JDS Capstone: Target Market Analysis

Group 4
Ahmad Azraai
Luqman Hakim
Mohd Hidzam

Objective

1. Design a predictive model to determine the potential customers who will purchase if you send the advertisement . The target variable is Potential_Customer.

2. Calculate the value and the revenue of your model. Fit your model on train set. Assume among the customers on your test set we only send advertisement to those your model predicted as Class1 and we ignore the rest. From the data you can calculate the average Cust_Last_Purchase for those who are in the train set and had the last purchase (Cust_Last_Purchase>0) . Assume sending advertisement to each customer costs 5$ and the average purchase you calculated on the train set remains the same for the test set. Calculate the value of your models to choose the best model.

3. Compare your best model's revenue with the revenue of the default solution which is sending advertisement to all the customers in X_test. Which solution would you choose?

4. Assume the next time you want to target a group of 30,000 customers similar to this group. And assume the purchase rate is 10 which means 10 out of 100 people who receive the advertisement will purchase the product. Also assume your model will have the same Precision and Recall for Class1 . Will you send the advertisement to everyone, or you use one of the models you have already created?

- calculate your model's revenue on this set of 30,000 customers based on the above assumptions

- calculate the revenue of the default model: send advertisement to everyone

Data Cleaning

- Remove $ and , sign, drop duplicates
- Change datatype each column to categorical or numerical
- Drop C_ID (not useful for analysis)

Dimensions of data after cleaning : (3618, 25)

First 5 records data

	Potential_Customer	Cust_Last_Purchase	Pur_3_years	Pur_5_years	Pur_3_years_Indirect	Pur_5_years_Indirect	Pur_latest	Pur_3_years_Avg
0	1	5.0	2.0	17.0	2.0	4.0	0.0	7.50
1	1	30.0	1.0	7.0	0.0	3.0	25.0	25.00
2	0	0.0	5.0	12.0	3.0	5.0	15.0	15.00
3	1	20.0	1.0	11.0	0.0	3.0	20.0	20.00
4	1	5.0	3.0	15.0	2.0	7.0	3.0	4.33

Exploratory Data Analysis (EDA)

Explore Categorical Variables
-How many categories in each categorical variables?
-What proportion/percentage from each category?

Gender

Category	Count of Gender
F	53.12
M	42.29
U	4.59

*Count typo should be percentage %

Potential_Customer

Category	Count of Potential_Customer
0	52.02
1	47.98

Status_Cust

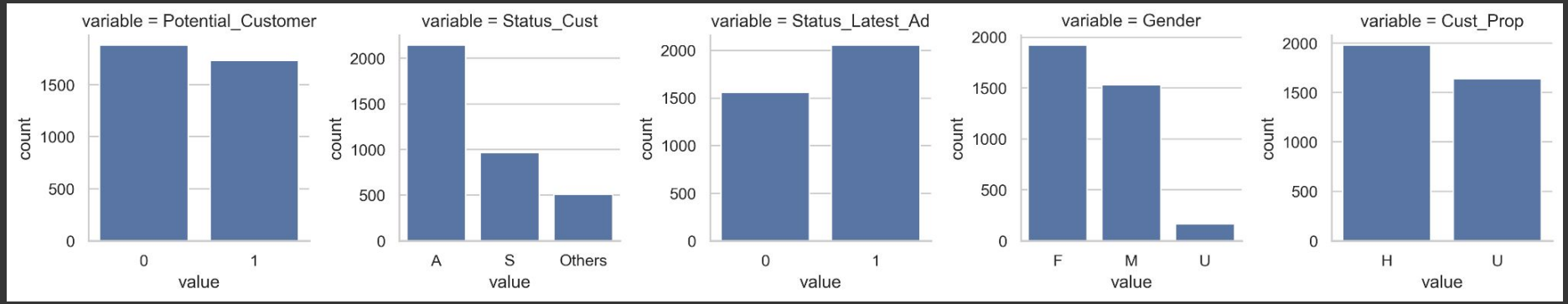
Category	Count of Status_Cust
A	59.31
S	26.64
Others	14.04

Status_Latest_Ad

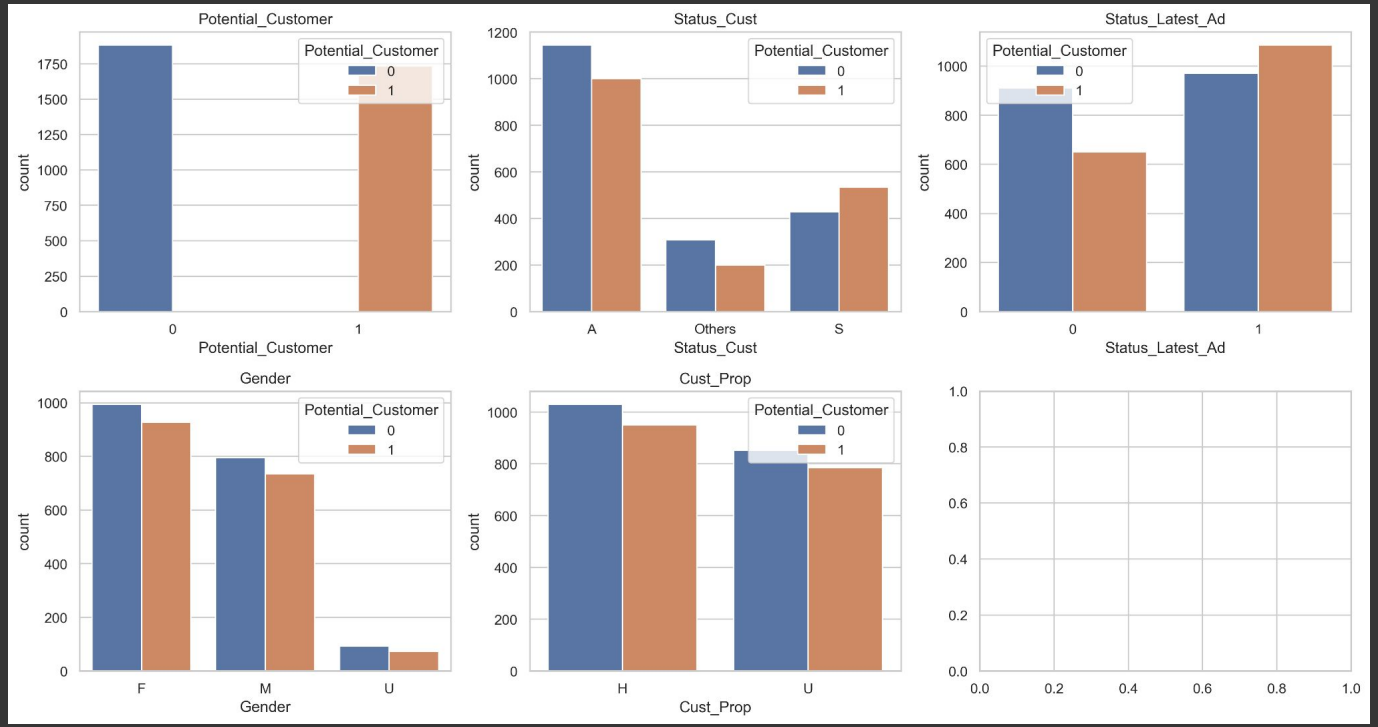
Category	Count of Status_Latest_Ad
1	56.85
0	43.15

Cust_Prop

Category	Count of Cust_Prop
H	54.75
U	45.25



Explore Relationship Between Categorical & Target Variable. Interpret the observation

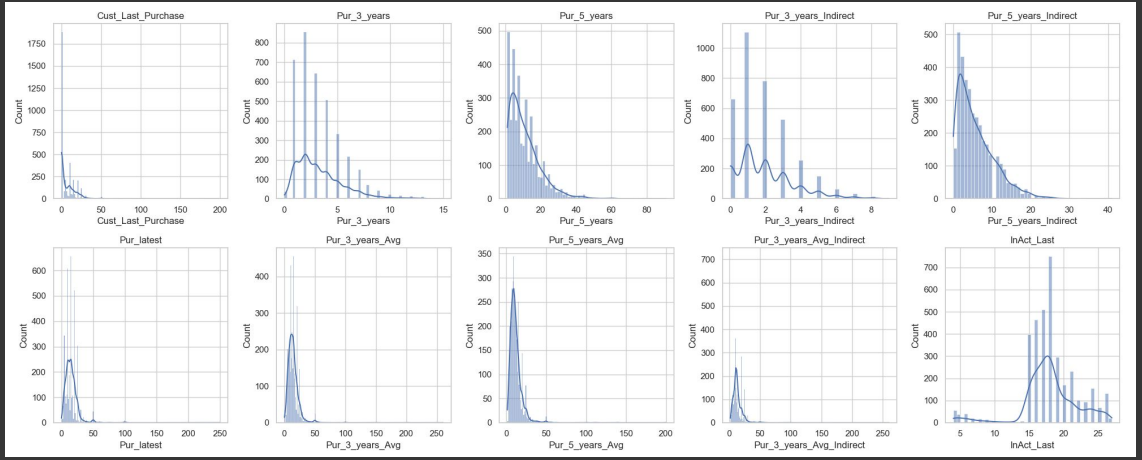


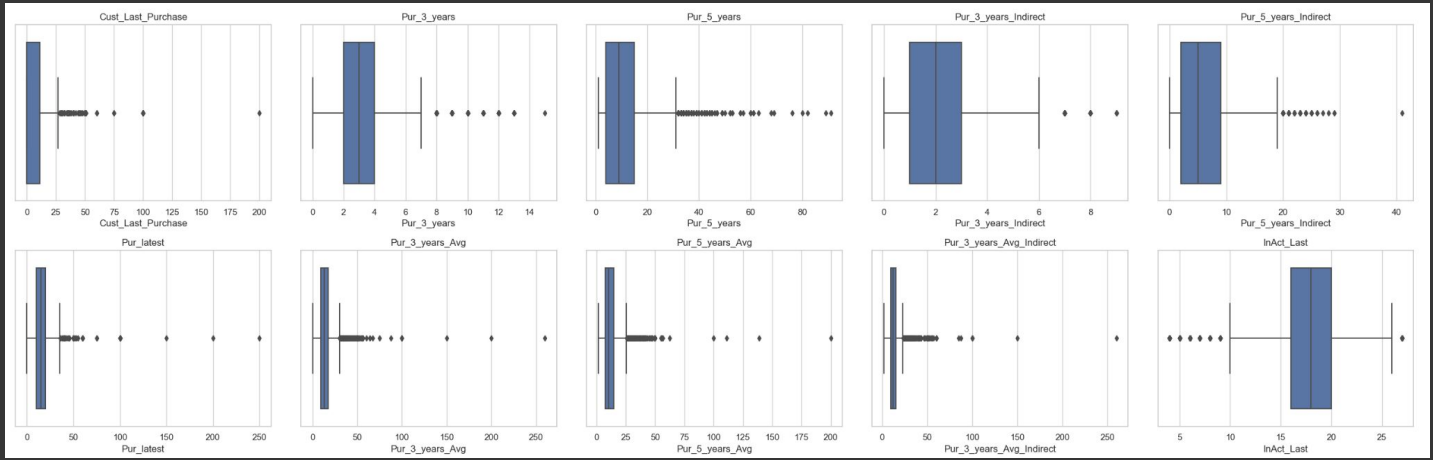
Insight

Status Customer (S) has relationship with Potential Customer

Status Latest Ad (1) has relationship with Potential Customer

Explore Numerical Variables





Insight

Based on histogram, the data has skewness and not normally distributed

Based on boxplot, there are outliers in the data

Explore the Relationship between the columns and try to answer the following questions:

Is there any significant difference between men/women's salary?

-p-value: 0.002949686546935098

-There is a significant difference between men and women's salaries.

Is there any significant difference between men/women's number of the purchase in the last three years?

-p-value: 0.053421072225418 28

-There is no significant difference between men and women's number of purchases in the last three years.

Is there any significant difference between men/women's average purchase in the last three years?

-p-value: 0.029119656654305 51

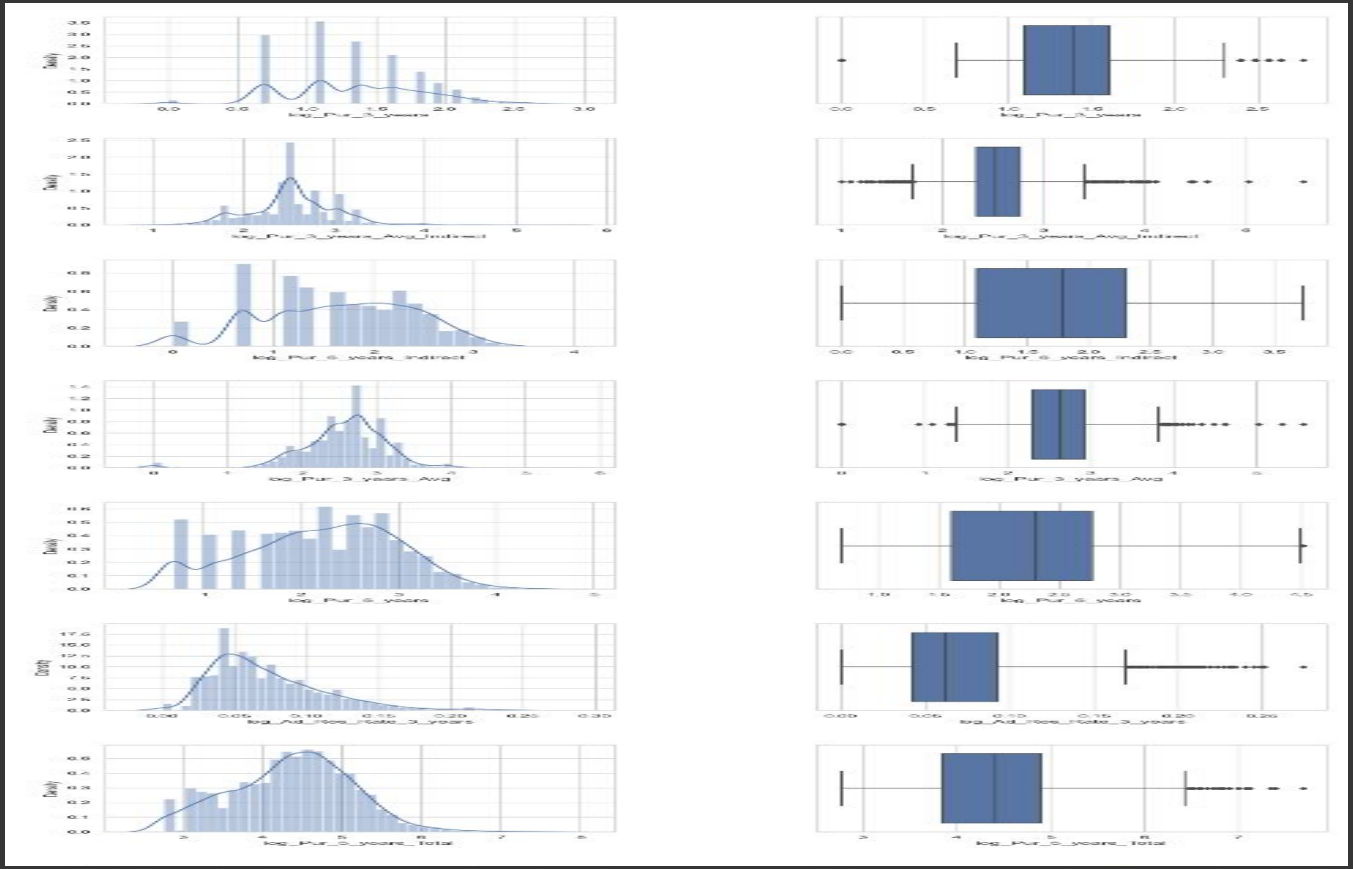
-There is a significant difference between men and women's average purchase in the last three years.

Is there any significant difference between men/women's total purchase in the last three years?

-p-value: 0.33404528691822555

-There is no significant difference between men and women's total purchase in the last three years.

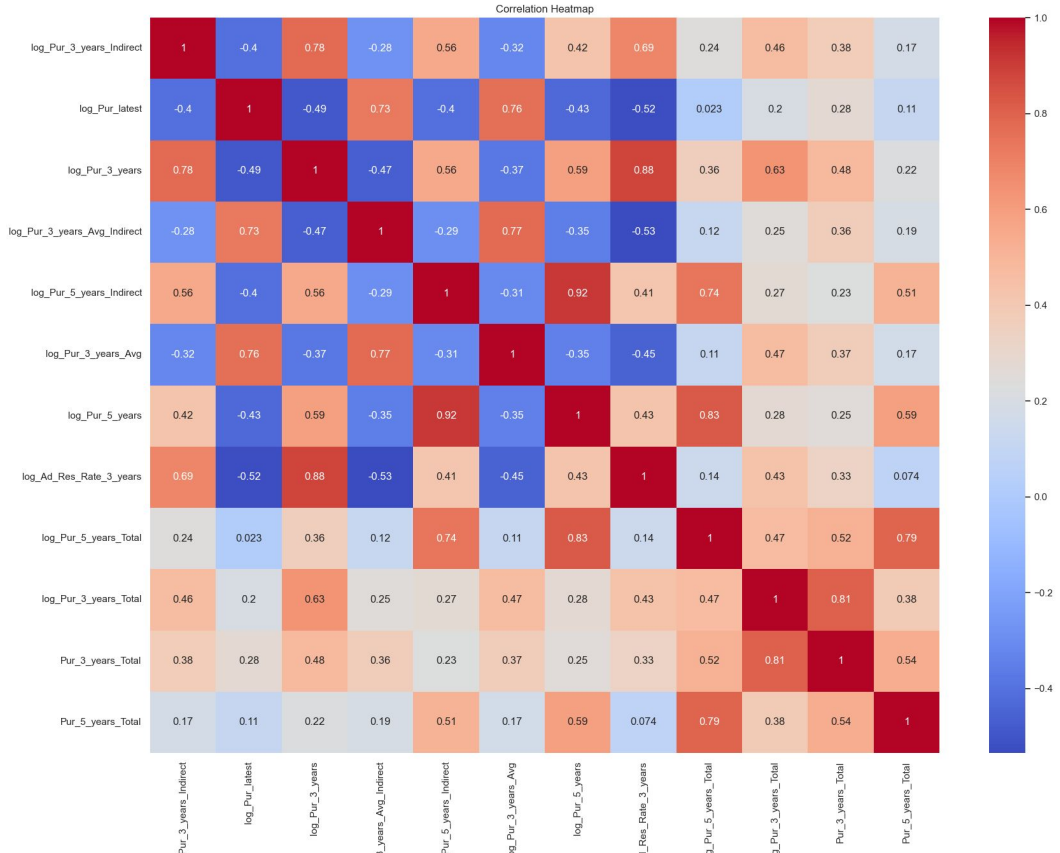
Add Some High Level Features and explore their relationship with the target variable



Insight

We use logarithm to normalize data/dimension reduction

Check Correlation between Numerical Variables

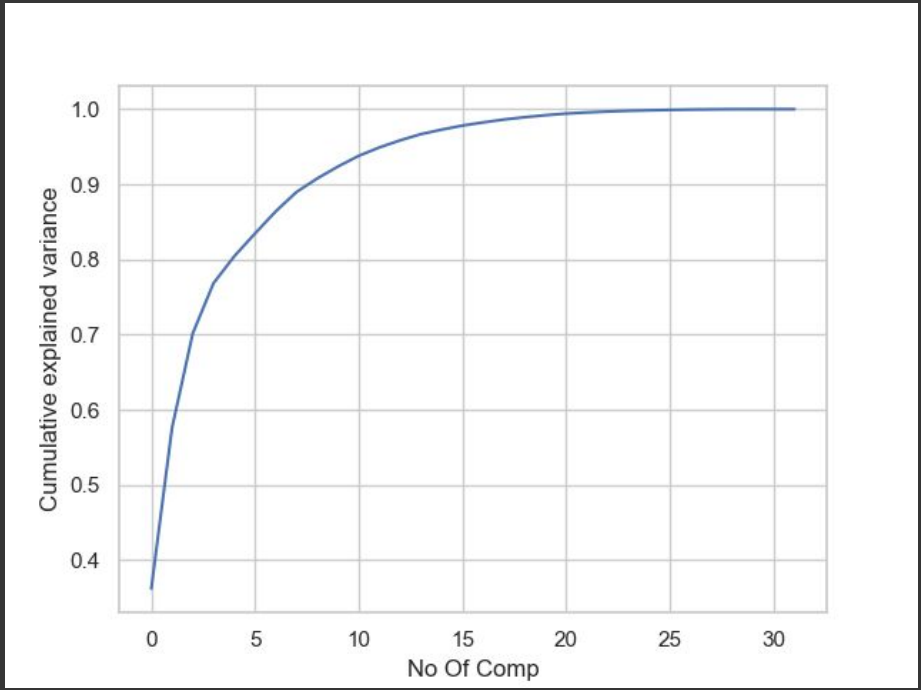


Insight

Correlation is a statistical measure that indicates the degree to which two variables are related to each other.

PCA on Numerical Columns only

No Of Comp Chosen: 20



Insight

We want to choose the number of components that will explain a sufficient proportion of the variance in the data while keeping the number of components as small as possible to avoid overfitting.

Objective 1: Machine Learning

1. Design a predictive model to determine the potential customers who will purchase if you send the advertisement . The target variable is Potential_Customer.

Apply various ML algorithms on the data, evaluate them after Grid Search and Cross Validation, and choose the best model.

Logistic Regression: 56.69%

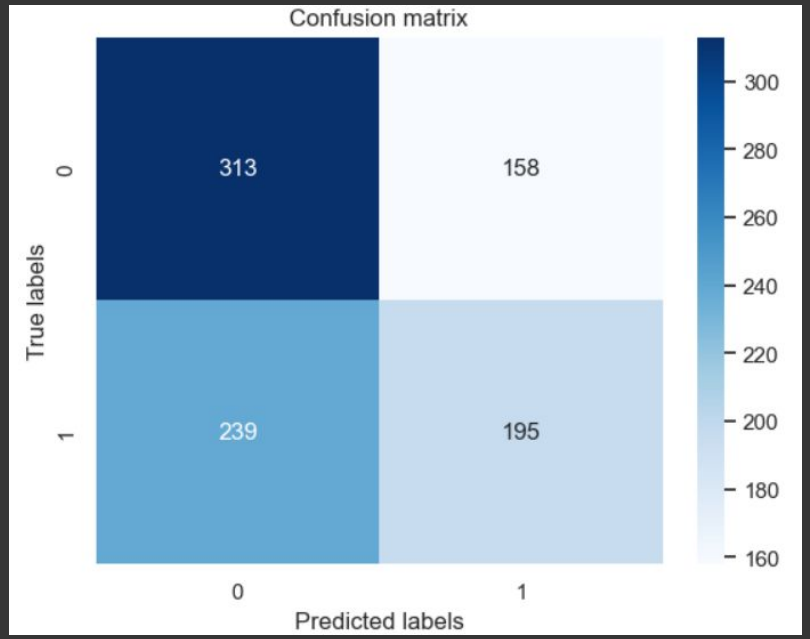
K-Nearest Neighbors: 50.00%

Decision Tree: 50.53%

Support Vector Machine (RBF Kernel): 51.30%

Random Forest: 54.06%

The best model is: Logistic Regression



Objective 2:

2. Calculate the value and the revenue of the model

1. Model Cost: The cost of the model is $1765.
2. Model Lost: The amount lost by the model is $3173.59.
3. Model Gain: The amount gained by the model is $2838.76.
4. Model Value: The value of the model is calculated by subtracting the cost and the amount lost from the amount gained, which results in a negative value of -$2099.83.
5. Model Revenue: The revenue generated by the model is $1073.76.

Metric	Value
Model Cost :	$1765
Model Lost :	$3173.59
Model Gain :	$2838.76
Model Value :	$-2099.83
Model Revenue :	$1073.76

The model has generated positive revenue, but high costs and losses outweighed the gains, resulting in a negative value. This data can help evaluate the model's performance and decide whether to continue investing or explore other options.

Objective 3:

3. Compare your best models' revenue with the revenue of the default solution which is sending advertisement to all the customers in X_test

1. The first row shows the cost of the default solution which is $4525.
2. The second row indicates the gain from the solution which is $6140.0.
3. The third row represents the revenue generated from the default solution which is $1615.0.

Metric	Value
Default Solution Cost:	$4525
Default Solution Gain:	$6140.0
Default Solution Revenue:	$1615.0

The default solution shows a positive revenue of **$1615.0**

Model Revenue: The revenue generated by the model is **$1073.76.**

Solution: Send ads to all customers

Objective 4:

4- calculate your model's revenue on this set of 30,000 customers

1. The revenue generated by the **"Model"** solution is **$200,056,** while the revenue generated by the **"Default"** solution is **$42,173.2.**

	Model	Default
Revenue	200056	42173.2

The table compares the total revenue generated by the model and the default solution, showing that the model solution generates significantly higher revenue of $200,056 compared to the default solution's $42,173.

This indicates that utilizing the model for decision-making could result in a more **positive outcome.**

Recommendations

For building an effective ML system, it is crucial to choose the appropriate model. To achieve this, **explore various models and compare their performance** to determine the best fit for the problem at hand.

Improving an ML model significantly relies on the quantity and quality of the training data. Enhancing the data can be accomplished by **obtaining more data or enhancing the quality of the existing dataset**.

Change the target variable from Potential Customer to other target such as Age, Annual Income, Gender and House Ownership


THANK YOU