Project Report Template

Please include at least the following content in your project executive summary.

A. Project goal and objective

Project Goal-

The goal of your project is likely to make sense of the data—to figure out what it's telling you and how you can use that information to achieve something meaningful.

Project Objectives

1.Analyze Outreach Data: Review trends in outreach activities and their effectiveness.

2. Optimize Opens: Understand patterns in email opens and engagement to improve communication strategies.

3. Evaluate Performance Metrics: Use KPIs to assess the success of outreach and engagement efforts.

4. Provide Recommendations: Based on analysis, propose actionable strategies to improve project outcomes.

B1. Data Description

Outreach summary:

- Time Sent: Time when the emails were sent
- Campaign Name: Name of the campaign Sends: Number of total emails sent
- Opens: Number of emails opened
- Open Rate: Ratio of opened emails to the email received by the company (Total emails sent - emails bounced)
- Mobile Open Rate: Rate of emails opened on a mobile
- Desktop Open Rate: Rate of emails opened on a desktop
- Clicks: Number of emails clicked on
- Click Rate: Rate of clicks
- Bounces: Number of emails bounced
- Bounce Rate: Rate of emails bounced to the email received by the company (Total emails sent - emails bounced)
- Unsubscribes: Number of emails unsubscribed
- Unsubscribe Rate: Rate of unsubscription

Sent Emails:

- Campaign Name: Name of the campaign
- Company: Name of the company
- Email status: To check if the company unsubscribed (Active or Unsubcribed)
- Email permission status: If the permission is granted (Implied or not)
- Confirmed Opt-Out Date: Date of unsubscribing
- Confirmed Opt-Out Reason: Reason for unsubscribing
- City - Home: City of the company
- State/Province - Home: State of the company
- Country - Home: Country of the companies (United States)
- Website: Website of the company
- Title: Position in the company email sent to
- Created At: Time email was created
- Updated At: Time the data was updated
- Sent At: Time the email was sent

Opens Emails:

- Campaign Name: Name of the campaign
- Company: Name of the company
- Email status: To check if the company unsubscribed (Active or Unsubcribed)
- Email permission status: If the permission is granted (Implied or not)
- Confirmed Opt-Out Date: Date of unsubscribing
- Confirmed Opt-Out Reason: Reason for unsubscribing
- City - Home: City of the company
- State/Province - Home: State of the company
- Country - Home: Country of the companies (United States)
- Website: Website of the company
- Title: Position in the company email sent to
- Created At: Time email was created
- Updated At: Time the data was updated
- Opened At: Time the email was opened -

Company Master Data:

- - **company_name**
- - **city_stevens**
- - **city_thomasnet**
- - **city_whitestone**
- - **company_age_whitestone**
- - **company_name_solar**
- - **company_name_stevens**
- - **company_name_thomasnet**
- - **company_name_whitestone**

- - **company_revenue_($m)_stevens**
- - **company_revenue_($m)_thomasnet**
- - **company_revenue_($m)_whitestone**
- - **contact_title_stevens**
- - **contact_title_whitestone**
- - **founded_year_thomasnet**
- - **headcount_range_whitestone**
- - **headcount_stevens**
- - **headcount_thomasnet**
- - **headcount_whitestone**
- - **major_products,_capabilities_stevens**
- - **major_products,_capabilities_whitestone**
- - **revenue_range_whitestone**
- - **state_stevens**
- - **state_thomasnet**
- - **state_whitestone**
- - **zip_code_stevens**
- - **zip_code_whitestone**
- - **companytype_thomasnet**
- - **description_thomasnet**

Merged

- Campaign Name
- Company
- City - Home
- Opened
- Mobile open Rate

B1. How did you clean and prepare your data? Please also provide the first 5 rows including the headers of your dataset.

Preprocessing for Outreach Summary, Opens Email and Sent Emails:

Handling Unsubscribed Emails:

- Removed rows where Email status was marked as "Unsubscribed" in both open and sent.

Dropping Unnecessary Columns:

- From outreach: Dropped Sends, Opens, Open Rate, Clicks, Click Rate, Bounces, Bounce Rate, Unsubscribes, Unsubscribe Rate
    - These attributes were deleted because they did not contribute to the objective, making it important to remove unnecessary columns like Bounces, Bounce Rate, Unsubscribes, and Unsubscribe Rate.
    - Sends, Opens, Open Rate, Clicks, and Click Rate were also removed as they did not explain why specific emails were opened. These were recorded after the emails were opened, and the goal was to narrow down the dependent variables as much as possible.
- From open and sent: Dropped Email status, Email permission status, Confirmed Opt-Out Date, Confirmed Opt-Out Reason, Website, Title, Country - Home.
    - Email status, Email permission status, Confirmed Opt-Out Date, and Confirmed Opt-Out Reason columns were unnecessary because companies marked as "Unsubscribed" had already been removed.
    - Website and Title columns were dropped because they did not provide insight into why specific emails were being opened. Additionally, these columns contained missing values, further justifying their removal.
    - Country - Home was redundant since all entries listed the USA, providing no variability or additional value to the analysis.

Merging Open emails and Sent emails:

- Performed a left merge using Campaign Name and Company as keys.
- Duplicates in both datasets were removed before merging based on key attributes like City - Home, State/Province - Home, Created At, Updated At, and timestamp columns.
- Following the merging of the Open and Sent datasets on Company and Campaign Name, unnecessary columns like City - Home_y, State/Province - Home_y, Created At_y, and Updated At_y were removed to eliminate redundancy. Columns from the left dataset were retained for consistency.

Changing Opened At column:

- Converted the Opened At column into a binary indicator:

        - 1: Email was opened.
        - 0: Email was not opened (default for unmatched rows during the merge from Sent Emails dataset).

Merging with Outreach Dataset

- Open_Sent dataset was further merged with the Outreach dataset using the Campaign Name column. This step integrated additional campaign details, enabling deeper insights into the email engagement data.

Datetime Processing:

- The Time Sent column was divided into two separate columns: Date Sent and Time Sent. This change allows for more detailed analysis of email engagement by specific dates or times of the day.

Standardizing State Names:

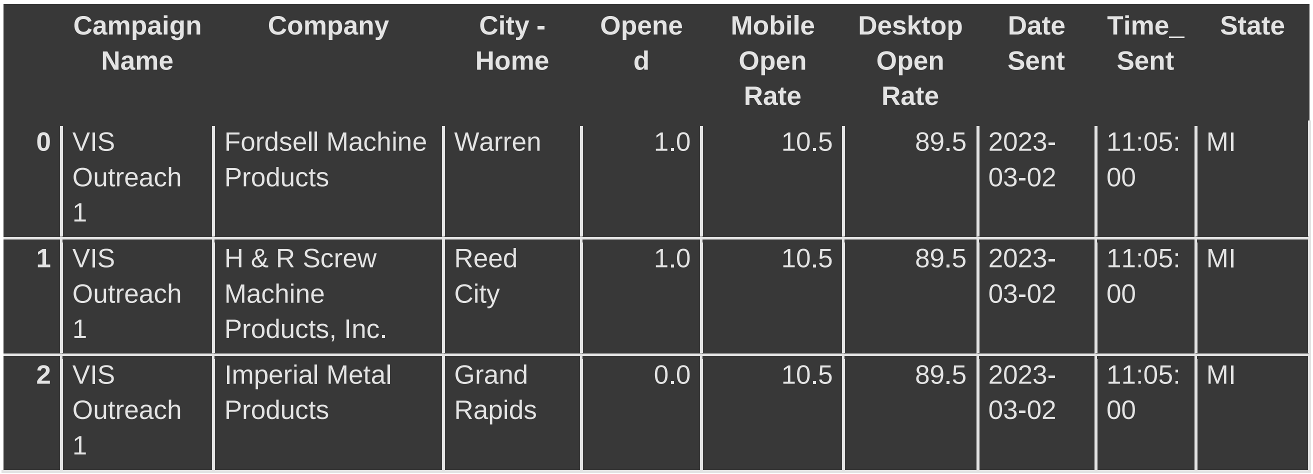
- State/Province - Home column values were mapped to standard state abbreviations (e.g., "Minnesota" to "MN") using a dictionary. The original column was then removed for consistency.

Fixing Rates Data

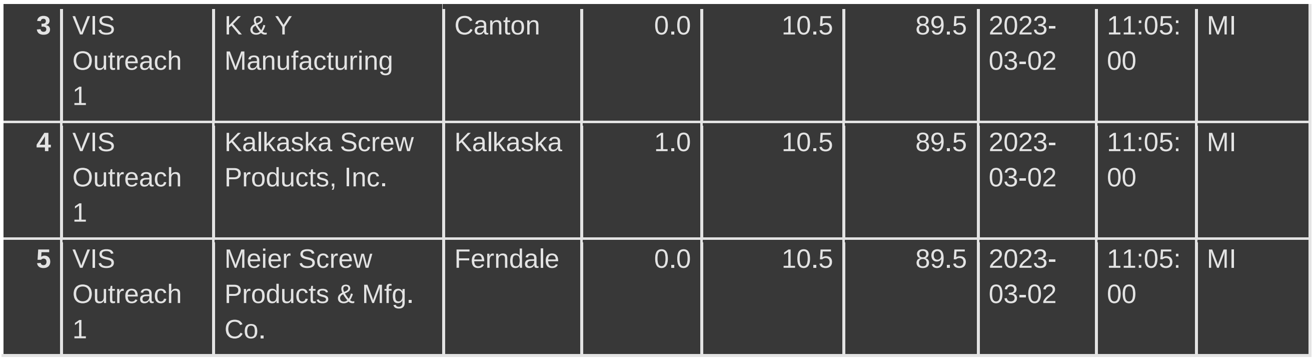
- Mobile Open Rate and Desktop Open Rate columns were cleaned by removing percentage symbols (%) and converting the values into numeric format. This ensures the data is ready for quantitative analysis.

Final Export

- The processed dataset was saved as Email_outreach.csv for further analysis.



	Campaign Name	Company	City - Home	Opened	Mobile Open Rate	Desktop Open Rate	Date Sent	Time_ Sent	State
0	VIS Outreach 1	Fordsell Machine Products	Warren	1.0	10.5	89.5	2023-03-02	11:05:00	MI
1	VIS Outreach 1	H & R Screw Machine Products, Inc.	Reed City	1.0	10.5	89.5	2023-03-02	11:05:00	MI
2	VIS Outreach 1	Imperial Metal Products	Grand Rapids	0.0	10.5	89.5	2023-03-02	11:05:00	MI



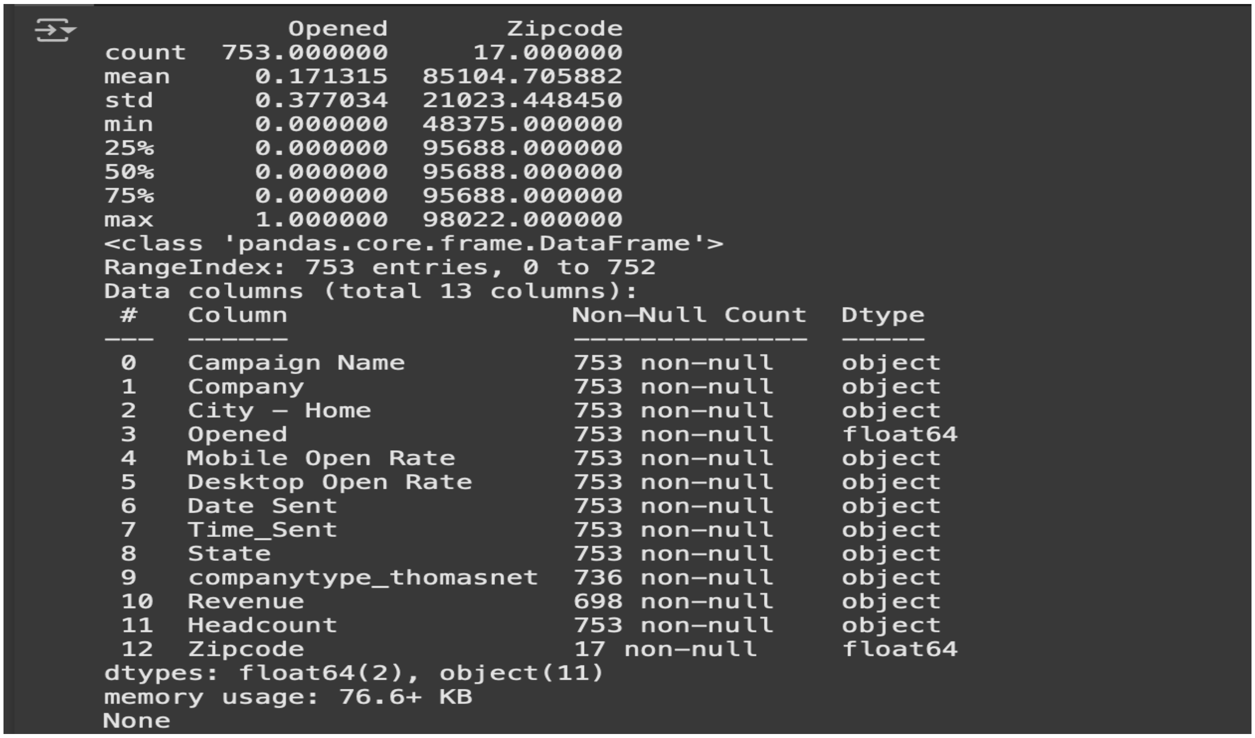
3	VIS Outreach 1	K & Y Manufacturing	Canton	0.0	10.5	89.5	2023-03-02	11:05:00	MI
4	VIS Outreach 1	Kalkaska Screw Products, Inc.	Kalkaska	1.0	10.5	89.5	2023-03-02	11:05:00	MI
5	VIS Outreach 1	Meier Screw Products & Mfg. Co.	Ferndale	0.0	10.5	89.5	2023-03-02	11:05:00	MI

B2. What is the dependent/output variable? What are the independent/input variables? Why do you choose these variables?

Dependent variable: Opened column which is binary and will tell us if the email was opened or not.

B3. Please provide the statistics table of your cleaned data. Below is an example, yours can be a little different.

Preprocessing for Outreach Summary, Opens Email and Sent Emails:



```
             Opened       Zipcode
count  753.000000     17.000000
mean     0.171315  85104.705882
std      0.377034  21023.448450
min      0.000000  48375.000000
25%      0.000000  95688.000000
50%      0.000000  95688.000000
75%      0.000000  95688.000000
max      1.000000  98022.000000
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 753 entries, 0 to 752
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Campaign Name          753 non-null    object 
 1   Company                753 non-null    object 
 2   City - Home            753 non-null    object 
 3   Opened                 753 non-null    float64
 4   Mobile Open Rate       753 non-null    object 
 5   Desktop Open Rate      753 non-null    object 
 6   Date Sent              753 non-null    object 
 7   Time_Sent              753 non-null    object 
 8   State                  753 non-null    object 
 9   companytype_thomasnet  736 non-null    object 
 10  Revenue                698 non-null    object 
 11  Headcount              753 non-null    object 
 12  Zipcode                17 non-null     float64
dtypes: float64(2), object(11)
memory usage: 76.6+ KB
None
```

C. Analyses and Result

C1. What models are you using for this project and why?

We are using four different machine learning models for this project:

1. **Random Forest:** This is a robust machine learning algorithm widely used for classification and regression tasks. It works by building a large number of decision trees during training. It is favored for accuracy as it has the ability to handle large dataset and reduced overfitting.

2. **Gradient Boosting Machine**: Unlike Random Forest (where trees are built independently), GBM builds trees sequentially, learning from the errors of previous trees. Each new tree focuses on correcting mistakes made by earlier models. GBM minimizes the loss function (error) using gradient descent to improve accuracy with each iteration.

3. **Support Vector Machine (SVM)**: This is a powerful model that can be used for both classification and regression problems. We chose SVM because it is a robust model that can handle high-dimensional data and non-linear relationships.
4. **Decision Tree**: This is a simple and interpretable model that can be used for both classification and regression problems. We chose the decision tree because it is a widely used model that can provide a baseline for comparison with other models.

We chose these models because they are all widely used and well-established in the field of machine learning, and they provide a good balance of simplicity, interpretability, and performance.

C2. What are the results? Which model works better

The results are as follows:

- Random Forest: Classification Report:
  - Precision: 0.84 (class 0), 0.60 (class 1)
  - Recall: 0.99 (class 0), 0.09 (class 1)
  - F1-score: 0.91 (class 0), 0.15 (class 1)
  - Accuracy: 0.83

- Gradient Boosting Machine: Classification Report:
  - Precision: 0.84 (class 0), 0.60 (class 1)
  - Recall: 0.99 (class 0), 0.09 (class 1)

    - F1-score: 0.91 (class 0), 0.15 (class 1)
    - Accuracy: 0.83

- SVM: Classification Report:
    - Precision: 0.83 (class 0), 0.00 (class 1)
    - Recall: 1.00 (class 0), 0.00 (class 1)
    - F1-score: 0.91 (class 0), 0.00 (class 1)
    - Accuracy: 0.83

- Decision Tree: Classification Report:
    - Precision: 0.84 (class 0), 0.60 (class 1)
    - Recall: 0.99 (class 0), 0.09 (class 1)
    - F1-score: 0.91 (class 0), 0.15 (class 1)
    - Accuracy: 0.83

Based on the results, it appears that the Gradient Boosting machine, Random Forest, and Decision Tree Regression models are similar in terms of accuracy of 0.83. However, the ROC-AUC Score is the highest for Gradient Boosting Machine as 0.6307. That's the reason we are moving further with Gradient Boosting Machine model.

D. Discussion

<u>D1. What are the implications of your analyses and results?</u>

Classification Problem:

1. **Imbalanced data**: The classification report shows that the models are biased towards class 0, with a high precision and recall for this class, but a low precision and recall for class 1. This suggests that the data is imbalanced, with more instances of class 0 than class 1.
2. **Limited generalizability**: The models' performance on the test data may not generalize well to new, unseen data, especially for class 1. This is because the models are not well-represented by the minority class.
3. **Need for class weighting or oversampling**: To improve the models' performance on class 1, it may be necessary to use class weighting or oversampling techniques to balance the data.

Overall Implications:

1. **Data quality issues**: The results suggest that there may be issues with the quality of the data, such as imbalanced classes or outliers. These issues need to be addressed before further modeling can be done.
2. **Model selection**: The results highlight the importance of selecting the right model for the problem at hand. Different models may be more suitable for different problems, and the choice of model should be based on the specific characteristics of the data and the problem.
3. **Hyperparameter tuning**: The results suggest that hyperparameter tuning may be necessary to improve the performance of the models. This can be done using techniques such as grid search or random search.

D2. What are some potential ways to improve your models in the future?

Improvements:

1. **Data preprocessing**: Improve data preprocessing techniques, such as handling missing values or outliers, to ensure that the data is clean and ready for modeling.
2. **Model interpretability**: Use techniques such as feature importance or partial dependence plots to improve model interpretability and understand how the models are making predictions.
3. **Model validation**: Use techniques such as walk-forward optimization or backtesting to validate the models' performance on unseen data.
4. **Continuous learning**: Continuously collect new data and update the models to ensure that they remain accurate and effective over time.
5. **Transfer learning**: Use transfer learning techniques to leverage pre-trained models and improve the models' performance on related problems.
6. **Reinforcement learning**: Use reinforcement learning techniques to improve the models' performance on problems with sequential decision-making.

D3. What are your main takeaways of this project?

Takeaways:

1. **Data quality is crucial**: Imbalanced classes, outliers, and missing values impact model performance.
2. **Model selection is key**: Choose the right model for the problem.
3. **Hyperparameter tuning is essential**: Optimize model performance through tuning.
4. **Ensemble methods improve performance**: Combine multiple models for better results.
5. **Model interpretability is important**: Understand how models make predictions.

D4. What are your suggestions/feedback on how to improve the class experience of this course?

Suggestions for Improving the Class Experience:
.

1. **Real-World Case Studies**: Use real-world case studies to illustrate key concepts and make the material more engaging and relevant.
2. **Online Resources**: Provide additional online resources, such as video tutorials and online forums, to supplement the course material.
3. **Project-Based Learning**: Consider incorporating project-based learning, where students work on a project throughout the course and receive feedback and guidance from the instructor.

Feedback:

- The course material was comprehensive and well-organized.
- The instructor was knowledgeable and provided clear explanations.
- The assignments and projects were challenging but helped to reinforce the material.
- Consider providing more opportunities for students to ask questions and receive feedback.