Predicting New Customers' Purchase Decision on Energy Product

Joanne Charles, Cristina Jiang, Kaya Manolt, Jaden Cho, Jaden Noh

Agenda

1. Intro to Our Dataset
2. EDA
3. Feature importance
4. Logistic Regression
5. Our Recommended Model
6. Testing and Conclusion

Our Main Goal

1.Factors influencing the purchase decision

2. Build prediction models to increase efficiency and success rate of the marketing campaign

Increasing Revenue

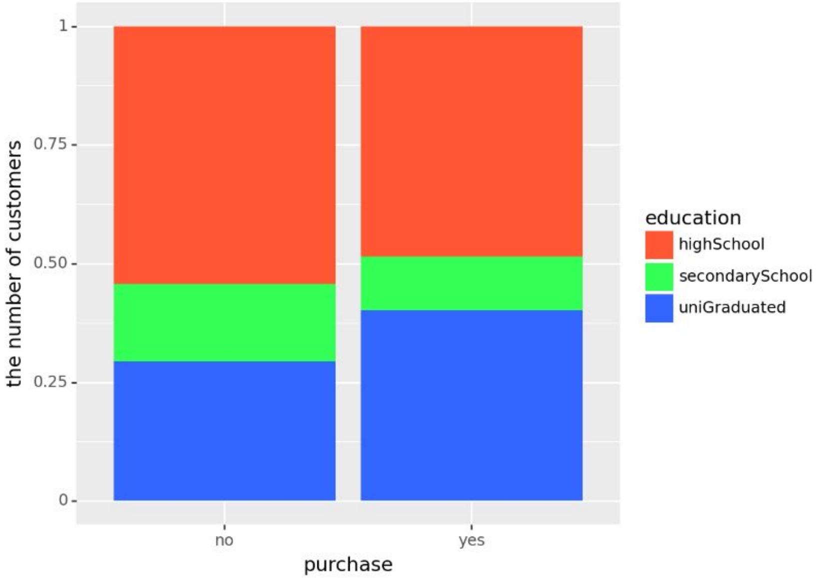
Our Dataset – 31,480 target customers

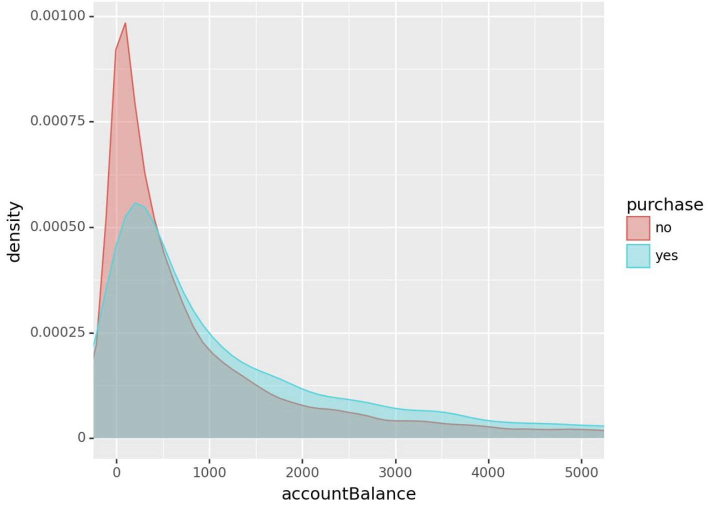
Features

ID	Education	Job	Account Balance	Credit Failure	Marital Status	Last Campaign Result

Day	Days Since Last Campaign	Contact Type	Age

People who are more educated with higher account balance are more likely to purchase





Our dataset is not perfect yet, preprocessing is important!

Preprocessing our dataset to gain more accuracy and clarity

Step 1

Drop Column?

For EDA: All dimensions are included

For Prediction Model: dropped the 'daySinceLastCampaign' 'lastCampaignResult', "contactid" And "id"

(too many null values and are irrelevant)

Step 2

Rename

Changed the name of target variable from "target" to "purchase"

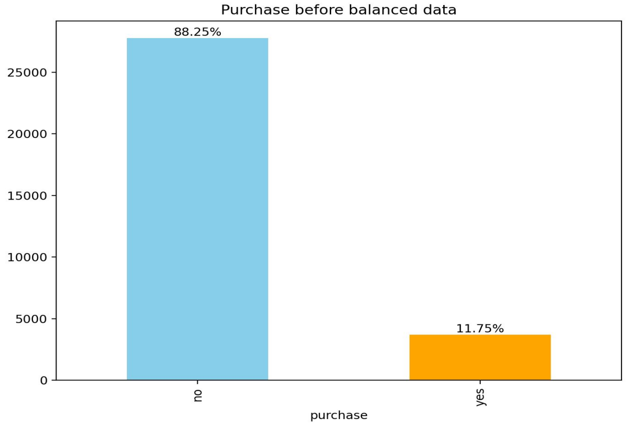
It measures whether or not a target customer made a purchase

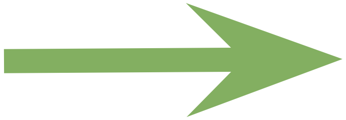
Step 3

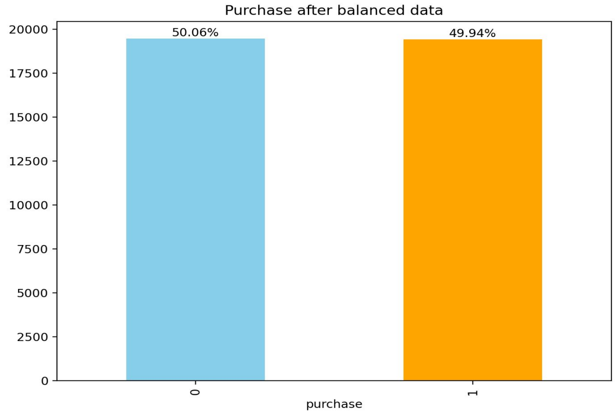
One Hot Encoder

Convert our categorical data into numerical values (binary variables)

Our dataset was imbalanced, “purchase_yes” was the minority







- **The original target variable (purchase) was not balanced, with 88.25% who did not make the purchase and 11.75% who made the purchase**

- **To ensure best prediction results, we used <u>“Random Over Sampler”</u>**

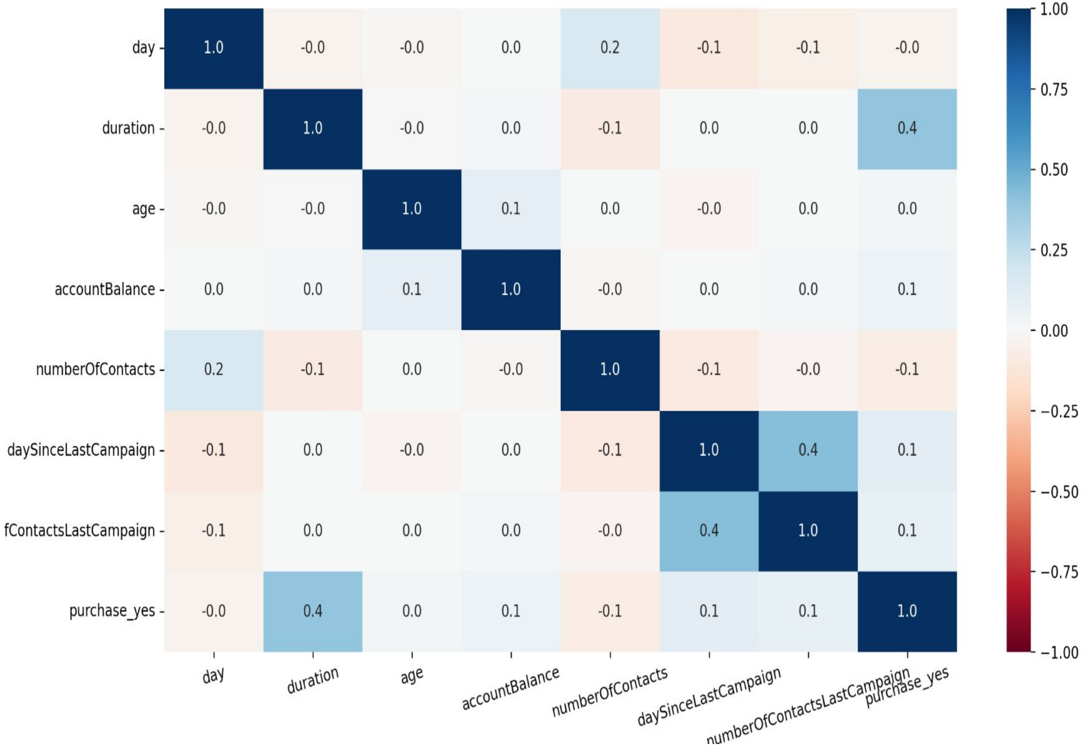
- **It balances the class distribution, so that each class has a similar proportion (as shown in the right graph)**

No strong correlations among variables are identified

To avoid multicollinearity that can affect our prediction, we checked the correlations

No strong correlation is identified among any variables

No variable need to be dropped!



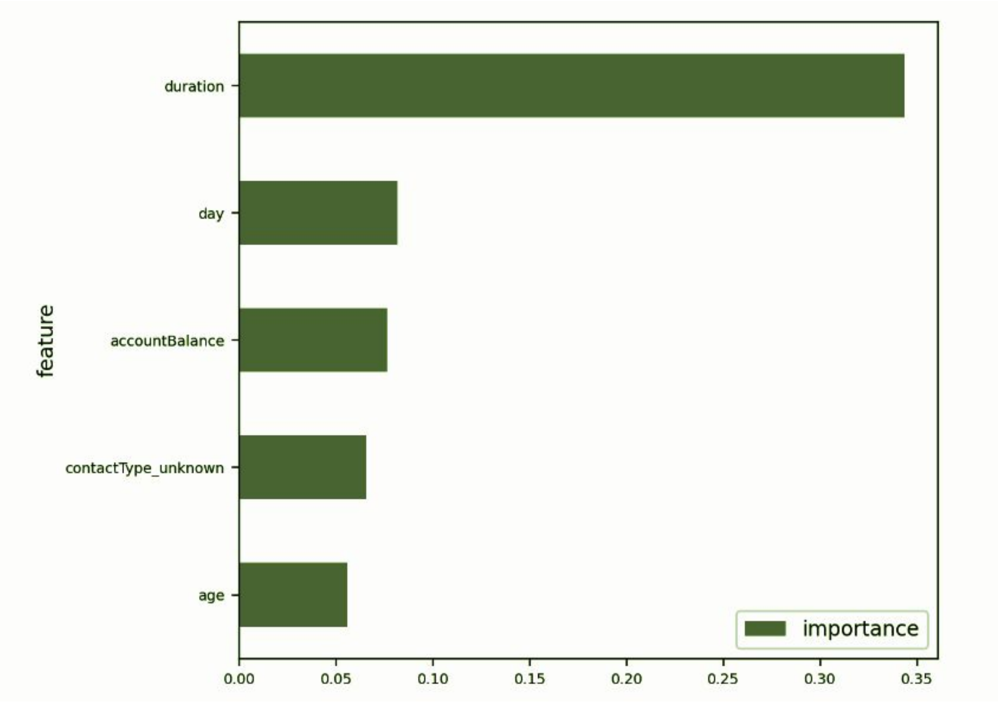
Determine the Most Important Variables We Should Focus on

Using Decision Tree Classifiers to Assess Importance

The **duration of a phone call** had the most significant impact on whether or not a targeted potential customer chose to make a purchase with our company.

Day, **Account Balance** and **Age** had a noticeable impact on purchasing decisions which can be attributed more to whether or not a target can afford the product rather than interest.

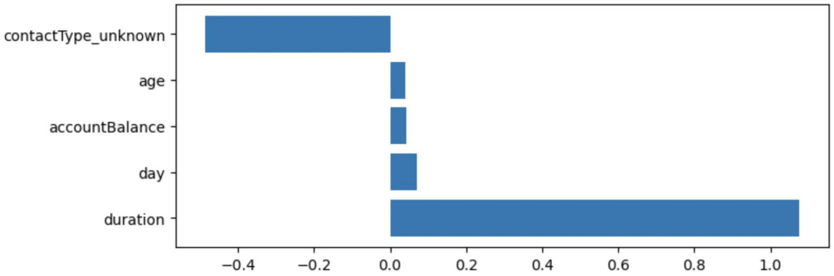
Using this classifier, we were able to narrow down our **41** dimensions to the **5** most important for our logistic regression analysis.



Logistic Regression Further Highlights the Impact of Duration

Using the dimensions highlighted in our decision tree classifier, we created a logistic regression which better showcases the direction and magnitude of important dimensions on purchase.

Duration and Unknown Contact Types have the most direct influence. In future contacts, sales representatives should place their focus on maintaining client interest and attention and familiarizing themselves with new forms of communication.

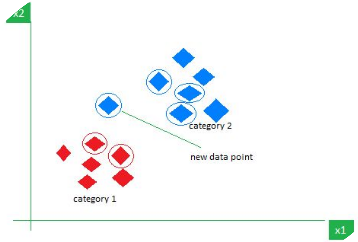


	features	coef	std err	z	P>|z|	[0.025	0.975]	exp_coef
0	day	0.0718	0.029	2.493	0.013	0.015	0.128	1.074440
1	duration	1.0763	0.023	47.611	0.000	1.032	1.121	2.933804
2	age	0.0388	0.033	1.158	0.247	-0.027	0.104	1.039563
3	accountBalance	0.0431	0.022	1.994	0.046	0.001	0.085	1.044042
4	contactType_unknown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

We Built Models to Optimize Marketing Campaign Success Rate

We built different prediction models and tested their accuracy

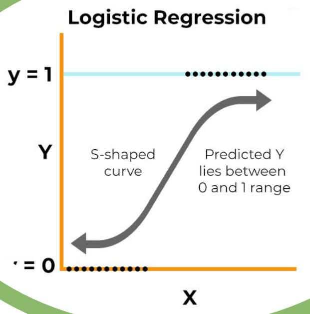
all higher than 88.25% (Naive rule)



88.49%

KNN

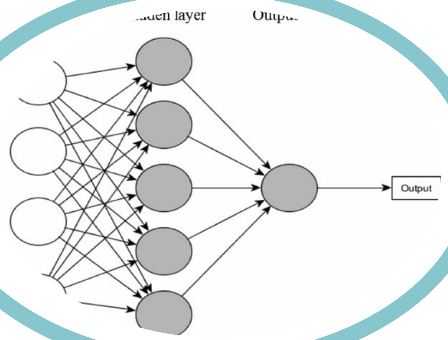
1.StandardScaler to normalize the variables
2.Test size: 30%
3.10 n_neighbors



89.92%

Logistic Regression

1.StandardScaler to normalize
2.Test size: 30%
3. Higher rate of false negatives
4. More conservative at predicting positives



89.72%

Neural Network

1.StandardScaler to normalize
2.Hidden layer with 5 neurons
3.Output layer: sigmoid activation
4.Underwent training over 100 epochs with a batch size of 10

Random forest turns out to be the best model

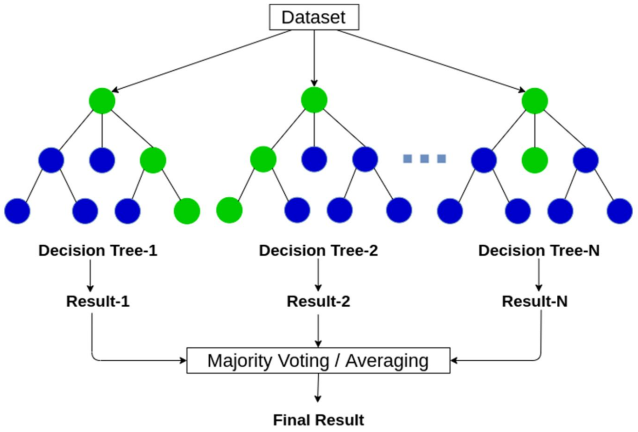
Data Preparation:

To prepare for modeling, the dataset was standardized using the StandardScaler

Random Forest Classifier with a random state of 42

1. Trained on the scaled training data
2. Testing: the model demonstrated a high level of accuracy at approximately **96.63%** on the test set.
3. Results: **533 false positives**: the model may be too lenient in predicting the positive class
4. Future: further fine tune this model

Random Forest



Marketing campaign strategies to increase success rate (Key Insights)

Likely to Increase Purchase Intent

Campaign

- Increasing Duration of Call

Target Consumer

- Older
- More Educated
- High Account Balance

Characteristics

- Willing to give their contact type
- Contacted at the end of the month

If you were our marketing manager, who would you target?

1 **Customer 1**
432184585

2 **Job**
Retired

3 **Account Balance**
$8,044

4 **Duration**
702 seconds

1 **Customer 2**
432146206

2 **Job**
Manager

3 **Account Balance**
$19

4 **Duration**
56 seconds

Results!

Customer 1

Expected **0.65**

Test ID	Expected	Job	Account Balance	Education	Duration
432184585	0.65	retired	8044	secondarySchool	702

LIKELY TO PURCHASE

Customer 2

Expected **0.01**

Test ID	Expected	Job	Account Balance	Education	Duration
432146206	0.01	manager	19	secondarySchool	56

NOT LIKELY TO PURCHASE

Methods of Improvement for Future Studies

1. **<u>Increase Focus on the *Quality* of Phone Calls:</u>** Who Are the Representatives that are Grabbing People's Attention and How Can We Develop That in Others?
2. **<u>Decrease Focus on Economic Factors:</u>** People Who Can't Afford the Product, Can't Buy It. Don't Base the Majority of Dimensions on Related Factors You Can't Control
3. **<u>Test for Power of Experiment:</u>** How Likely Are We to Draw the Correct Conclusion From this Data?

Conclusion

We have designed new marketing campaign strategies to optimize revenue by:

1. Identifying the target consumer characteristics
2. Using random forest as prediction model with high accuracy rate

This project was a creative way to apply what we have learned in class to a real world situation. We've learnt to analyze data and communicate analysis effectively.

Any Questions?