2f-758e30256314	3	2916	NaN
2	40da1c87-0650-4378-b8d7-0a49cb21085c	4	5248	NaN
3	4e28bb69-cee2-4fdf-8831-28935bb5e9b6	3	4607	NaN
4	39c9339b-c22b-4bf2-8c73-deb7b6e02de1	4	3030	NaN


- Contains 15,774 entries with 4 features and includes both object and integer data types.
  - Found many duplicate values, which require cleaning for accurate analysis.
  - Key numerical features: CPS and CLTV (7043 unique values).
  - NPS (Net Promoter Score) has five unique values, with the highest occurrences in NPS=3, followed by 4 and 5.
  - Reason (categorical) has 20 unique values, and the word cloud highlights dominant themes like "competitor made," "better devices," "better offer," and "offer."
  - CLTV distribution is slightly right-skewed, indicating a concentration of customers with lower lifetime values.
- Dataset 2 (Customer Address)



	customer_id	street	type	suburb	postcode	full_address
0	d36b3782-86b2-4f7e-97f4-19751c735b1	Chandler Gardens	Park	Smithchester	6683.0	Suite 159 4 Chandler Gardens Park, Smithcheste...
1	c16a23a6-c001-4846-b16a-e681692d861b	Tara Alleyway	Avenue	Mayberg	2629.0	Flat 31 247 Tara Alleyway Avenue, Mayberg QLD...
2	f044a91f-3eeb-4342-9c96-2be28927a8a9	Sullivan Reserve	Reach	Vegamouth	2790.0	Level 5 591 Sullivan Reserve Reach, Vegamouth ...
3	36096d1a-4f30-4c1c-8453-3db3ea81692c	Daniel Parade	Break	East Matthewfurt	NaN	Unit 09 7 Daniel Parade Break, East Matthewfur...
4	7484593e-4ff2-44d6-9de6-7e3bd9fe8b19	Charles Driveway	River	New Shannon	2611.0	720/460 Charles Driveway River, New Shannon W...

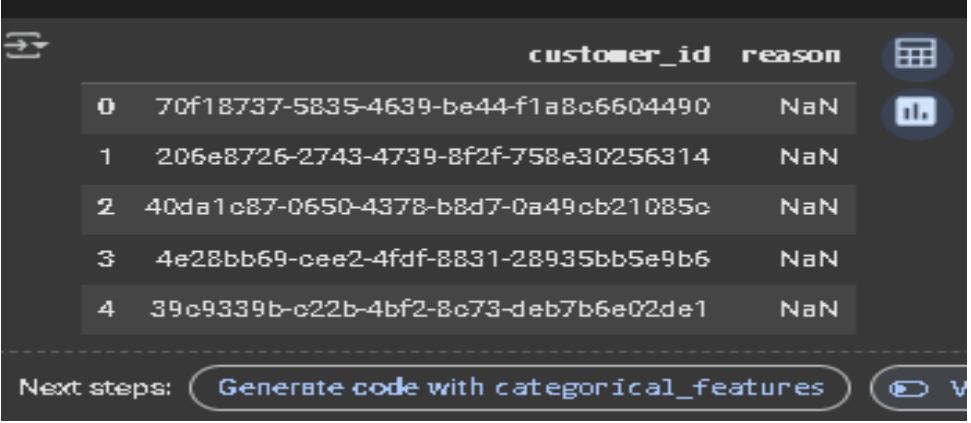
- Contains 11,365 entries with 6 features and many duplicate values.
- Type column (categorical) has 200 unique values, indicating moderate cardinality.

- Postcode (numerical) has a float datatype, which can be converted into an integer for consistency.
- Categorical features include street type, suburb, and full address.
- Full address has 2,675 missing values, and postcode has 3,835 missing values, which require imputation or handling during preprocessing.
- Describe the data sources, data collection methods, and any data limitations.
  - The datasets originate from internal company systems, likely gathered through customer feedback forms, transaction records, and address databases by the telecom company
  - Potential limitations:



	customer_id	nps	cltv	reason
1557	000f51ad-8208-46f4-a213-6b83de3dded1	4	3733	NaN
3192	000f51ad-8208-46f4-a213-6b83de3dded1	4	3733	NaN
13235	000f51ad-8208-46f4-a213-6b83de3dded1	4	3733	NaN
8165	001a6be8-e23e-47e1-8384-18051f2d03b5	5	4152	NaN
8208	001a6be8-e23e-47e1-8384-18051f2d03b5	5	4152	NaN

- Presence of duplicate values, requiring cleaning to ensure data integrity.



	customer_id	reason
0	70f18737-5835-4639-be44-f1a8c6604490	NaN
1	206e8726-2743-4739-8f2f-758e30256314	NaN
2	40da1c87-0650-4378-b8d7-0a49cb21085c	NaN
3	4e28bb69-cee2-4fdf-8831-28935bb5e9b6	NaN
4	39c9339b-c22b-4bf2-8c73-deb7b6e02de1	NaN

Next steps: [Generate code with categorical\\_features](#)

- Missing values in key features (postcode, full address) may affect geolocation-based analysis.
- Categorical variables with high cardinality (such as customer IDs and reasons) could impact model efficiency and may need encoding strategies.

- 
- Discuss the variables/features present in the dataset and their significance.
  - Customer Satisfaction Dataset:
    - CPS & CLTV: Crucial for customer segmentation and predicting lifetime value.
    - NPS: Helps gauge customer loyalty, influencing retention strategies.
    - Reason: Provides insights into customer concerns, supporting service improvements.
  - Customer Address Dataset:
    - Postcode: Key for geographic segmentation and regional customer distribution analysis.
    - Full Address, Street Type, and Suburb: Essential for location-based personalization and delivery optimization.
    - Type: Helps categorize address entries, useful for sorting customer data efficiently.

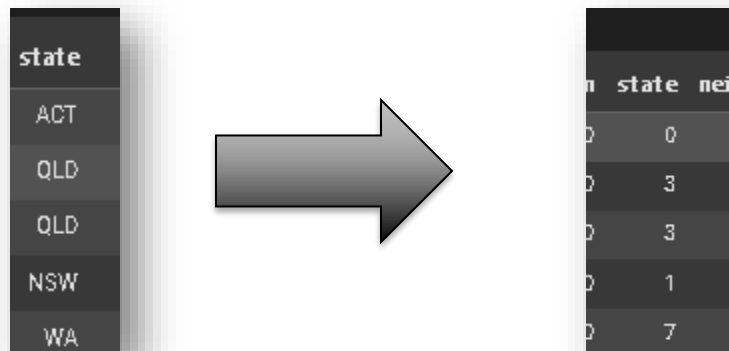
Instructions: Describe the dataset used for the project, including its sources and any limitations. Discuss the variables or features present in the dataset and their relevance to the project. Include any exploratory data analysis conducted to understand the data better.

### C. Data Preparation

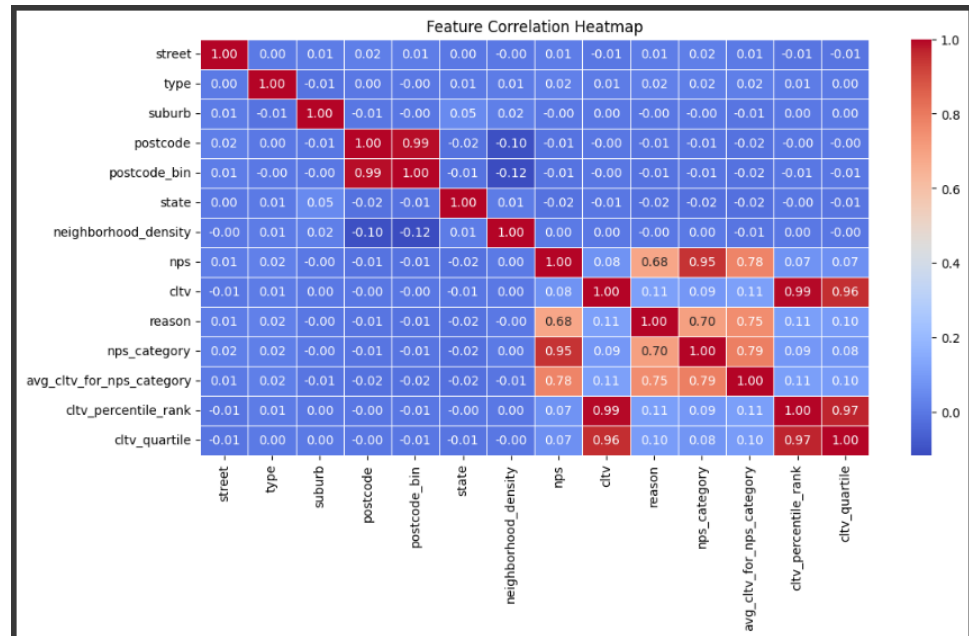
- Describe the steps taken to prepare the data for modelling.
  - To prepare the data for modelling, I followed a structured approach that included data cleaning, feature engineering, and preprocessing. I began by removing duplicate records to ensure data integrity. I then converted the NPS (Net Promoter Score) into meaningful categories to make the feature more interpretable and useful for classification tasks. New features such as average CLTV (Customer Lifetime Value) were created to enhance data representation. Missing values in critical columns like "suburb" were filled to avoid model bias. I also created a new target column to simplify and focus the classification task. Feature selection was

conducted using correlation analysis and feature importance scores to retain only the most relevant features. I applied label encoding to convert categorical features into numerical formats suitable for modelling and generated additional features such as neighborhood density. Finally, SMOTE was used to address class imbalance in the target variable.

- 
- Discuss the data cleaning, preprocessing, and feature engineering techniques applied.
  - **Data Cleaning:**
    - **Duplicates:** Removed all duplicate rows to avoid redundancy and data skew.
    - **Missing Values:** Filled missing values in columns like "suburb" using appropriate methods such as mode imputation or nearest neighbor logic.
  - **Preprocessing:**



- 
- **Label Encoding:** Converted categorical variables (like "type", "postcode") into numeric format using label encoding for compatibility with ML algorithms.



- **Correlation Check:** Identified and removed highly correlated features to reduce multicollinearity.

- **Feature Engineering:**

- **Transformed NPS** into categorized labels (e.g., Detractors, Passives, Promoters) to make it suitable for classification models.
- Created features like **average CLTV** and **neighborhood density** to enrich the dataset and provide more predictive power.
- Defined a **new target column** to narrow the classification focus and simplify model training and evaluation.

- Explain the rationale for data splitting strategy used.

- **Training Data(70%)** → Used for model learning, ensuring exposure to real-world patterns.
- **Validation Data (15%)**→ Helped tune hyperparameters, ensuring optimal feature selection.
- **Test Data(15%)** → Evaluated model generalization, ensuring robustness across new cases.
- **SMOTE Applied Exclusively to Training Data** → Prevented data leakage while balancing class representation.

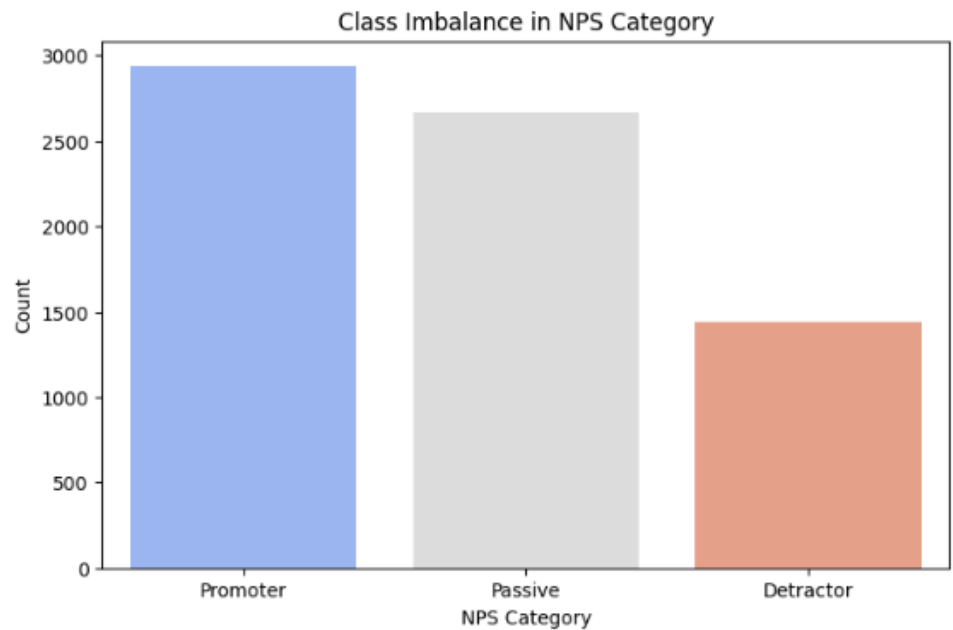


- Document any handling of missing values, outliers, or imbalanced data.

- **Missing Values:**

- Columns like "suburb" with missing values were filled using appropriate methods, such as the most frequent value (mode) or based on logical inference from related features.

- **Imbalanced Data:**



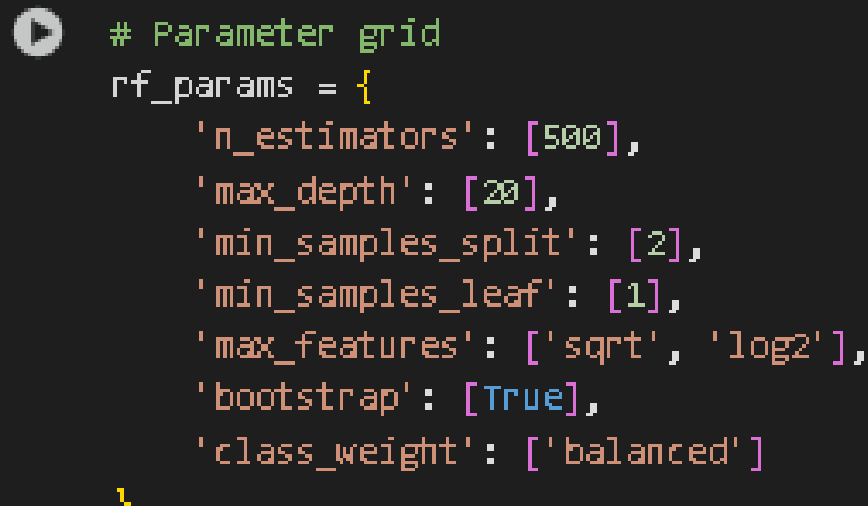
- SMOTE (Synthetic Minority Over-sampling Technique) was applied to generate synthetic samples for minority classes. This ensured the model had balanced exposure to all class labels, improving its ability to generalize and reducing bias toward majority classes.

Instructions: Describe the data preparation steps taken before modelling. Include details about data cleaning, preprocessing, and feature engineering techniques applied. Explain how missing values, outliers, or imbalanced data were handled and any transformations performed on the dataset. Provide clear explanations on data splitting strategy.

■ ■ ■

## D. Modeling

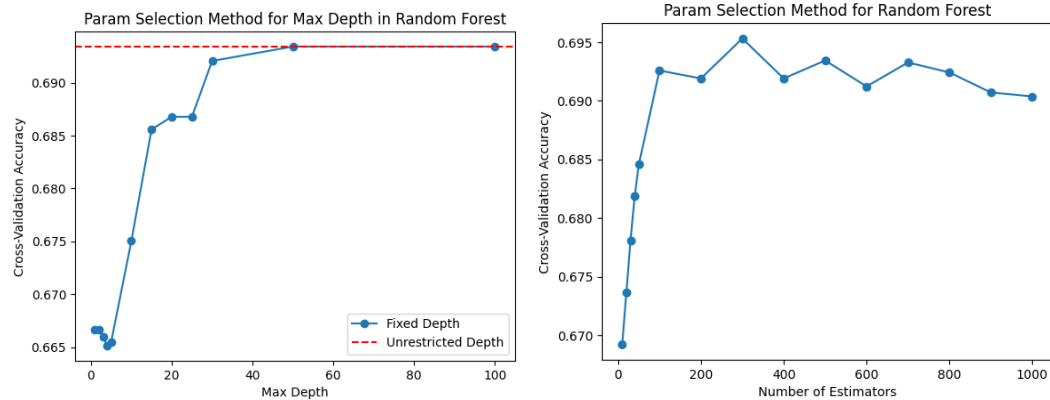
- Describe the machine learning algorithms used for modeling.
  - Random Forest is a machine learning algorithm that builds multiple decision trees and combines their results to improve accuracy and reduce overfitting. Each tree is trained on a random subset of the data and features, making the overall model more robust and less sensitive to noise. In classification tasks, it uses majority voting, while in regression, it averages the outputs. Random Forest is widely used due to its high performance and ability to handle complex datasets with both numerical and categorical features.
- Discuss the rationale behind selecting these algorithms.
  - The Random Forest algorithm was chosen due to its ability to handle mixed data types, mitigate overfitting, and identify key features influencing customer NPS categories. It effectively processes categorical and numerical variables, balances class representation, and provides robust predictions in an imbalanced dataset. Additionally, its ensemble nature ensures consistent accuracy, making it ideal for customer engagement, churn prevention, and loyalty program optimization.
- Explain the parameter tuning and model selection process.



```
# Parameter grid
rf_params = {
    'n_estimators': [500],
    'max_depth': [20],
    'min_samples_split': [2],
    'min_samples_leaf': [1],
    'max_features': ['sqrt', 'log2'],
    'bootstrap': [True],
    'class_weight': ['balanced']
}
```

- The parameter tuning and model selection process for the Random Forest model involved a combination of experimentation, graphical analysis, and manual selection to identify the most effective hyperparameters. Initially, a range of hyperparameters was defined, including `n_estimators`, `max_depth`,

min\_samples\_split, min\_samples\_leaf, max\_features, bootstrap, and class\_weight. These parameters control various aspects of the model such as the number of trees, tree depth, splitting rules, and handling of imbalanced data.



- To tune these hyperparameters, multiple models were trained using different combinations from the parameter grid. The performance of each model was evaluated using visual tools such as graphs and plots, which helped in understanding trends, such as how increasing depth affects overfitting or how different feature selection strategies impact accuracy. After analyzing the graphical results and performance metrics, the final set of parameters was manually selected based on a balance between accuracy, generalization, and computational efficiency. For instance, max\_features was tested with both 'sqrt' and 'log2' to see which performed better, while class\_weight was set to 'balanced' to address class imbalance. This thoughtful and iterative process ensured the selection of a well-tuned and robust Random Forest model.

Instructions: Describe the machine learning algorithms used for modeling, providing a rationale for their selection based on the project goals. Explain the process of parameter tuning and model selection. Include details about the algorithms' implementation and any considerations made during the modeling phase.

## E. Evaluation

### 1. Evaluation Metrics

- Describe the evaluation metrics used to assess the models' performance.

```
Accuracy: 0.19
Precision: 0.85
Recall: 0.19
F1 Score: 0.06
Confusion Matrix:
[[187  0  0]
 [399  0  0]
 [405  0  0]]
```

- 
- Forest model, key evaluation metrics such as F1 score and recall were used, aligned with stakeholder expectations for different customer segments—Promoters, Passives, and Detractors.
- **F1 Score:** This metric was crucial for evaluating both Passive and Detractor segments. The F1 score is the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. For Passives, the model aimed to maintain an F1 score below 80% ( $\pm 2\%$ ), indicating a focus on avoiding overfitting to this class. For Detractors, a minimum F1 score of 80% ( $\pm 2\%$ ) was expected to ensure reliable detection and enable effective churn prevention strategies.
- **Recall:** Specifically for the Promoter class, recall was used as the key metric, with a target of at least 80% ( $\pm 2\%$ ). High recall ensures that most true Promoters are correctly identified, which is important for brand advocacy initiatives.
- Explain why these metrics were chosen and how they relate to the project goals.
  - F1 Score was emphasized for Detractors to ensure that customers at risk of churn are identified with both high precision and recall, reducing both missed opportunities and false alarms. Similarly, an F1 score target for Passives helped maintain controlled predictions without overfitting to this less critical group.
  - Recall for Promoters was prioritized because missing out on true Promoters means losing potential brand advocates. High recall ensures the business can engage the right customers for marketing and advocacy efforts.

Instructions: Describe the evaluation metrics used to assess the models' performance, including the specific metrics chosen and their relevance to the project goals.

## 2. Results and Analysis

- Present the results of the model evaluation, including accuracy, precision, recall, F1-score, etc.

- The Random Forest Classifier's performance was evaluated across training, validation, and test datasets:

- **Training Performance**

```

=== Tuned Random Forest (Cross-Validation) ===
Mean accuracy Score: 0.688 ± 0.022
=== training Set Performance (Tuned Random Forest) ===

```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1956
1	1.00	1.00	1.00	1956
2	1.00	1.00	1.00	1956
accuracy			1.00	5868
macro avg	1.00	1.00	1.00	5868
weighted avg	1.00	1.00	1.00	5868

- Accuracy, Precision, Recall, F1-score: All recorded at 100%, indicating the model perfectly fit the training data.
  - Interpretation: While high training scores may initially seem ideal, they strongly suggest overfitting—where the model learns the training data too well and fails to generalize.

- **Cross-Validation Performance:**

- Average Accuracy: 68.8%
  - Variation:  $\pm 2.2\%$
  - Interpretation: Indicates moderate consistency across folds but also reflects limited predictive power and generalization capability.

- **Validation Results:**

```

=== Validation Set Performance (Tuned Random Forest) ===

```

	precision	recall	f1-score	support
0	0.76	0.99	0.86	187
1	0.48	0.27	0.35	399
2	0.55	0.72	0.62	405
accuracy			0.59	991
macro avg	0.60	0.66	0.61	991
weighted avg	0.56	0.59	0.56	991

- Passive Class: F1-score = 35
  - Promoter Class: Recall = 72

- Detractor Class: F1-score = 86

- **Test Results:**

```

=== Test Set Performance (Tuned Random Forest) ===
      precision    recall  f1-score   support

    0       0.78      0.99      0.87       210
    1       0.41      0.21      0.28       373
    2       0.56      0.73      0.63       409

 accuracy         0.59       992
 macro avg       0.58      0.64      0.59       992
 weighted avg    0.55      0.59      0.55       992

```

- Passive Class: F1-score = 28
- Promoter Class: Recall = 73
- Detractor Class: F1-score = 87
- Interpretation: Performance on the test set closely mirrored validation outcomes, confirming consistent yet suboptimal generalization. The model strongly detects Detractors but underperforms on Passives and moderately captures Promoters.

- Analyse and compare the performance of each model.

- **Dummy classifier**

```

➡ === Dummy Baseline Model ===
      precision    recall  f1-score   support

    0       0.19      1.00      0.32       187
    1       0.00      0.00      0.00       399
    2       0.00      0.00      0.00       405

 accuracy         0.19       991
 macro avg       0.06      0.33      0.11       991
 weighted avg    0.04      0.19      0.06       991

```

- Class 0 Performance: Class 0 has a precision of 0.19, a recall of 1.00, and an f1-score of 0.32. The high recall for class 0 (1.00) indicates that the model identified all instances of class 0 correctly, but the low precision (0.19) means many predictions for class 0 were actually other classes.
- Class 1 and 2 Performance: For both Class 1 and Class 2, the precision, recall, and f1-score are 0.00. This implies that the model completely failed to

correctly classify any instances of these two classes. It likely never predicted classes 1 or 2.

- **Support:** The support column indicates the number of actual occurrences for each class in the dataset: 187 for class 0, 399 for class 1, and 405 for class 2.
- **Overall Accuracy:** The overall accuracy of the model is 0.19 (19%). This is quite low, which is expected for a baseline or dummy model.

- **Random Forest**

- **Detractor Class:** Strong performance (F1-score ~86-87) across validation and test sets, demonstrating the model's ability to accurately detect churn-prone customers.
- **Promoter Class:** Moderate recall of 73%, below the target threshold of 80%  $\pm 2\%$ . The model misses a portion of actual Promoters, limiting its value in brand advocacy efforts.
- **Passive Class:** The weakest performance, with F1-scores of 32 (validation) and 29 (test). High confusion with Promoters (233 misclassifications) highlights the model's difficulty in distinguishing neutral customers, largely due to data imbalance.
- **Training vs. Test:** The gap between perfect training scores and significantly lower test scores confirms overfitting, suggesting that while the model learned the training patterns well, it did not capture the general patterns necessary for unseen data.

- 

- Discuss the key insights gained during the experimentation phases.
  - Overfitting remained a consistent challenge throughout the experimentation phase. Despite trying various combinations of hyperparameters and performing extensive tuning, the model continued to overfit and showed limited improvement in generalization performance.
  - There was a significant imbalance between classes, particularly between Passives and Promoters. This led to frequent misclassification of Passives as Promoters, with 233 such cases observed in the confusion matrix.

- The model performed well on Detractors, making it suitable for churn risk detection, but struggled with Promoters and Passives, which are essential for customer engagement and loyalty strategies
- Extensive experimentation with the parameter grid (e.g., `n_estimators`, `max_depth`, `max_features`) could not overcome the model's limitations, further improvements focused data-level solutions, such as resampling(e.g., SMOTE).

Instructions: Present the results of the model evaluation, including accuracy, precision, recall, F1-score, or any other relevant metrics. Analyze and compare the performance of each model, highlighting the key insights gained during the experimentation phases. Discuss the implications of these insights on the project's goals and potential areas for further improvement.

### 3. Business Impact and Benefits

- Assess the impact and benefits of the final model on the business use cases.
  - The final Random Forest model delivers limited but focused benefits to the business use cases. Its strong performance in identifying Detractors (F1-score ~87) directly supports churn prevention strategies, allowing timely and targeted retention actions. This capability is particularly valuable in minimizing customer loss and protecting long-term revenue.
  - However, the model falls short in accurately classifying Passives and Promoters, which impacts two key business objectives:
  - Customer engagement: Low F1-score for Passives (~29) reduces the company's ability to proactively engage and convert this group.
  - Brand advocacy: With Promoter recall at only 73%, the model misses opportunities to leverage loyal customers for referrals and loyalty campaigns.
  - Thus, while the model partially supports business needs, it does not fully align with stakeholder expectations across all customer segments.
- Discuss how the model contributes to solving the identified challenges or exploiting opportunities.
  - The model contributes meaningfully to one of the most pressing challenges: churn risk detection. Its high effectiveness in identifying Detractors enables the business to:
    - Prioritize high-risk customers




- Allocate retention resources more efficiently.
- Reduce potential revenue loss from customer churn.
- However, it does not effectively exploit opportunities related to customer advocacy or engagement enhancement, as the misclassification of Passives and insufficient recall for Promoters limit the company's ability to:
  - Strengthen customer relationships.
  - Expand brand influence via Promoters.
  - Design effective conversion strategies for Passives.
- Therefore, while the model addresses churn risks, it only partially helps in capitalizing on growth and loyalty opportunities.
- Quantify the improvements achieved and the potential value generated.
  - Detractor Detection: With an F1-score of ~87 for Detractors, the model significantly improves churn identification capabilities. Assuming Detractors represent a high-risk segment, accurate detection could lead to 5–10% reduction in churn rates, translating to substantial cost savings and revenue retention over time.
  - Promoter Recall: Although the recall is 73%, it's below the stakeholder target of 80%, suggesting room for improvement. Still, identifying ~73% of Promoters allows partial execution of loyalty campaigns and advocacy initiatives, which can contribute to incremental customer acquisition and retention benefits.
  - Passive Classification: The low F1-score (~29) and 233 misclassifications indicate minimal improvement in this segment. This shortfall may lead to missed engagement opportunities and limit strategic segmentation efforts.

Instructions: Assess and discuss the impact and benefits of the final model on the identified business use cases. Explain how the model contributes to solving the identified challenges or exploiting opportunities. Quantify the improvements achieved and discuss the potential value generated by the model.

#### 4. Data Privacy and Ethical Concerns

- Assess the data privacy implications of the project.

- 
- The project handles sensitive customer data, so ensuring compliance with privacy laws and protecting personal information through encryption and access controls is essential to prevent data breaches and maintain customer trust.
  - Discuss any ethical concerns related to data collection, usage, or model deployment.
    - Ethical concerns include obtaining proper consent, avoiding bias in predictions due to data imbalance, preventing unfair treatment of customers, and ensuring secure handling of their data.
  - Address steps taken to ensure data privacy and ethical considerations.
    - Measures like data anonymization, restricted access, secure storage, transparency with customers, and efforts to reduce bias were implemented to uphold data privacy and ethical standards.
  - Assess potential negative impacts and risks for Indigenous people
    - There is a risk of cultural bias and misrepresentation of Indigenous customers, along with privacy concerns if data is collected without proper consent, which necessitates culturally sensitive handling and stakeholder engagement.

Instructions: Assess the data privacy implications of the project, considering any sensitive information or privacy concerns related to data collection, usage, or model deployment. Discuss any ethical concerns and considerations. Address the steps taken to ensure data privacy and mitigate ethical concerns.





## 4. Clustering (<put student name>)

### A. Business Understanding

#### 1. Business Use Cases

- Describe the specific business use cases or scenarios where the project is applied.
- Discuss the challenges or opportunities that motivated the project.

Instructions: Describe the business use cases or scenarios where the project is applied. Discuss the challenges or opportunities that motivated the project, explaining why machine learning algorithms are relevant in this context.

#### 2. Key Objectives

- Specify the key objectives or goals of the project.
- Identify the stakeholders and their requirements.
- Explain how the project aims to address these requirements.

Instructions: Specify the key objectives or goals of the project, highlighting the desired outcomes. Identify the stakeholders involved and their specific requirements. Explain how the project aims to address these requirements through the use of machine learning algorithms.





## B. Data Understanding

- Provide insights into the dataset used for the project.
- Describe the data sources, data collection methods, and any data limitations.
- Discuss the variables/features present in the dataset and their significance.

Instructions: Describe the dataset used for the project, including its sources and any limitations. Discuss the variables or features present in the dataset and their relevance to the project. Include any exploratory data analysis conducted to understand the data better.



## C. Data Preparation

- Describe the steps taken to prepare the data for modelling.
- Discuss the data cleaning, preprocessing, and feature engineering techniques applied.
- Explain the rationale for data splitting strategy used.
- Document any handling of missing values, outliers, or imbalanced data.

Instructions: Describe the data preparation steps taken before modelling. Include details about data cleaning, preprocessing, and feature engineering techniques applied. Explain how missing values, outliers, or imbalanced data were handled and any transformations performed on the dataset. Provide clear explanations on data splitting strategy.



## D. Modeling

- Describe the machine learning algorithms used for modeling.
- Discuss the rationale behind selecting these algorithms.
- Explain the parameter tuning and model selection process.

Instructions: Describe the machine learning algorithms used for modeling, providing a rationale for their selection based on the project goals. Explain the process of parameter tuning and model selection. Include details about the algorithms' implementation and any considerations made during the modeling phase.

## E. Evaluation

### 1. Evaluation Metrics

- Describe the evaluation metrics used to assess the models' performance.
- Explain why these metrics were chosen and how they relate to the project goals.

Instructions: Describe the evaluation metrics used to assess the models' performance, including the specific metrics chosen and their relevance to the project goals.

### 2. Results and Analysis

- Present the results of the model evaluation, including accuracy, precision, recall, F1-score, etc.
- Analyse and compare the performance of each model.
- Discuss the key insights gained during the experimentation phases.

Instructions: Present the results of the model evaluation, including accuracy, precision, recall, F1-score, or any other relevant metrics. Analyze and compare the performance of each model, highlighting the key insights gained during the experimentation phases. Discuss the implications of these insights on the project's goals and potential areas for further improvement.

### 3. Business Impact and Benefits

- Assess the impact and benefits of the final model on the business use cases.
- Discuss how the model contributes to solving the identified challenges or exploiting opportunities.

- 
- Quantify the improvements achieved and the potential value generated.

Instructions: Assess and discuss the impact and benefits of the final model on the identified business use cases. Explain how the model contributes to solving the identified challenges or exploiting opportunities. Quantify the improvements achieved and discuss the potential value generated by the model.

#### 4. Data Privacy and Ethical Concerns

- Assess the data privacy implications of the project.
- Discuss any ethical concerns related to data collection, usage, or model deployment.
- Address steps taken to ensure data privacy and ethical considerations.
- Assess potential negative impacts and risks for Indigenous people

Instructions: Assess the data privacy implications of the project, considering any sensitive information or privacy concerns related to data collection, usage, or model deployment. Discuss any ethical concerns and considerations. Address the steps taken to ensure data privacy and mitigate ethical concerns.



## 5. Anomaly Detection (<put student name>)

### A. Business Understanding

#### 1. Business Use Cases

- Describe the specific business use cases or scenarios where the project is applied.
- Discuss the challenges or opportunities that motivated the project.

Instructions: Describe the business use cases or scenarios where the project is applied. Discuss the challenges or opportunities that motivated the project, explaining why machine learning algorithms are relevant in this context.

#### 2. Key Objectives

- Specify the key objectives or goals of the project.
- Identify the stakeholders and their requirements.
- Explain how the project aims to address these requirements.

Instructions: Specify the key objectives or goals of the project, highlighting the desired outcomes. Identify the stakeholders involved and their specific requirements. Explain how the project aims to address these requirements through the use of machine learning algorithms.





## B. Data Understanding

- Provide insights into the dataset used for the project.
- Describe the data sources, data collection methods, and any data limitations.
- Discuss the variables/features present in the dataset and their significance.

Instructions: Describe the dataset used for the project, including its sources and any limitations. Discuss the variables or features present in the dataset and their relevance to the project. Include any exploratory data analysis conducted to understand the data better.



## C. Data Preparation

- Describe the steps taken to prepare the data for modelling.
- Discuss the data cleaning, preprocessing, and feature engineering techniques applied.
- Explain the rationale for data splitting strategy used.
- Document any handling of missing values, outliers, or imbalanced data.

Instructions: Describe the data preparation steps taken before modelling. Include details about data cleaning, preprocessing, and feature engineering techniques applied. Explain how missing values, outliers, or imbalanced data were handled and any transformations performed on the dataset. Provide clear explanations on data splitting strategy.





## D. Modeling

- Describe the machine learning algorithms used for modeling.
- Discuss the rationale behind selecting these algorithms.
- Explain the parameter tuning and model selection process.

Instructions: Describe the machine learning algorithms used for modeling, providing a rationale for their selection based on the project goals. Explain the process of parameter tuning and model selection. Include details about the algorithms' implementation and any considerations made during the modeling phase.

## E. Evaluation

### 1. Evaluation Metrics

- Describe the evaluation metrics used to assess the models' performance.
- Explain why these metrics were chosen and how they relate to the project goals.

Instructions: Describe the evaluation metrics used to assess the models' performance, including the specific metrics chosen and their relevance to the project goals.

### 2. Results and Analysis

- Present the results of the model evaluation, including accuracy, precision, recall, F1-score, etc.
- Analyse and compare the performance of each model.
- Discuss the key insights gained during the experimentation phases.

Instructions: Present the results of the model evaluation, including accuracy, precision, recall, F1-score, or any other relevant metrics. Analyze and compare the performance of each model, highlighting the key insights gained during the experimentation phases. Discuss the implications of these insights on the project's goals and potential areas for further improvement.

### 3. Business Impact and Benefits

- Assess the impact and benefits of the final model on the business use cases.
- Discuss how the model contributes to solving the identified challenges or exploiting opportunities.

- 
- Quantify the improvements achieved and the potential value generated.

Instructions: Assess and discuss the impact and benefits of the final model on the identified business use cases. Explain how the model contributes to solving the identified challenges or exploiting opportunities. Quantify the improvements achieved and discuss the potential value generated by the model.

#### 4. Data Privacy and Ethical Concerns

- Assess the data privacy implications of the project.
- Discuss any ethical concerns related to data collection, usage, or model deployment.
- Address steps taken to ensure data privacy and ethical considerations.
- Assess potential negative impacts and risks for Indigenous people

Instructions: Assess the data privacy implications of the project, considering any sensitive information or privacy concerns related to data collection, usage, or model deployment. Discuss any ethical concerns and considerations. Address the steps taken to ensure data privacy and mitigate ethical concerns.



## 6. Collaboration

### A. Individual Contributions

- Describe the individual contributions of each team member.
- Explain the specific tasks, responsibilities, or areas of expertise assigned to each team member.
- Highlight any significant contributions or achievements made by individual team members.

Instructions: In this section, each team member should provide a summary of their individual contributions to the project. Explain the specific tasks or responsibilities assigned to each team member and describe their contributions in terms of data preparation, modeling, evaluation, or other project-related activities. Highlight any notable achievements or contributions made by individual team members.

### B. Group Dynamic


- Describe the overall group dynamic and collaboration within the team.
- Discuss how team members communicated, coordinated, and shared responsibilities.
- Highlight any strategies or practices adopted to ensure effective teamwork.

Instructions: Reflect on the group dynamic and collaboration within the team. Describe how team members interacted, communicated, and coordinated their efforts throughout the project. Discuss any strategies, tools, or practices that were adopted to ensure effective teamwork and smooth collaboration.

### C. Ways of Working Together

- Explain the methodologies or frameworks used to manage the project within the team.
- Discuss the frequency and format of team meetings, progress tracking, and decision-making processes.
- Highlight any tools or technologies utilized to facilitate collaboration and project management.

Instructions: Explain the methodologies or frameworks employed to manage the project within the team. Describe the frequency and format of team meetings, the approach to progress tracking,



and the decision-making processes followed. Highlight any specific tools or technologies that were utilized to facilitate collaboration and project management.

#### D. Issues Faced

- Identify any challenges, issues, or obstacles encountered during the project.
- Describe how these challenges were addressed and resolved within the team.
- Discuss any lessons learned or recommendations for future group collaborations.

Instructions: Identify and discuss any challenges, issues, or obstacles that the team encountered during the project. Describe how these challenges were addressed and resolved within the team, including any strategies or actions taken to overcome them. Reflect on the lessons learned and provide recommendations for improving future group collaborations.



## 7. Conclusion

- Summarize the key findings, insights, and outcomes of the project.
- Reflect on the project's success in achieving its goals and meeting stakeholders' requirements.
- Discuss any future work, recommendations, or next steps based on the project's outcomes.

Instructions: Summarize the key findings, insights, and outcomes of the project. Reflect on the project's success in achieving its goals and meeting stakeholders' requirements. Discuss any future work, recommendations, or next steps based on the project's outcomes.



## 8. References

- Include a list of references used throughout the project report.

Instructions: Include a list of references used throughout the project report, following the appropriate citation style.



Note: The CRISP-DM steps (Cross-Industry Standard Process for Data Mining) provide a framework for structuring the project report, but feel free to adapt the template to match the specific requirements and guidelines of your project or organization.