BA 305

Forecasting the Target Market: Predicting New Customers Purchase Decisions

Team B5: Joanne Charles, Cristina Jiang, Kaya Manolt, Jaden Cho, Jaden Noh

1. Introduction

1.1 Selecting A Dataset

For our project, we decided to create a model that would accurately predict if a customer would make a purchase in the energy sector based on the data we have from a previous marketing campaign. Our major question is **how can we increase revenue?** We wanted to identify the factors that have the largest influence on the purchase decision. Using this model, future campaigns can be more efficient and effective in targeting consumers.

At the beginning of this project, we considered different datasets and project ideas around predicting a target variable. We narrowed our selection to two options: a dataset about factors that contribute to whether or not someone rents an apartment and our chosen dataset on energy. We decided on the dataset on energy because of both our shared interest in climate change and sustainability and the fact that this dataset was more complete with less missing data and a better balance between numerical and categorical variables. We also were successful in contacting the owner of the dataset and learned more about the industry behind this series of telemarketing campaigns.

1.2 The Energy Sector Marketing Analytics Data

After choosing the dataset, we wanted to solidify what the data meant, what we were predicting and how we could use the data to make the model. Our first idea was to create a model that determines a customer's likelihood to purchase. However, we shifted our focus to creating a model that would accurately predict purchases after learning more about factors through exploratory data analysis.

The dataset we got from Kaggle was provided by Dr. Maryam Khanian and after doing some research and speaking with Dr. Maryam we learned that this data was from and centered around the energy industry. This dataset has 31,480 observations and 20 features with 11 clear categorical variables such as gender, job, marital status, education, and credit failure. There were also 9 numerical variables, however we decided to make some categorical variables into dummy variables to increase our number of numerical variables. Again, the target variable is the most important for our analysis because it records the amount of customers who end up purchasing the products or not. We were confident in using the data to make predictions and proceeded to clean and preprocess the data.

2. Data Preprocessing

2.1 Cleaning and Reformatting the Data

After assessing the data quality of all potential datasets to aid in our decision over which set to analyze, we began data preprocessing. Compared to all the other datasets that we looked at, our dataset

did not require as much cleaning. All of the dimensions were relevant and had minimal correlation so we did not need to combine or drop any of our columns. One of our dimensions, “daySinceLastCampaign”, which classified the amount of time between the current campaign and the last one completed, had a lot of missing data. We decided to fill up these missing values with -1 to signify that these target customers were not part of the target audience in previous campaigns. While there were a lot of nulls, we assumed that this missing data might have some significance and should be kept when examining correlation matrix and EDA. Our final step in cleaning our dataset was to relabel the “target” column which measures whether or not a customer within the campaign made any purchases. We changed the column name to “purchase” which we felt was a more accurate representation of what the dimension represented and easier to understand when working with the data.

2.2 Reviewing our Correlation Matrices

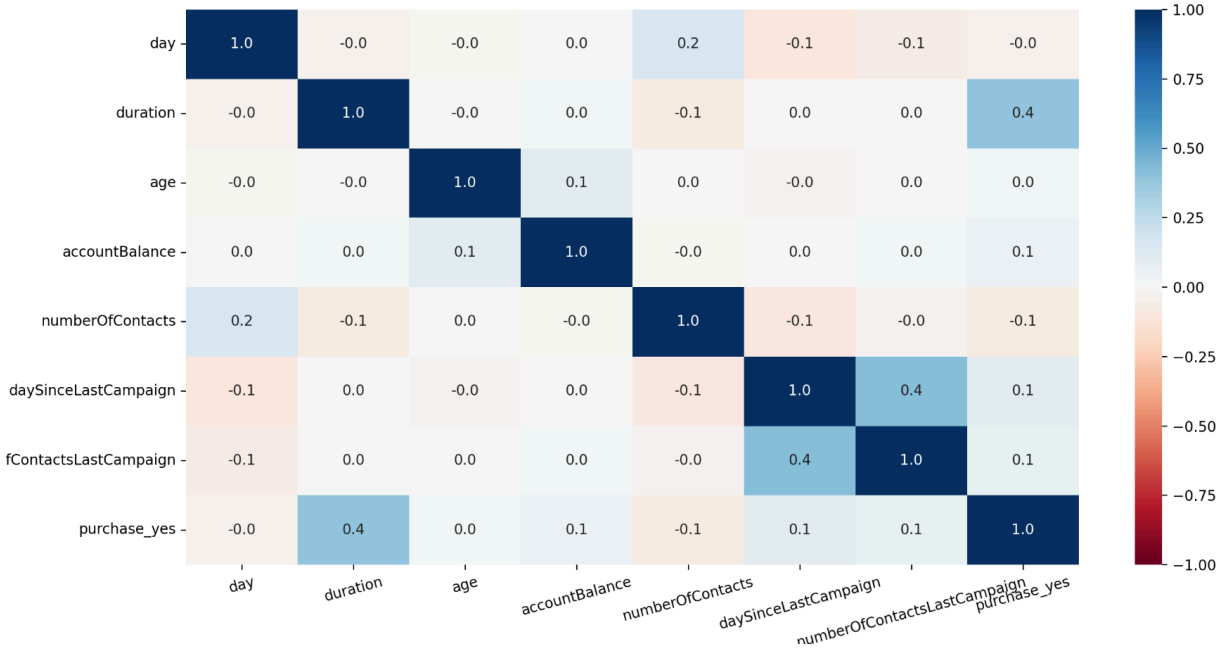


Figure 1: Correlation matrix of few features

We created this correlation matrix to check the correlations amongst the numeric variables. After analysis, we found that there is no strong correlation (Figure 1) among any variables. The highest correlation we found was between duration (how long a person stayed on the phone with a representative) and people who made the purchase (purchase_yes), which is 0.4; our team thinks this is acceptable. In the end, because there are no strong correlations, we didn’t drop any variables and we don’t think it’s necessary to perform PCA either.

2.3 Balancing the Datasets

The original target variable (purchase) was not balanced, with 88.25% who did not make the purchase and 11.75% who made the purchase (shown in Figure 2). Because of the imbalance nature, we decided to make the dataset more balanced to ensure the best results of our prediction models. We

decided to use a random over sampler. By applying this sampling method, it balances the class distribution, so that each class has a similar proportion (shown in Figure 3). This makes the dataset more balanced, which can lead to better model performance for later.

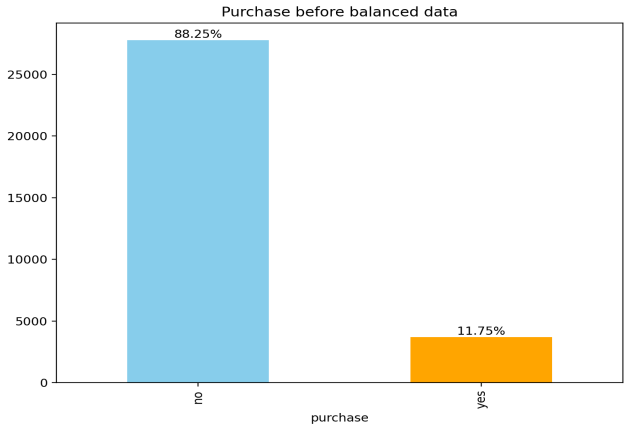


Figure 2: Purchase data with imbalance

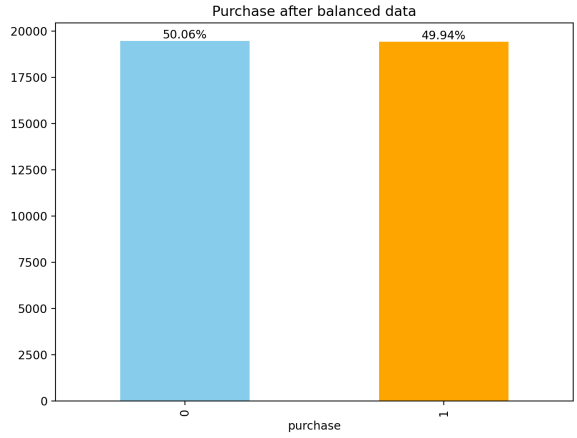


Figure 3: Purchase with balanced data

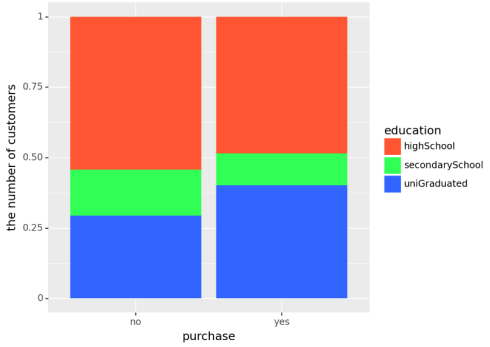
2.4 Reformatting for the Prediction Model

Once we began creating our machine-learning algorithm, we needed to convert our categorical data into a numerical format utilizing One Hot Encoder to transfigure these categories into binary values. Because aforementioned “daySinceLastCampaign” column turned out to be insignificant after EDA and correlation check, we decided to drop the ‘daySinceLastCampaign’ and ‘lastCampaignResult’ to increase the performance of our model since they had a lot of null values. We also dropped ‘contactid’ and ‘id’ because they aren’t relevant to our model.

3. EDA

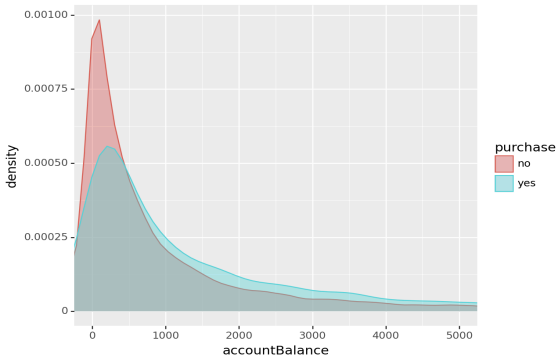
For our exploratory data analysis, we focused on account balance and education as factors that we were interested in learning more about and seeing if there were any patterns. While correlations were not very significant, we wanted to see if there was any information that we could gain from exploring the data.

Looking at the correlation between education and purchase, we can see that within our last campaigns, persons whose highest form of education was high school were more likely to not purchase the product while those we targeted with a university education made purchases in a higher quantity than those who did not. Based on these observations,



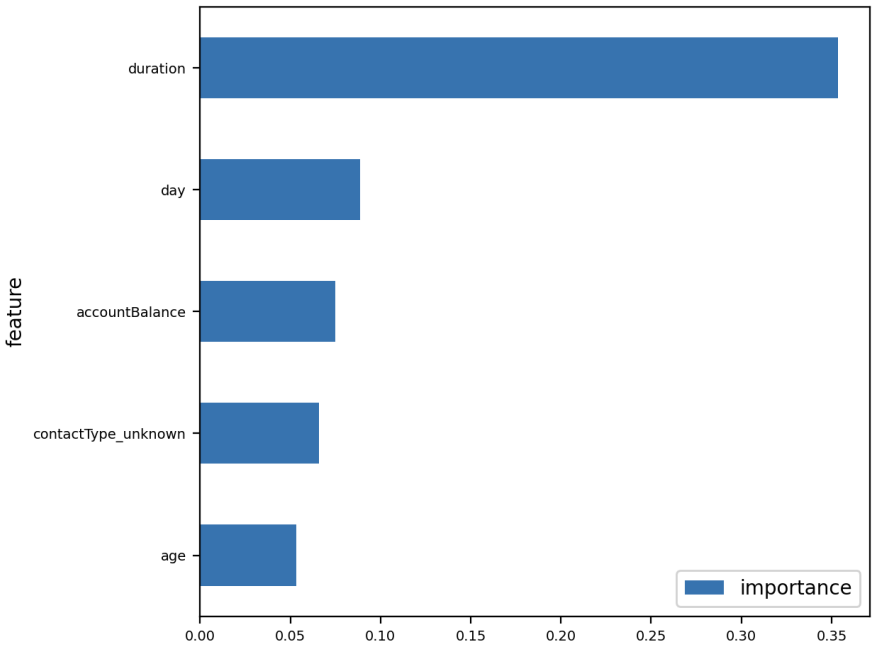
we would advise a focus on the next marketing campaign around highly educated persons.

Even though there are discrepancies of whether or not someone will make a purchase, once we reach the point where people have $500 in their account balance it seems to show that a higher concentration of individuals in the 'no' group have lower account balances compared to the 'yes' group. Even though there is a small percentage of those that said yes, we believe that the campaign should be focused on those who have higher than 500 in their account balance. It would be more efficient to focus our efforts on this group of people.



4. Decision tree classifier feature selection/ Logistic Regression

In our approach, we utilized a decision tree classifier to assess the significance of features within our dataset. Recognizing the imbalance in our target variable, "purchase," we addressed this issue by employing the RandomOverSampler technique to balance the dataset, ensuring a 50:50 ratio between classes.

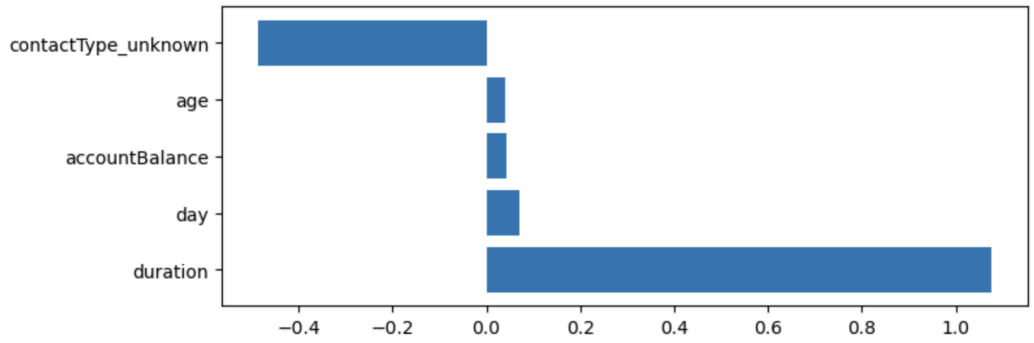


Post-balancing, we conducted feature importance analysis using the decision tree model. "Duration" emerged as the most crucial feature, followed by "day" (contact day-day of a month- in the previous campaign) and "account balance," which exhibited similar levels of importance. Noteworthy features also include "age" and “contactType_unknown” among others.

While "Contact Types" is a single variable, its categorical nature required dummy coding, resulting in the separate identification of contactType_unknown.

Moving forward to logistic regression, we selected these features due to their prominence in the decision tree analysis. In our logistic regression analysis, we identified several key features expected to significantly influence the likelihood of a purchase. Among these features, the coefficients for 'duration' and 'contactType_unknown' stood out as particularly impactful.

For 'duration', the coefficient of 1.076307 implies that for every one-unit increase in duration, the log-odds of making a purchase increase by 1.076307. This translates to the odds increasing by approximately 2.93 times. We can expect the longer duration of the marketing campaign on customers to increase the probability of making a purchase. With the p_value of 0.000, it is statistically significant, and it is the main factor we should take into account for successful marketing and purchase boosts.



Similarly, the coefficient of -0.487028 for 'contactType_unknown' indicates that having the contact type labeled as 'unknown' decreases the log-odds of making a purchase. The exponential of this coefficient, approximately 0.61446, suggests that the odds of making a purchase decrease by a factor of about 0.61446 when the contact type is 'unknown'.

We should consider the two different contact types that target consumers might have for the marketing campaign and consider the purchase intent because an unknown contact type has less than 1 coefficient and is interpreted as the negative factor on the marketing campaign and the customers. With its p-value of 0.000, it is statistically significant, so a valid factor that affects the customers' decision making that we can trust. This may be a unique behavioral difference. It is assumed that for the consumers who have not specified whether they are using a cellphone or a landline, they have either not been contacted previously or may be more suspicious people. People who are more cautious and reluctant to give their information to strangers are also less likely to purchase our product after a phone banking campaign.

	features	coef	std err	z	P>|z|	[0.025	0.975]	exp_coef
0	day	0.0718	0.029	2.493	0.013	0.015	0.128	1.074440
1	duration	1.0763	0.023	47.611	0.000	1.032	1.121	2.933804
2	age	0.0388	0.033	1.158	0.247	-0.027	0.104	1.039563
3	accountBalance	0.0431	0.022	1.994	0.046	0.001	0.085	1.044042
4	contactType_unknown	-0.4870	0.043	-11.362	0.000	-0.571	-0.403	0.614467

While other features like 'day', 'age', and 'accountBalance' also play a role, their coefficients suggest a less pronounced impact on purchase likelihood. For instance, a one-unit increase in 'day' results in a modest increase in the odds of purchase by a factor of approximately 1.074440, and so forth.

5. Random Forest (predictive model)

In this project, we have tested 4 different prediction models: Random Forest Classifier, Logistic Regression, KNN, and Neural Network.

First, we applied a **Random Forest Classifier** to a predictive modeling task. The Random Forest algorithm was selected for its proficiency in handling complex datasets with numerous features and its capability to mitigate overfitting. To prepare for modeling, the dataset was standardized using the StandardScaler. This step normalizes the feature set to ensure that each feature contributes equally to the distance computations in the model.

We initialized the Random Forest Classifier with a random state of 42 to ensure reproducibility. The model was then trained on the scaled training data. Moreover, we did not set up the limits of the depth of trees used, which would be beneficial to include for a comprehensive report.

A division of the dataset into feature sets and the target variable, 'purchase', was critical for training the model effectively. The feature set, composed of all columns except for the target, provided a comprehensive foundation for the model to learn and predict purchasing outcomes.

Upon testing, the model demonstrated a high level of accuracy at approximately 96.63% on the test set. This accuracy serves as a primary performance metric indicating the model's efficacy in classifying potential purchases. Moreover, the low number of false negatives 9 suggests that the model is exceptionally good at predicting the positive class. However, there were 553 false positives, which could be a focus for improvement. This could imply that the model may be too lenient in predicting the positive class and might benefit from further tuning of its probability threshold.

```
Confusion Matrix:
[[7757  553]
 [   9 8349]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.93      0.97      8310
           1       0.94      1.00      0.97      8358

    accuracy                           0.97     16668
   macro avg       0.97      0.97      0.97     16668
weighted avg       0.97      0.97      0.97     16668
```

Second, we have employed a **Logistic Regression model** to predict the binary outcome represented by the 'purchase' variable. It is suitable for binary classification tasks due to its efficiency and the interpretability of its coefficients.

In the data preprocessing, data also underwent a feature scaling using StandardScale to normalize the variables. After scaling, the dataset was split into a training set and a testing set with a test size of 30% and a random seed set for reproducibility.

As a result, the model's performance was evaluated on the scaled testing set, achieving an accuracy

```
Confusion Matrix:
[[8157  194]
 [ 767  326]]

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.98      0.94      8351
           1       0.63      0.30      0.40      1093

    accuracy                           0.90      9444
   macro avg       0.77      0.64      0.67      9444
weighted avg       0.88      0.90      0.88      9444
```

of approximately 89.82%. Looking at the confusion matrix and classification report, the matrix indicates that while the model is proficient at predicting the no purchase, it has a higher rate of false negatives than false positives, which implies that it is more conservative in predicting positive outcomes, and there is a notable discrepancy in the ability to predict actual purchases accurately, as seen in the lower recall and F1-score for class 1.

Third, we also utilized a **K-Nearest Neighbors (KNN) Classifier**, which is known for its effectiveness in classification tasks that have a reasonable number of dimensions. It classifies new cases based on a similarity measure with a specified number of nearest neighbors.

Data Processing and Model training are similar compared to Logistic regression except KNN model was set to 10 n_neighbors, indicating that the classification of a point is based on the majority vote of its ten nearest neighbors.

The KNN classifier was evaluated on the scaled test set, achieving an accuracy of approximately 88.49%. Moreover, from the confusion matrix and classification report, we can deduce that the model is more adept at identifying no purchases than purchases, as indicated by the higher precision, recall, and F1-score for class 0. The model exhibits low recall and F1-score, demonstrating a significant disparity in performance between the two classes

```
Confusion Matrix:
[[8205  146]
 [ 943  150]]

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.98      0.94      8351
           1       0.51      0.14      0.22      1093

    accuracy                           0.88      9444
   macro avg       0.70      0.56      0.58      9444
weighted avg       0.85      0.88      0.85      9444
```

Lastly, we used **Neural Network**, a deep learning known for its ability to model complex patterns and relationships within data.

This model also scaled the features using the StandardScaler to normalize the data ensuring that the model's input variables contributed equally to the analysis without any single feature disproportionately influencing the model's predictions.

Then, the model was constructed with three layers: input, hidden, and output. An input layer with a number of neurons matching the feature set dimensions and “relu” activation function to capture non-linear relationships. I set up a hidden layer with 5 neurons, and I set up an output layer with a sigmoid activation function which is mostly used for binary classification.

In addition, the model underwent training over 100 epochs with a batch size of 10, an approach that allows for sufficient model updating without being computationally intensive.

```
Confusion Matrix:
[[7921  430]
 [ 541  552]]

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.95      0.94      8351
           1       0.56      0.51      0.53      1093

    accuracy                           0.90      9444
   macro avg       0.75      0.73      0.74      9444
weighted avg       0.89      0.90      0.89      9444
```

As a result, the neural network achieved an accuracy of around 89.72% on the test data, indicating a high level of predictive accuracy. This suggests that the model is proficient in generalizing from the training data to unseen data. However, this model also has shown moderate precision and recall for purchases (class 1).

6. Testing Dataset

According to those accuracy scores, we selected the Random Forest Classifier as the best prediction model. Utilizing the model, the energy company is now equipped to more accurately forecast customers' purchase intentions. This allows for a refinement of the marketing strategies, directing efforts towards customers who exhibit a higher likelihood of purchasing the products.

In this section, we used a new test dataset (different from our original test dataset) with no target variable to run our prediction on to see what it looks like when we apply the prediction model in a real world setting. For example, customer 432176974 with the manager job, university degree, an account balance of 76, and a contact duration of 283 seconds was given a 65% of likelihood of purchasing intent. On the other hand, customer 432157692, despite similar demographics, only showed a 2% likelihood of purchasing intent, possibly due to a shorter contact duration.

	Test ID	Expected	Job	Account Balance	Education	Duration
0	432176974	0.65	manager	76	uniGraduated	283
1	432157692	0.02	manager	557	uniGraduated	111
2	432170850	0.59	technical	1274	uniGraduated	475
3	432151613	0.02	worker	986	highSchool	209
4	432167744	0.46	worker	3845	highSchool	459
...	...	...	...	...	...	...
13726	432162117	0.06	worker	409	highSchool	137
13727	432173177	0.06	worker	474	secondarySchool	152
13728	432159672	0.01	manager	57	uniGraduated	66
13729	432147170	0.00	manager	666	uniGraduated	253
13730	432184830	0.76	retired	3025	secondarySchool	166

13731 rows × 6 columns

If the expected value is greater than 50%, customers have a higher likelihood of purchasing a new product.

These predictions allow the energy company to prioritize their marketing initiatives, ensuring that efforts are concentrated on customers most receptive to their messaging. As a result, the company can more effectively allocate marketing resources, optimize engagement strategies, and maximize the potential for increased revenue stream, which is our ultimate goal.

id	Predicted
432176974	1
432157692	0
432170850	1
432151613	0
432167744	0
...	...
432162117	0
432173177	0
432159672	0
432147170	0
432184830	1

7. Potential Areas of Improvement

7.1 Standardization Before EDA

In the future, when analyzing large datasets like this one. It would be beneficial to standardize the variables before starting the exploratory data analysis. When looking at our graphs after we completed our EDA, it was often difficult to draw conclusions because the data was so unbalanced with significantly more people choosing not to purchase the product. By standardizing the data prior to this

step in analysis, we would have a more accurate idea of how different demographics react to this campaign and have a greater capacity to make comparisons.

7.2 Finding a Higher Quality Dataset

One of the biggest challenges when analyzing this dataset was that a lot of important information was missing. We found out what industry we were working in by contacting the creator of the dataset, but were given little information on how this data was collected or even what product was being purchased.

Another big hindrance within our dataset was the diversity of dimensions. Almost all dimensions focus on the demographics of a consumer rather than differences in how a person received information or what happened during the contact period. With telemarketing, every interaction is different so what happens during a call and differences between sales representatives are just as important as differences between target consumers. Most of our data pointed us to similar conclusions: people who can afford more can buy more. Having a greater diversity of dimensions and focusing more on what is going right in these phone conversations rather than just demographics will give us a better understanding of how our company can improve. Why do some people want to stay on the call longer and how do we make people interested?

7.3 Power Calculations

A final way that we would have been able to better analyze this data is if we calculated the power of this campaign and our analysis. Calculating power would have given us an understanding of the probability that we were coming to the correct conclusion while using this data. Power also allows for us to determine what number of subjects we would need to see results with significance. In future campaigns, it would be advised to figure out what the company wants to learn and include enough people in the campaign to go beyond that designated threshold.

8. Conclusion

This project has been an interesting and creative way to apply what we have learned in class to a real world situation. We have learned how to navigate issues that arise during data analysis and how to communicate the important information that we found through data analysis. We hope to bring these skills we have learned to other projects and when we move into the workforce.

Looking back on this entire project, it was very satisfying to see everything come together and see our predictive model work in real time. Even when we had to navigate through not knowing what the product was, it was still really interesting to see how our model would predict whether or not a new customer would purchase based on certain factors. For example, increasing the duration of the call with the target consumer, targeting customers who are older, more educated (with college degree), with a high account balance (higher than 500), who are also more willing to give their contact type, and contacting at the end of the month (greater day) might increase the chance of purchase intent. We are all very proud of how the project turned out and we worked really hard to complete it.

9. Appendix

Feature	Type	Description
id	Numerical	record ID
purchase	Nominal	target value (making new purchase or not after the marketing campaign)
day	Numerical	contact day in previous campaign
month	Nominal	contact month in previous campaign
duration	Numerical	contact duration in previous campaign
contactId	Numerical	contact ID
age	Numerical	age of the customer
gender	Nominal	customer gender
job	Nominal	customer occupation
maritalStatus	Nominal	customer marital status
education	Nominal	customer educational degree
craditFailure	Nominal	if the customer has a default credit
accountBalance	Numerical	customer account balance
house	Nominal	if the customer owns a house
credit	Nominal	if the customer has a credit
contactType	Nominal	contact media
numberOfContacts	Numerical	number of contacts during the current campaign
daySinceLastCampaign	Numerical	days after the last contact of the previous campaign
numberOfContactsLastCampaign	Numerical	number of contacts during the previous campaign
lastCampaignResult	Nominal	result of the previous campaign

Dataset Sources:

Background: https://www.kaggle.com/code/khanimar/bi-marketing-campaign-eda-analysis-prediction
Training Dataset:
https://www.kaggle.com/code/khanimar/bi-marketing-campaign-eda-analysis-prediction/input?select=train.cs