Target Market Analysis

ADCP Capstone Project

Objectives

- The **default option** is to **send advertisements to everybody**.
  - Given the cost of **sending one advertisement (RM5.00)**
  - The average purchase of **RM14.56**

- **Smaller** sample size revenue (~3,000): **RM1,693.25**.
- **Larger** sample size revenue (30,000): **-RM106,182.93**

Findings

- **Model statistics**:
  - Precision: 0.50
  - Recall: 0.82
- **Deep Learning:**
  - Precision: 0.53
  - Recall: 0.87

- **Smaller** sample size revenue (~3,000): **RM1,660.019**
- **Larger** sample size revenue (30,000): **RM21,879.99**

THE PROCESS

HOW WE GOT HERE...

INTRODUCTION

Identify the problem.

EDA

Getting to know our data.

FEATURE SELECTION

Preparing our data.

MODEL BUILDING

Process our data.

ANALYSIS

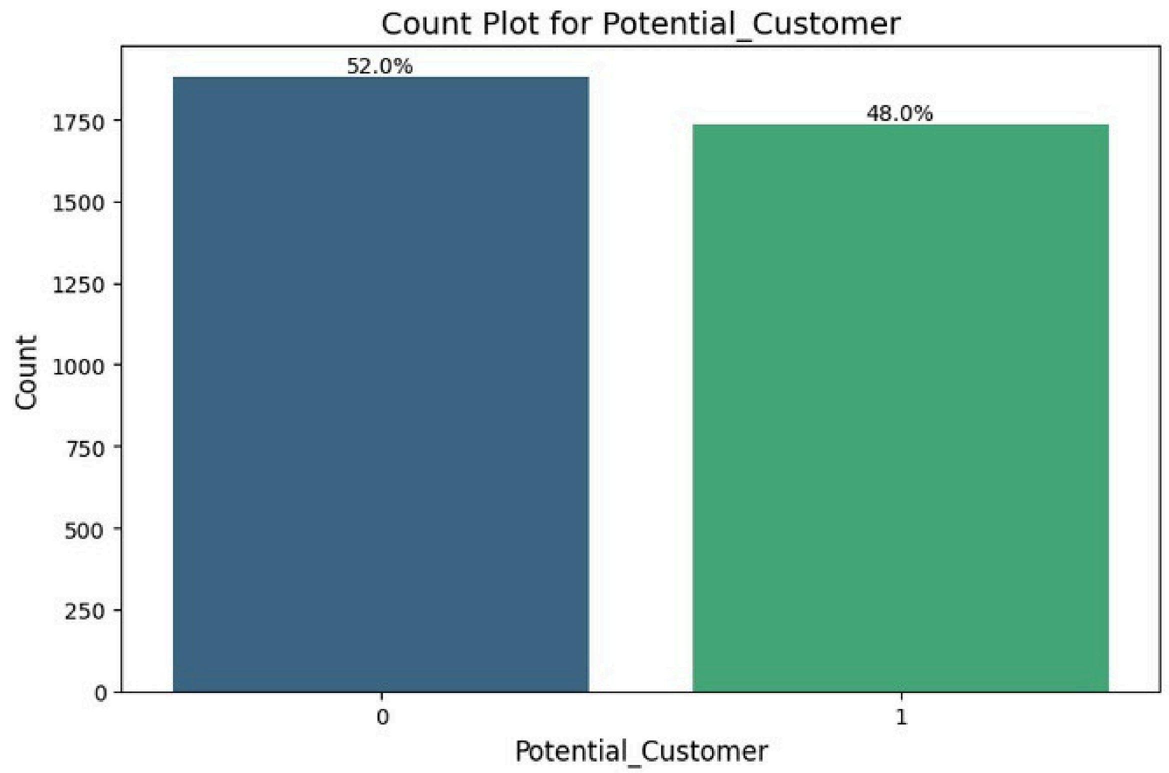
Providing actionable insights.

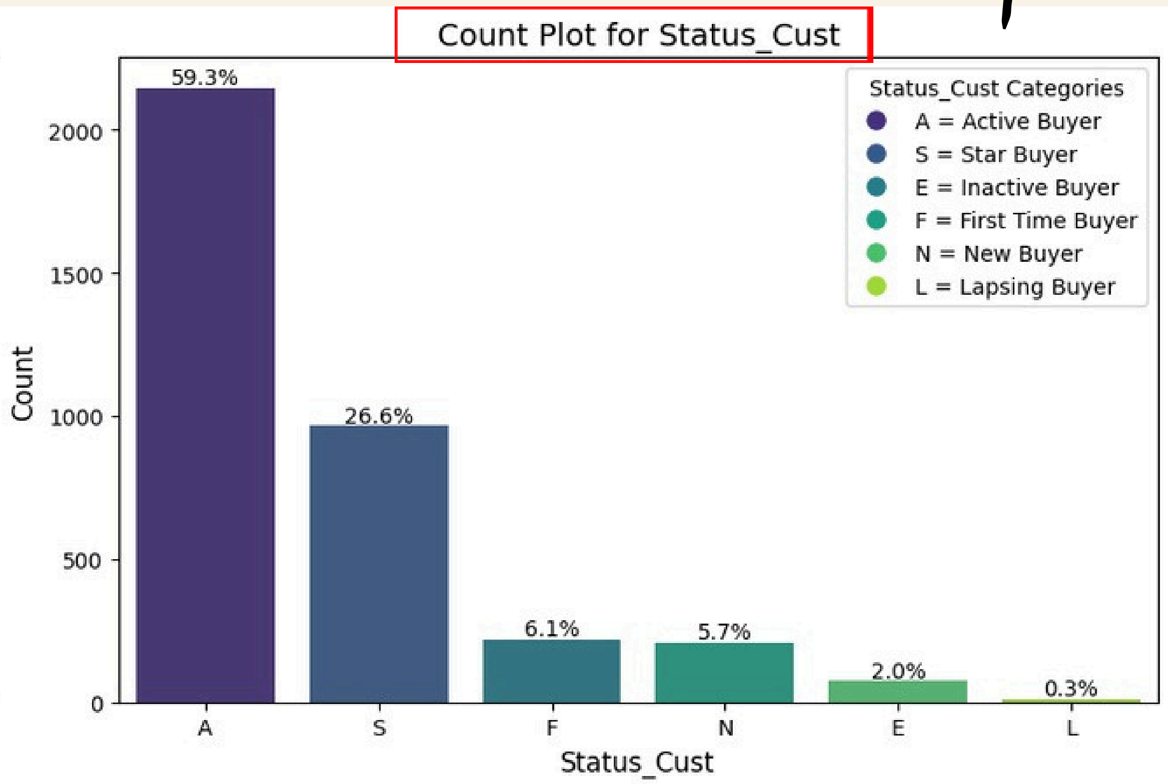
DESCRIPTIVE STATISTICS (#1)

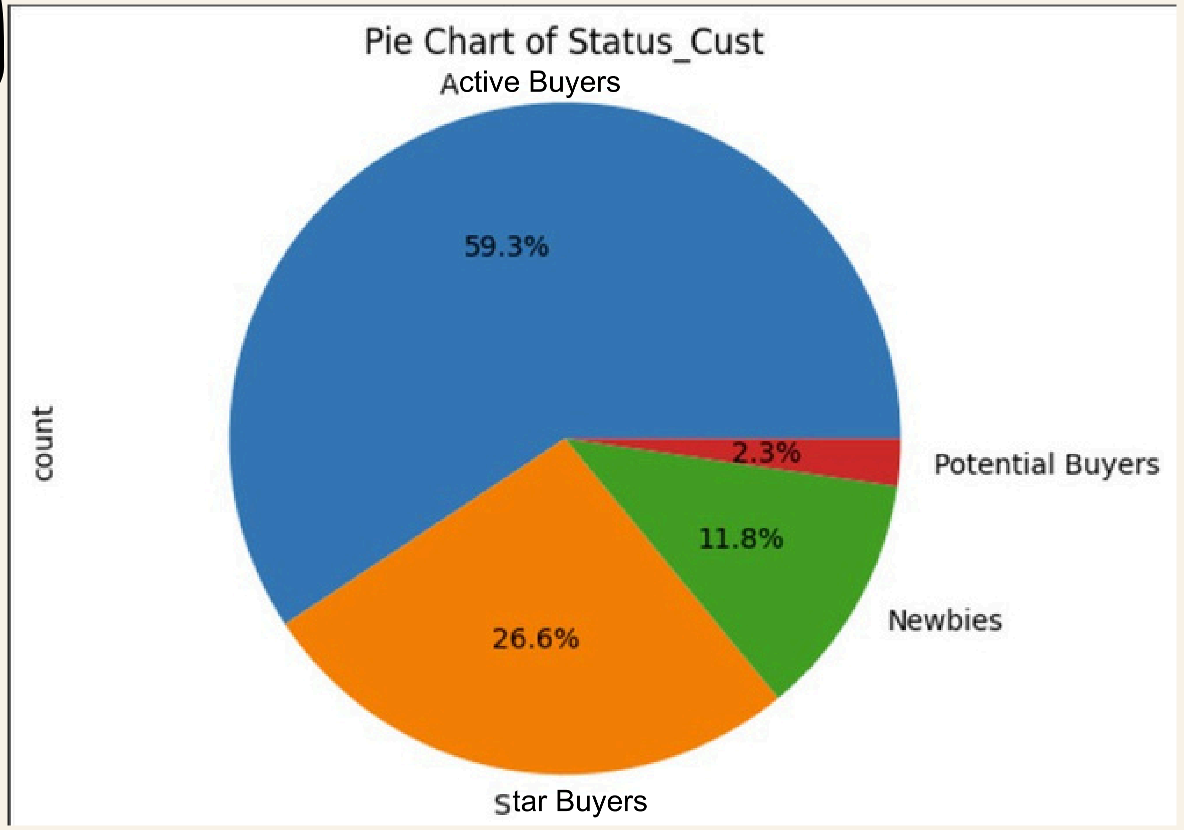
CATEGORICAL FEATURES

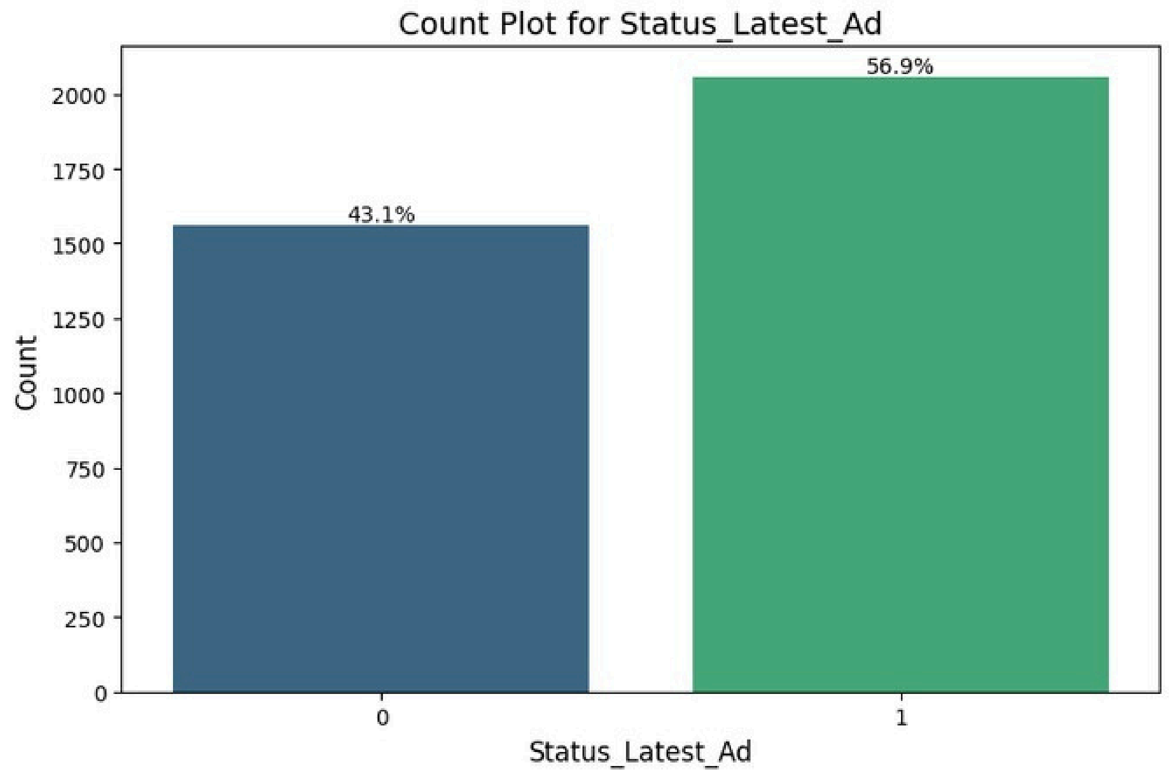
Data Size
- 4469 Rows
- 25 Features

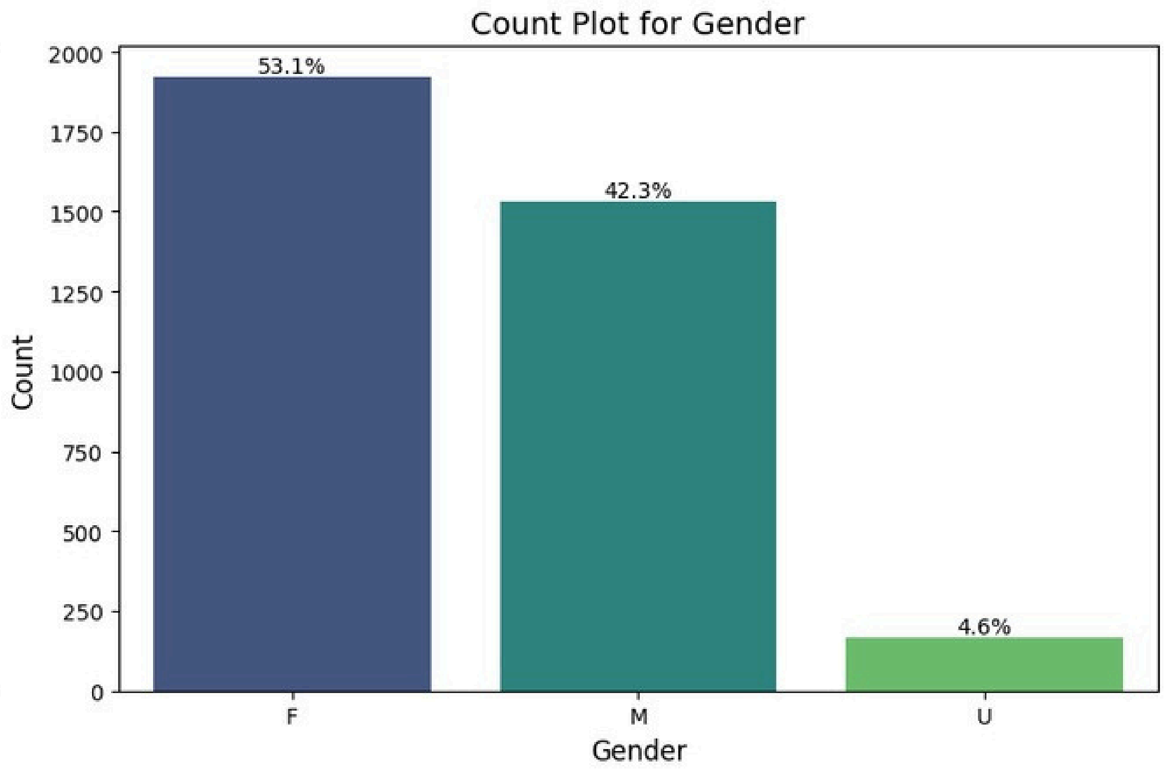
E & L → Potential Buyers
F & N → Newbies

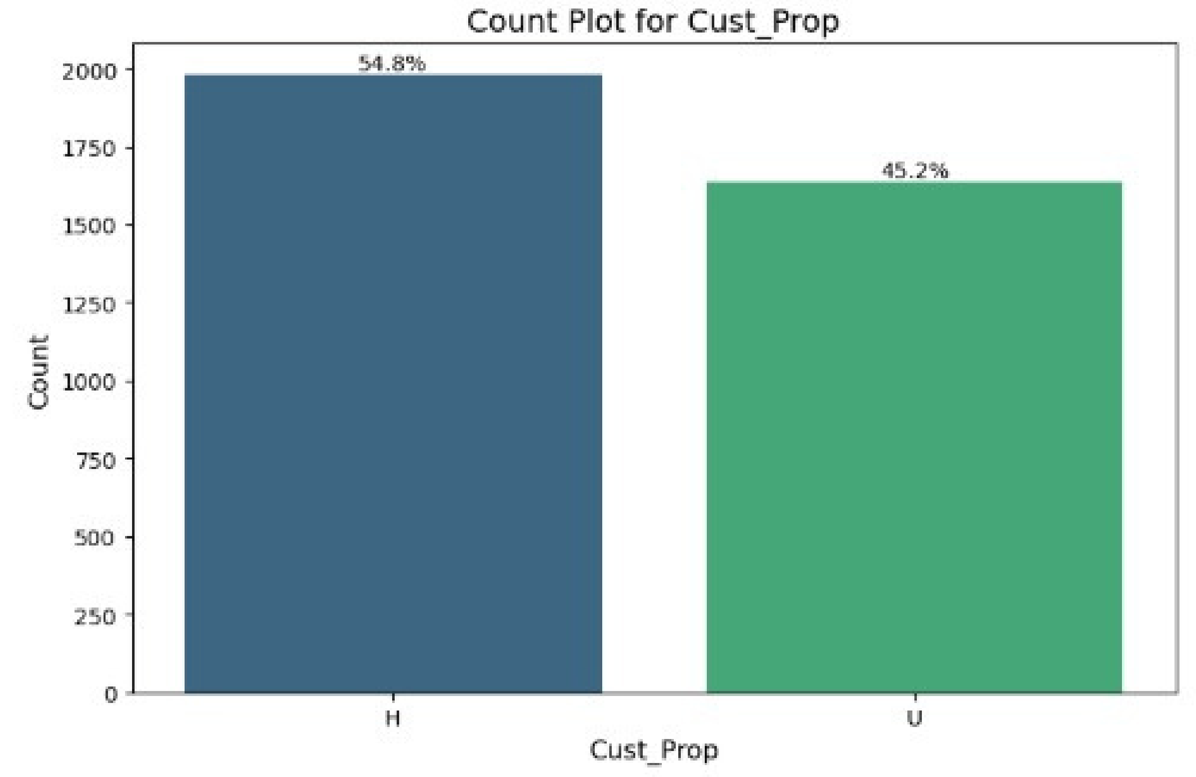










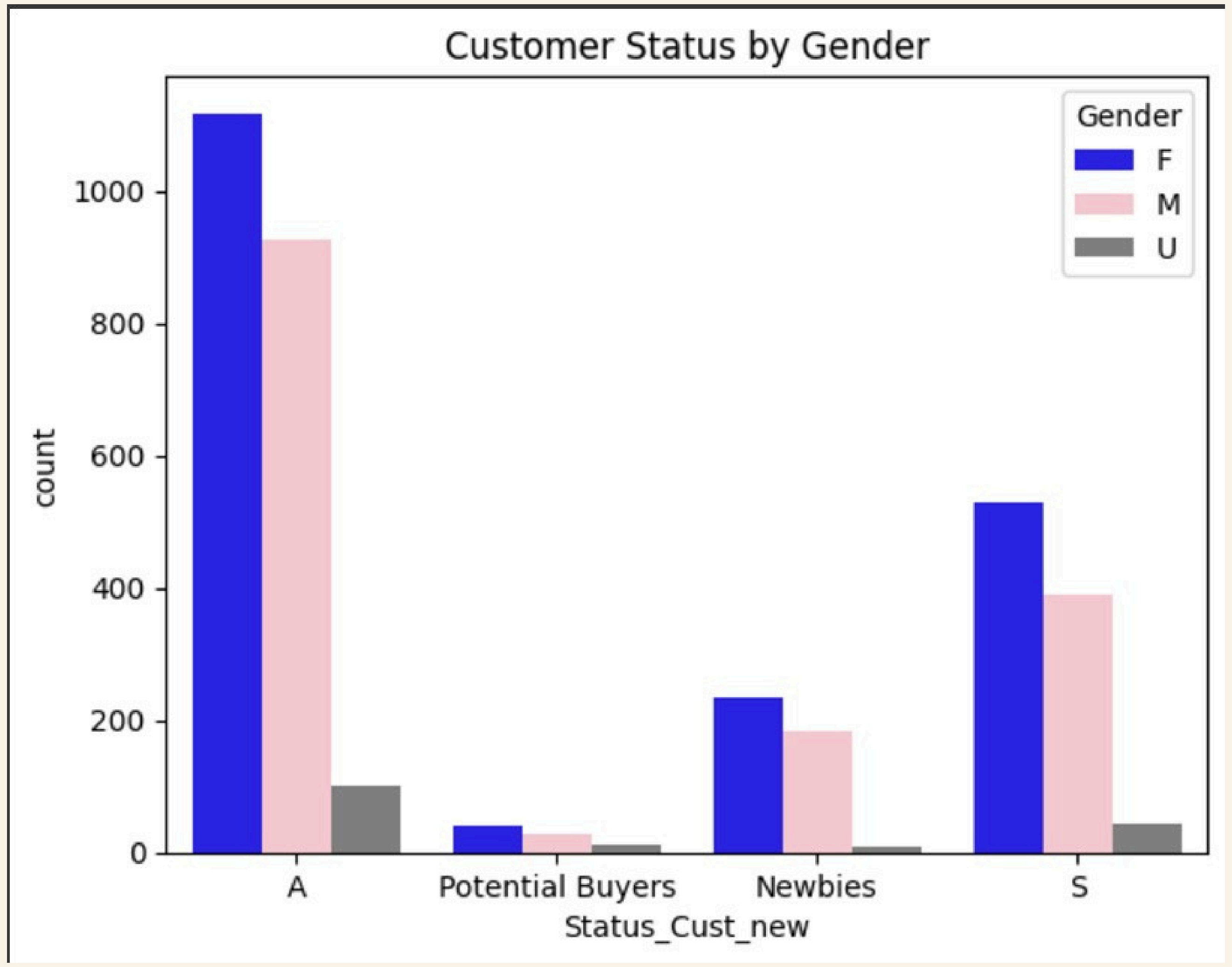


GETTING TO KNOW OUR DATA...

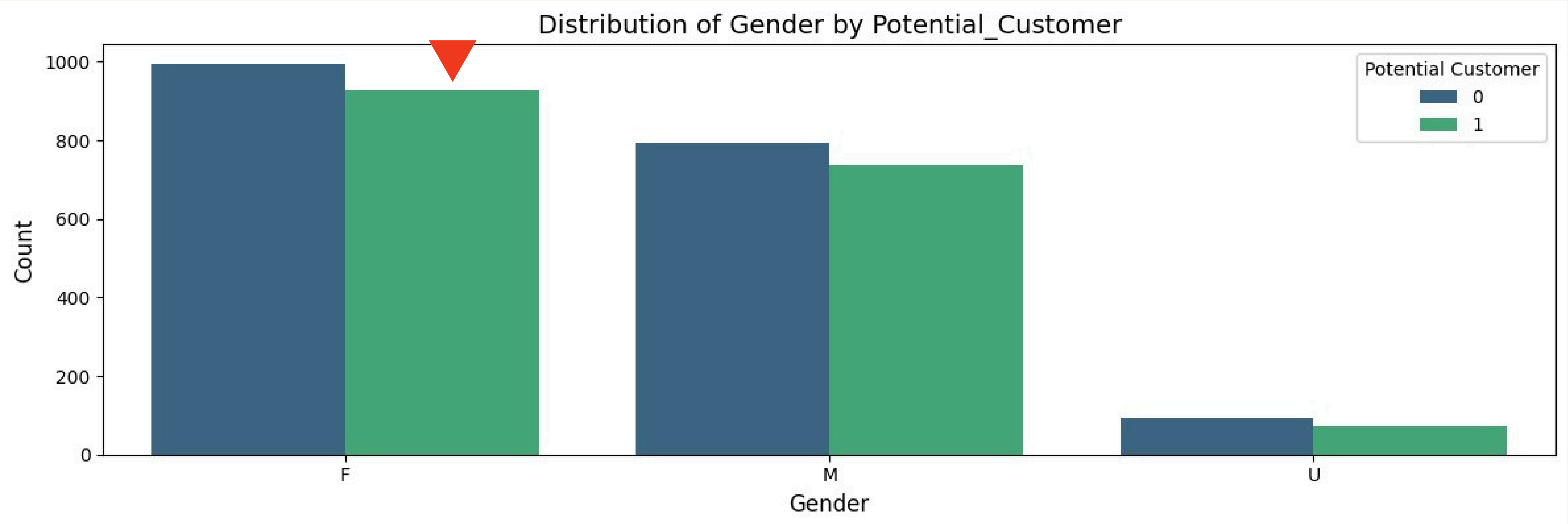
- Is there any significant difference between men/women's salary?
- Is there any significant difference between men/women's number of the purchase in the last three years?
- Is there any significant difference between men/women's average purchase in the last three years?
- Is there any significant difference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

- **ANOVA TEST**
- **TUKEY'S TEST**

Females dominate all the customer segments, #GIRL POWER !!

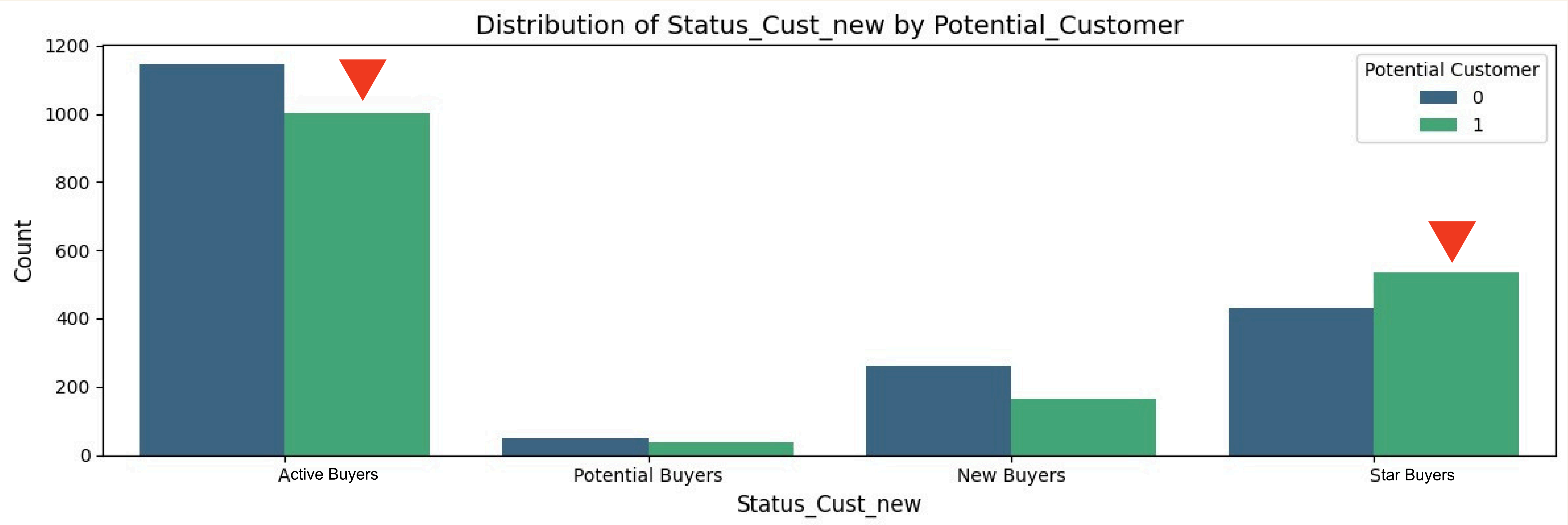


INSIGHTS



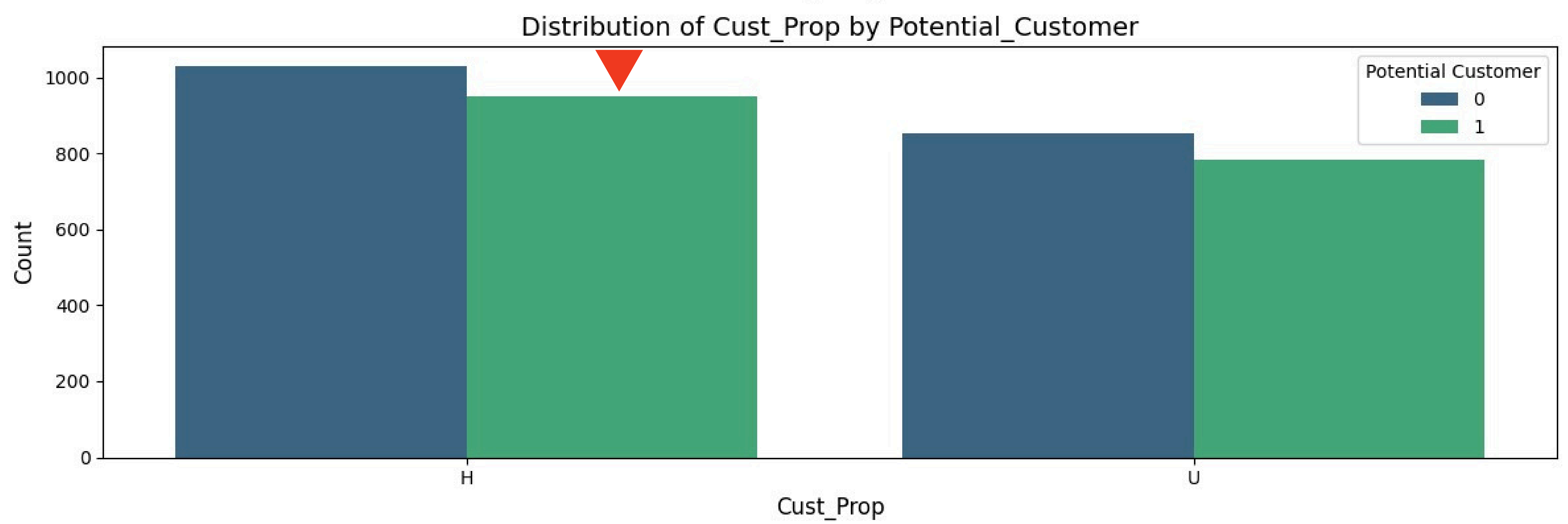
By classifying potential customers based on gender, we see **more conversions** made on **female customers.**

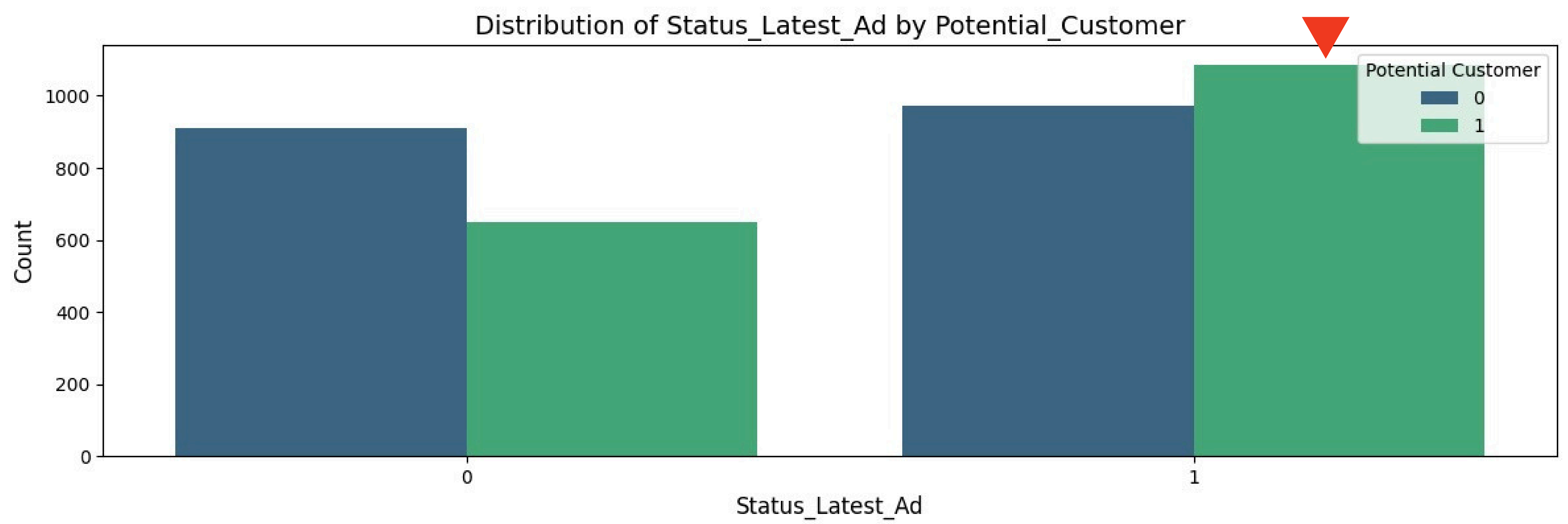
Active and Star Buyers are more convertible compared to other categories.



INSIGHTS

Houseowners are more **prone** to become our **potential customers.**





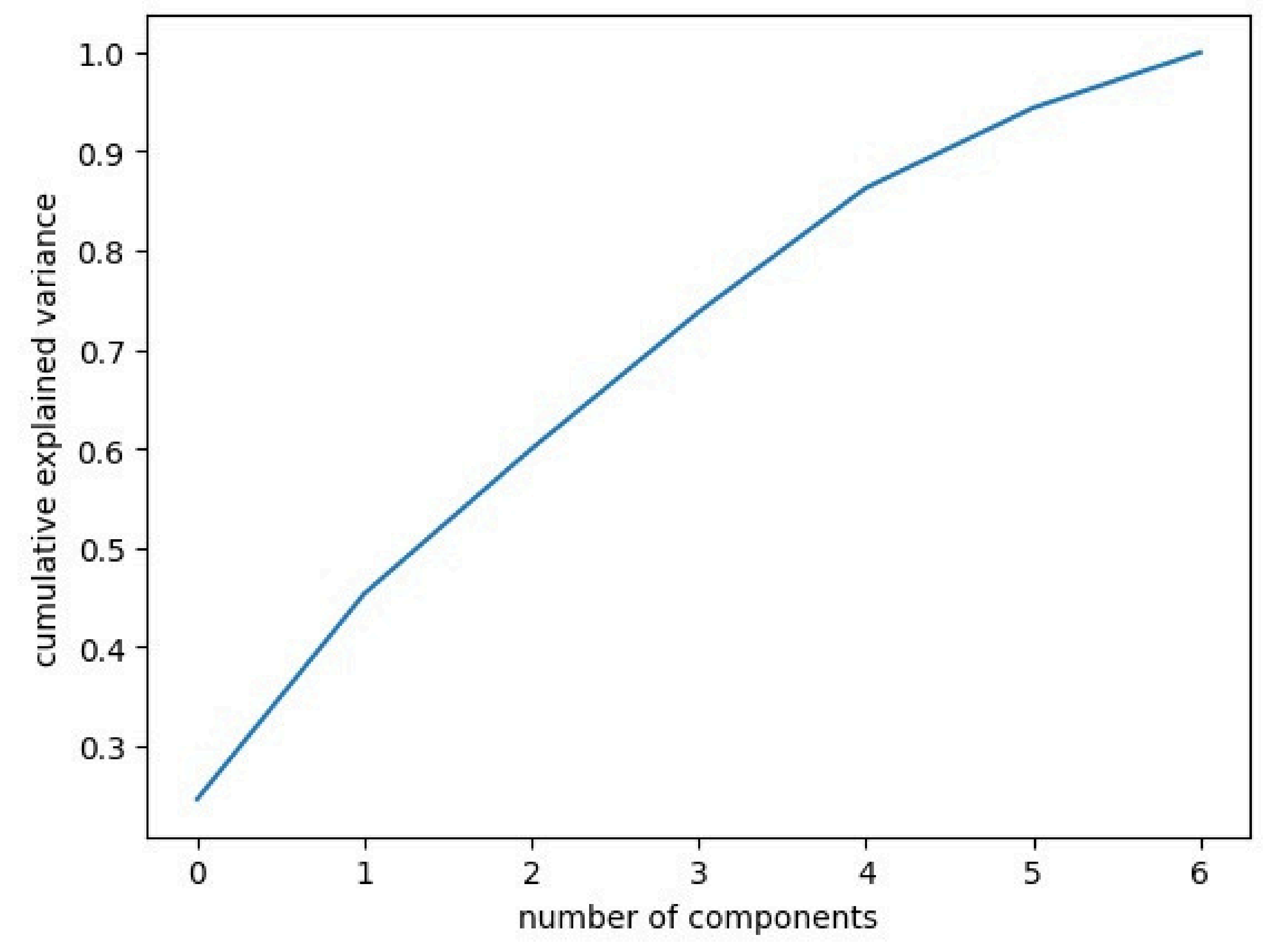
Solid proof that **our ads are efficient in promoting conversion**. Thus, we only need to **find the right customer group.**

FEATURE SELECTION

Sl.No	Variable	Description
1	Potential_Customer	ResponseVariable
2	C_ID	Cutomer Identification Number
3	Cust_Last_Purchase	Amount purchased in $ (most latest / recent purchase)
4	Pur_3_years	No of purchases made in the recent 3years
5	Pur_5_years	No of purchased made in the last five years
6	Pur_3_years_Indirect	No of purchases made in 3 years through link from other websites (indirect buys)
7	Pur_5_years_Indirect	No of purchases made in the last five years through link from other websites
8	Pur_latest	The latest purchase amount (in thousands)
9	Pur_3_years_Avg	Average Purchase over the last 24 months
10	Pur_5_years_Avg	Average Purchase over the last 5 years
11	Pur_3_years_Avg_Indirect	Average Indirect Purchase through link from other sources for the last 24 months
12	InAct_Last	Inactive no of months since the customers made the last purchase
13	InAct_First	Inactive no of months since the customers made the first urchase
14	Ad_Res_1_year	No of Promotional Ads by MyPurchase responded by the customer online in the last one year
15	Ad_Res_3_Year	No of Promotional Ads responded by the customer online in the last 3 years
16	Ad_Res_5_Year	No of Promotional Ads responded by the customer online in the last 5 years
17	Ad_Res_Ind_1_Year	No of Ads responded to the other sources (indirect) which directed to MyPurchase in the last 1 year
18	Ad_Res_Ind_3_Year	No of Calls made by References to the individual for the last 36 months
19	Ad_Res_Ind_5_Year	No of Calls made by References to the individual over the period of few years
20	Status_Cust	A if active buyer, S if star buyer, N if new buyer, E if inactive buyer, F if first time buyer, L if lapsing buyer
21	Status_Latest_Ad	1 if individual has purchased in response to the last promotional sale , 0 if not
22	Age	Age of the individual
23	Cust_Prop	Owns a House H- Owner / U-Unknown
24	Gender	Sex of the individual
25	Cust_Ann_Income	Customer_Annual_Income

No	Variable	Description
1	Potential_Customer	Response variable
2	Total_Purchase_Amount_3_years	Total Purchases (direct and indirect) made in 3 years
3	Pur_latest	The latest purchase amount (in thousands)
4	InAct_Last	Inactive for number of months since the last purchase
5	InAct_First	Inactive for number of months since the first purchase
6	Total_Ad_Res_3_Year	Total Ad Responses (direct and indirect) in last 3 years
7	Status_Cust_New	E,L (Potential Buyers); N,F (New Buyers); A (Active); S (Star)
8	Status_Latest_Ad	1 if purchased due to the last promotional sale; 0 otherwise
9	Age	Age of the individual
10	Cust_Prop	House ownership status: H (Owner), U (Unknown)
11	Gender	Gender of the individual
12	Cust_Ann_Income	Customer's annual income

PCA ON NUMERICAL COLUMNS ONLY



Cumulative Explained Variance to number of components

First 6 principal components

PC1	PC2	PC3	PC4	PC5	PC6
-0.800303	-0.078809	-1.148542	-1.326875	0.263001	-0.001646
0.856304	-0.505942	-0.580340	0.348188	0.291722	0.092944
2.026358	0.631294	0.371297	0.070716	0.336966	-0.945797
1.013115	-0.069233	-0.353553	-0.415438	1.339440	-1.230904
0.296507	0.005161	1.226534	0.390485	0.562335	-0.901171

MODEL BUILDING AND SELECTION

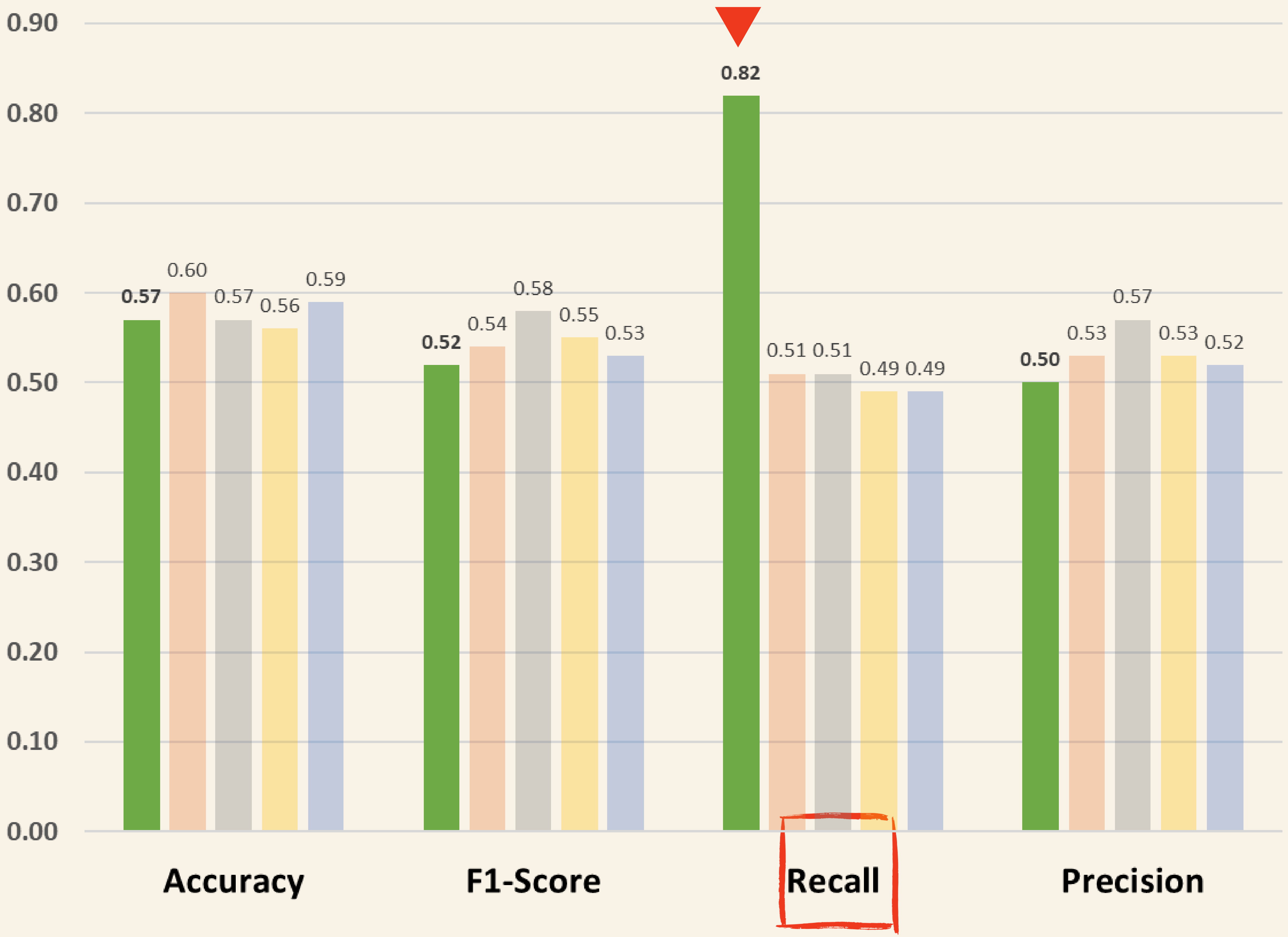
Using GridSearchCV :

- Decision Tree Classifier
- Support Vector Classifier
- Random Forest Classifier
- KNeighbors Classifier
- Logistic Regression

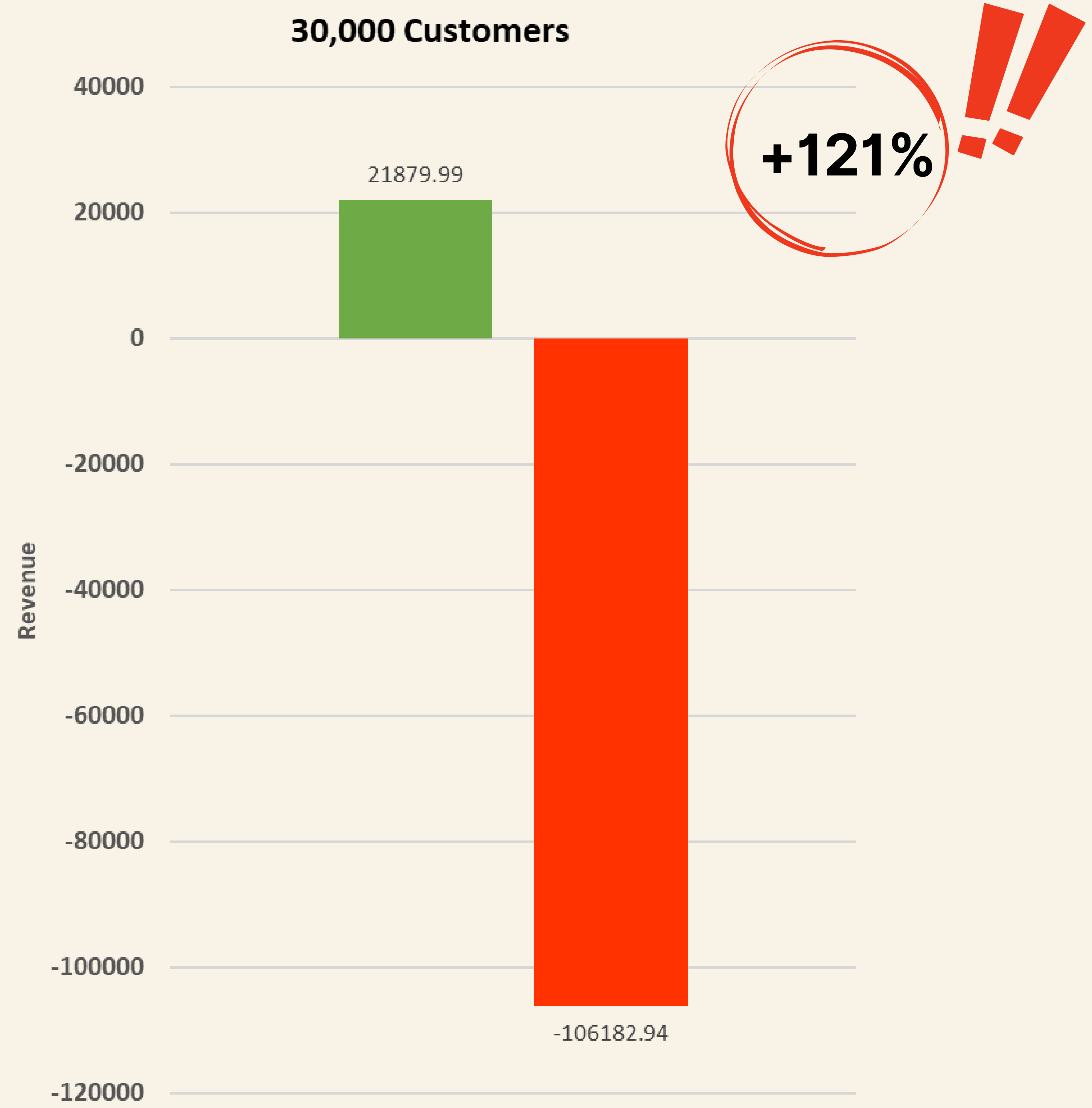
Hyperparameter

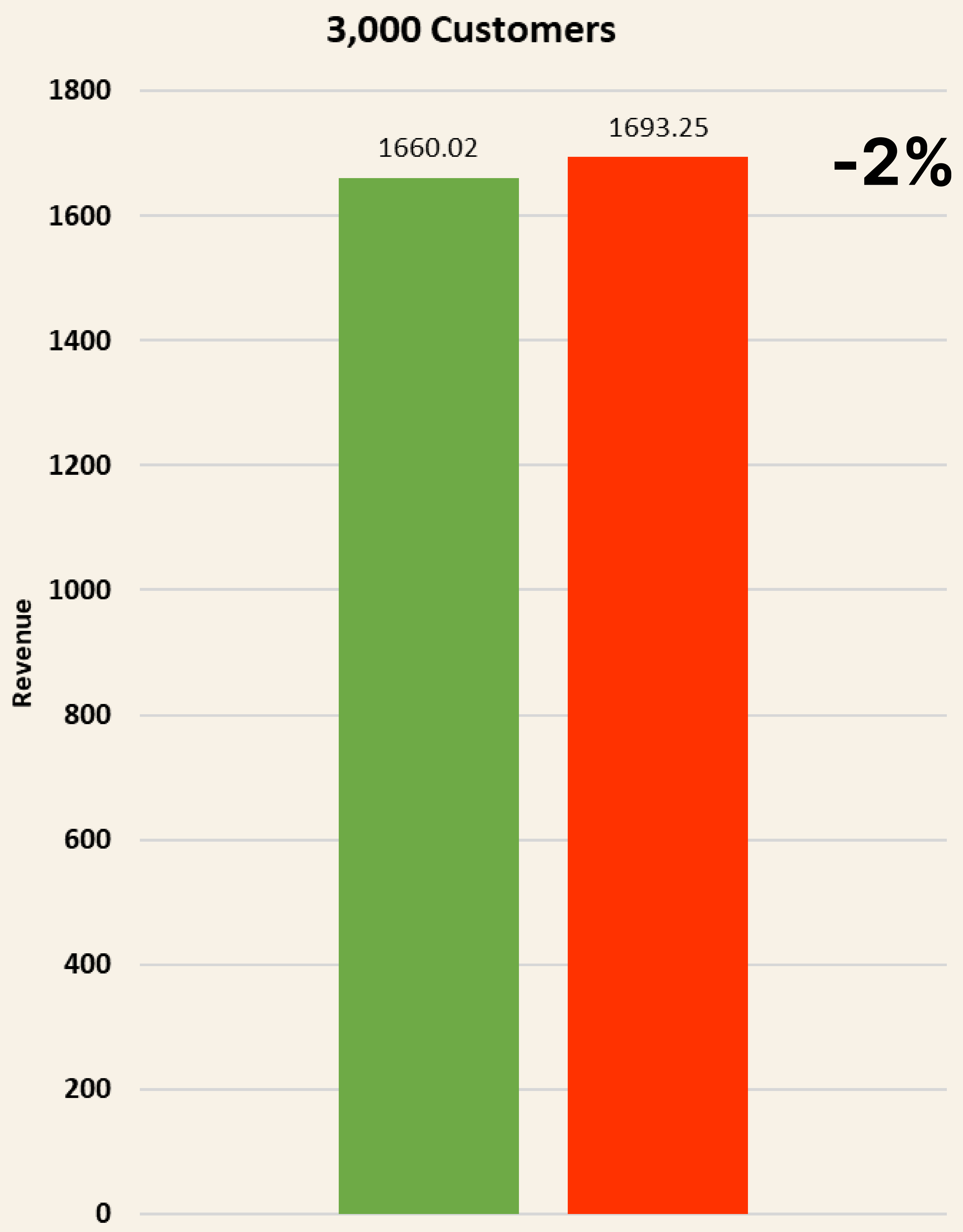
- max_depth = 2,
- max_features = 'log2',
- min_samples_leaf = 2,
- min_samples_split = 20

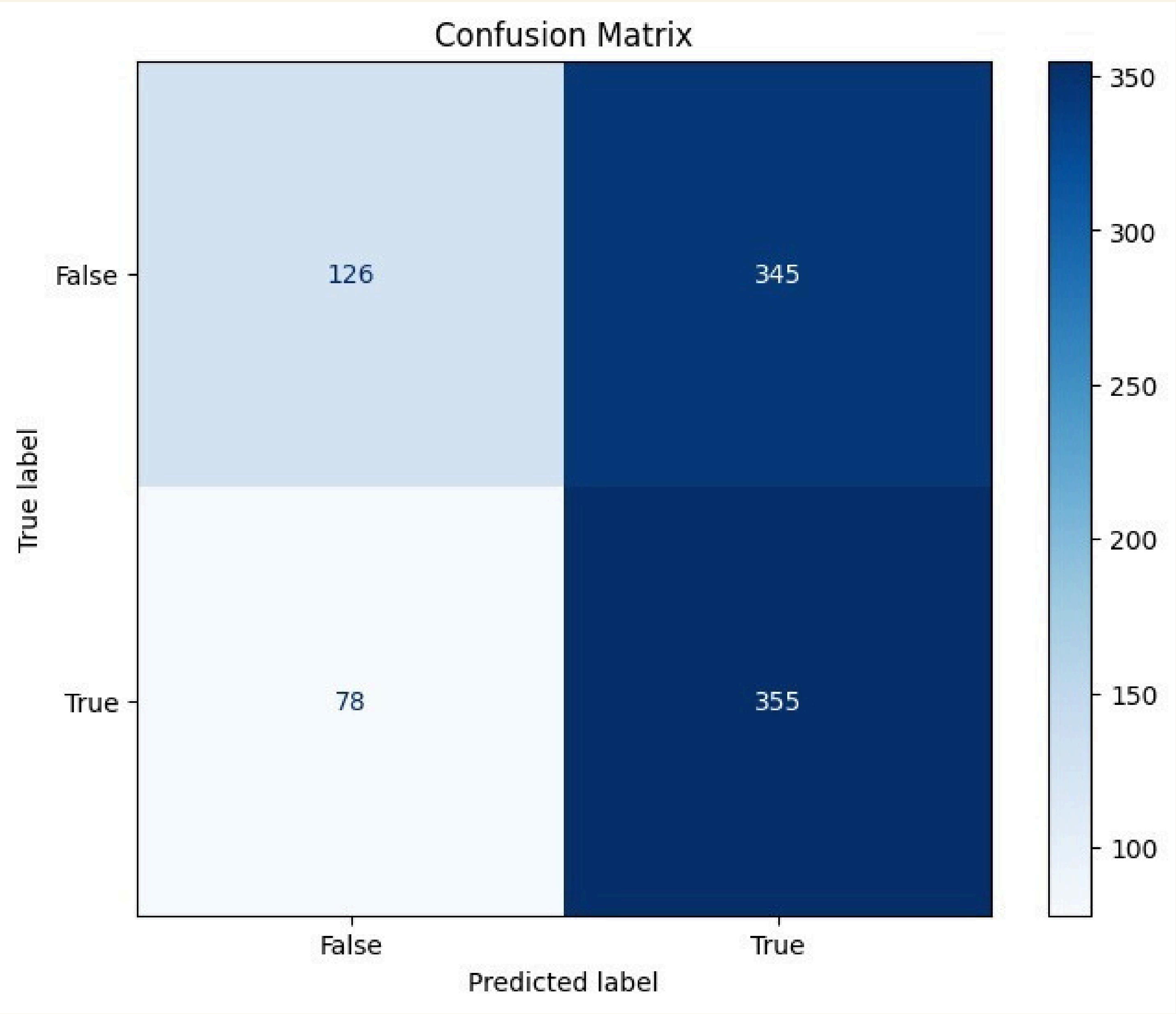
Here, Recall is preferred as we want to get **as many positives as possible**. We are not worried about false positives; instead, we want to **reduce the number of false negatives**. Thus, **recall is preferred scoring metric**.



The model shows **better performance** than **the default option** in a larger number of customers and a *marginal difference* in a smaller number of customers







Classification Report:

	precision	recall	f1-score	support
0	0.66	0.22	0.33	471
1	0.51	0.87	0.64	433
accuracy			0.53	904
macro avg	0.58	0.55	0.49	904
weighted avg	0.59	0.53	0.48	904

OUR FURTHER ENDEAVOUR

Recall = 0.87

Using Deep Learning frameworks like TensorFlow. We are able to get a **better recall score** for the model. Thus, with better tuning of models, we believe that we can get better results with better quality data and time.

Deep Learning architecture:

```
model = Sequential([
    Dense(11, activation='relu', input_shape=(X_train.shape[1],)),
    Dropout(0.4),  # Increased dropout for regularization
    Dense(32, activation='relu'),
    Dropout(0.4),
    Dense(16, activation='relu'),
    Dense(1, activation='sigmoid')  # Sigmoid activation for binary c
])
```

Conclusion

Customer Profile :

Female, Active and Star Buyers, and Latest Ad Responders are more likely to be potential customers

Model Performance :

Decision Tree is a good enough model as compared to other models to predict potential buyers without incurring loss.

The Future Way:

Involve industry experts for better data understanding & feature scaling

Use deep learning for better modeling