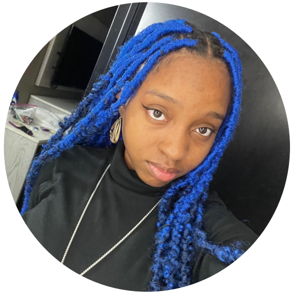
KPMG 1A: Driving Donations
AI Studio Final Presentation

December 5th, 2024

Meet our team!

Julia Husainzada
San Jose State University



Imani Cage
Seattle University

Janelle Chan
University of Washington

Ezuma Ekomo Ble
San Jose State University

Our AI Studio TA and Challenge Advisors

Yi Tong
AI Studio TA

Morgan Abbitt
Challenge Advisor

Sam X. Tan
Challenge Advisor

Table of Contents

01

Background on Project Focus

About C5LA

- 501(c)(3) charitable non-profit organization
- Youth Leadership
- Summer Camp + Hiking
- College/Career access and success program
- Push students out of comfort zone with fun outdoor activities and foster higher education and initiative in the community
- Fun Fact: Founded by Coca Cola CEO!

The goal was simple: help C5LA increase donations. How? Identify who is likely to donate again.

Business Impact

- Improved donor retention strategies
- Targeted Marketing and Outreach Campaigns
- Review of past decade’s success in growing donations

Our Approach

Explore data and
possible features

**Sprint 1:
Data Exploration**

Sprint 2: Data Preprocessing

- Clean and preprocess data
- Handle missing values
- Feature engineering

- Select techniques for ML model
- Validate and enhance model

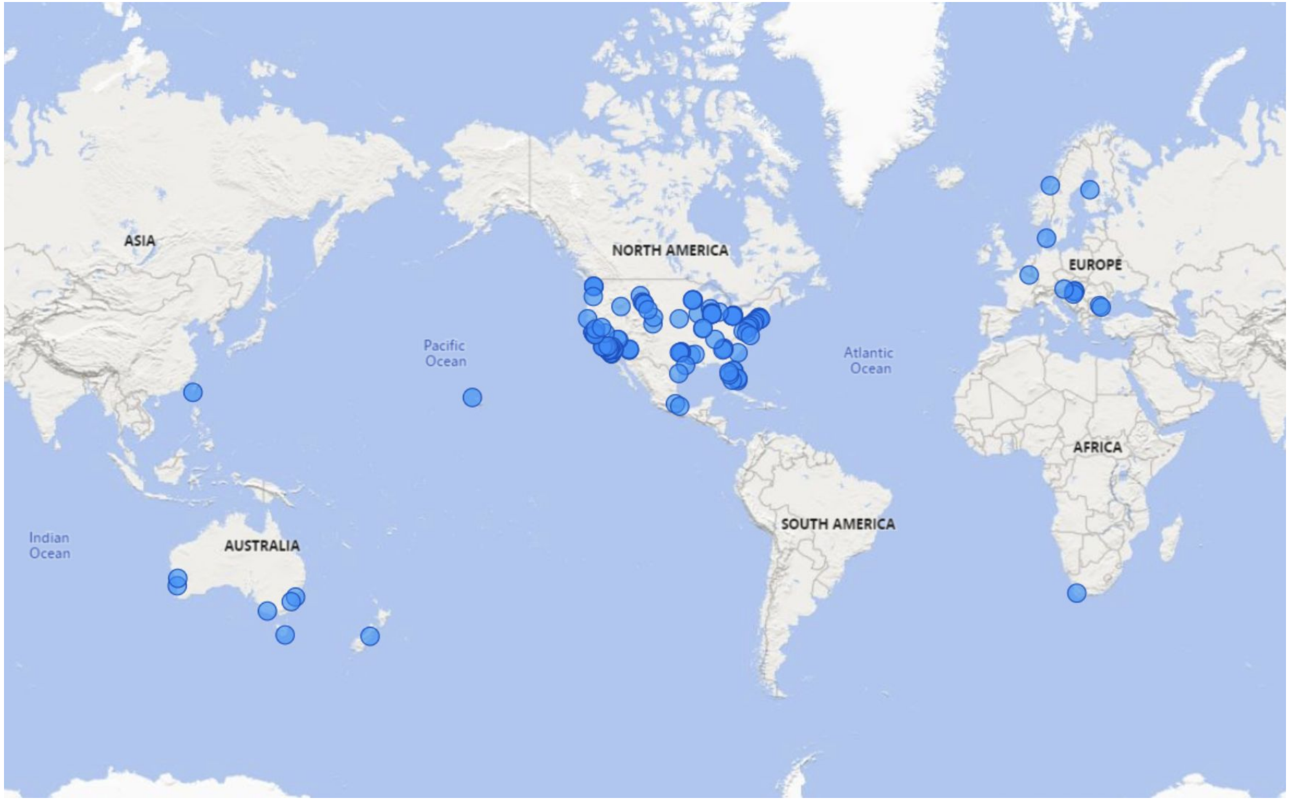
**Sprint 3: Data Analysis
and Modeling**

Sprint 4: Finalize Product

02

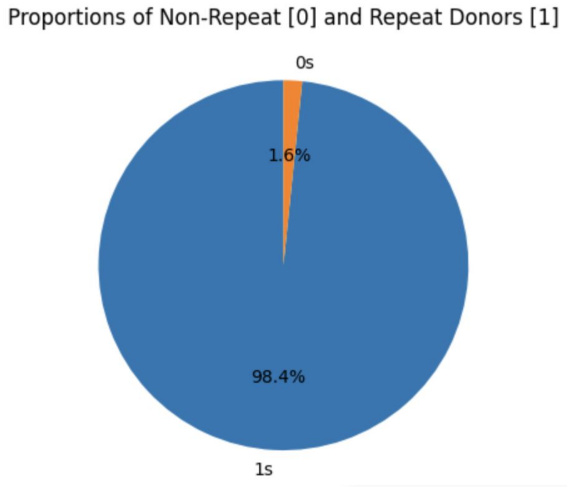
Data Preparation and Analysis

Where are the Donations From?

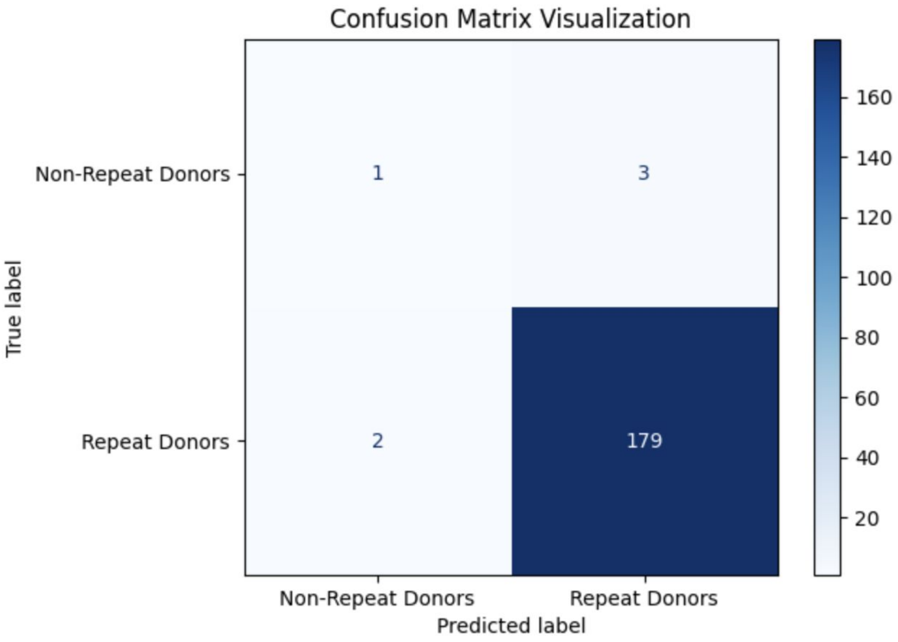


Summary of Data Collected

- 3122 donations collected between 2014 and 2021
- 616 unique donors
- 606 repeat donors
- 10 non-repeat donors



Summary


Confusion Matrix Visualization
Non-Repeat Donors
Repeat Donors
True label
1
3
2
179
Non-Repeat Donors
Repeat Donors
Predicted label
160
140
120
100
80
60
40
20

Data Analysis Key Takeaways

- Mapped ZCTA and zip codes to see how many times a donor's name popped up, which gave us intel on whether they were a repeat donor
- Did feature engineering to create repeat_donor feature
- Amount donated was highly correlated with whether someone would be a repeat donor
- Processed features to get percentage instead of absolute number for demographics data

03

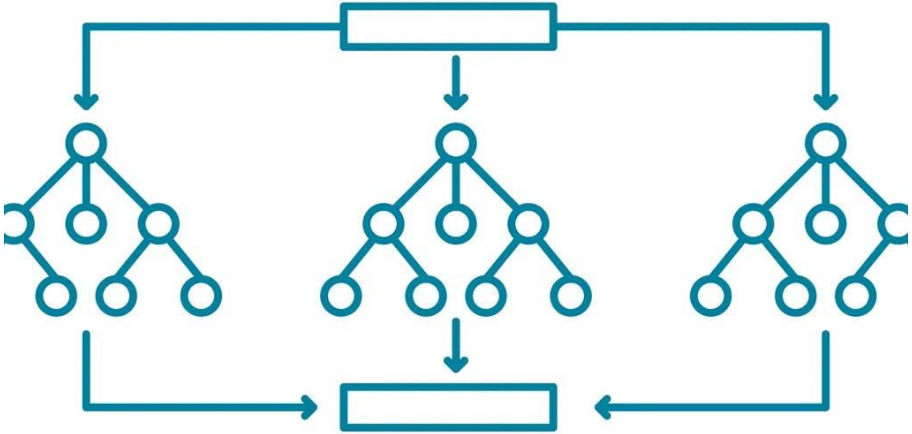
Building and Improving the Model

Overview of RFE

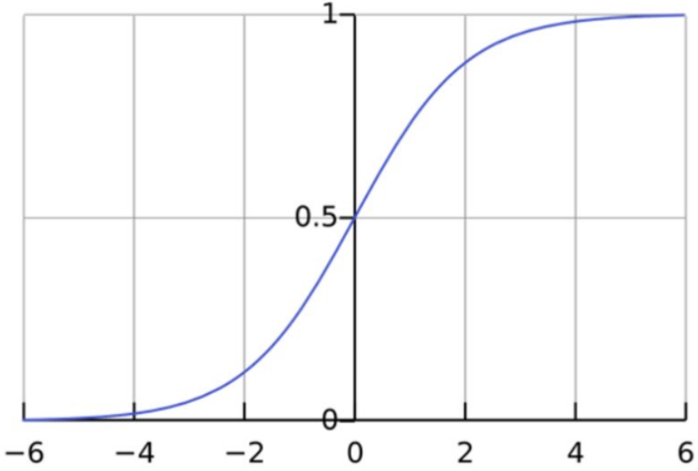
- Popular approach to choosing features is selecting ones that have high correlation with label
- RFE (Recursive Feature Elimination) trains model multiple times to eliminate the weakest label each time
- More powerful than selecting features with high correlations because it guarantees the features that remain will improve model performance
- Has to be performed during training stage, not data exploration stage

Modeling Components and Comparisons

'Billing Zip/Postal Code', 'ZCTA', 'Education Years', 'High school %', 'large_donation_flag', 'Account Type_Household'



Decision Tree



Logistic Regression

Model Comparison

Model Name	Description	Results	Pros	Cons
RF Decision Tree	Creates many branches to evaluate if a donor will donate again	98% precision :) Highly predicted people would donate again	Handles large datasets well	Prone to overfitting We have a small dataset
Logistic Regression	Does regression to classify beyond a threshold if someone will donate again	60% average accuracy...	Easy to set up Efficient training	Assumes linearity in relationships Bad for large datasets

Modeling Key Takeaways

- People are very much likely to donate again
- The amount donated is confirmed to be the best determinant of whether someone donates again
- %educated and %went to college were also important factors

04

Summary and Next Steps

What We Learned

- Initial data processing is most important
- Determining the right features can make or break a model's performance
- Overfitting can easily happen with smaller datasets (like in our case)
- If our model just always predicted a donor will donate, it would be correct 98% of the time. C5LA is killing it!

Improving Model Performance

- Removing outliers … lowered model performance
- Using mean to impute null values improved performance by 1x10^-16

Controlled experiment - change one for each model

1. No outliers
2. Using mean to impute nulls
3. Features
   a. Try model 1: remove campaigns
   b. Try model 2: remove close year, close day
   c. Try model 3: reduce features to 6, or keep reducing features to see optimized performance
4. param_grid
   a. change parameters

Next Steps

- More data points focused on donors who did not donate again could help us identify trends in those who don't become repeat donors
- We have yet to explore how the following 3 changes may affect model performance:

3. Features
    a. Try model 1: remove campaigns
    b. Try model 2: remove close year, close day
    c. Try model 3: reduce features to 6, or keep reducing features to see optimized performance
4. param_grid
    a. change parameters

Resources we Leveraged

We primarily used Jupyter Notebook, GitHub, Python, scikit-learn, PowerBI, and pandas

Bonus: Appendix

Precision/Recall Data

	precision	recall	f1-score	support
0	$\boxed{0.33}$	0.25	0.29	4
1	0.98	0.99	0.99	181
accuracy			0.97	185
macro avg	0.66	0.62	0.64	185
weighted avg	0.97	0.97	0.97	185

AUC-ROC: 0.951657458563536

Out of all the predictions where the model predicted non-repeat donors, only **33%** were actually correct.

Precision/Recall Data

	precision	recall	f1-score	support
0	0.33	0.25	0.29	4
1	0.98	0.99	0.99	181
accuracy			0.97	185
macro avg	0.66	0.62	0.64	185
weighted avg	0.97	0.97	0.97	185

AUC-ROC: 0.951657458563536

Only 25% of the actual non-repeated donors were correctly identified by the model.

Precision/Recall Data

	precision	recall	f1-score	support
0	0.33	0.25	0.29	4
1	0.98	0.99	0.99	181
accuracy			0.97	185
macro avg	0.66	0.62	0.64	185
weighted avg	0.97	0.97	0.97	185

AUC-ROC: 0.951657458563536

Out of all predictions labeled as repeat donors, 98% were correct.

Precision/Recall Data

	precision	recall	f1-score	support
0	0.33	0.25	0.29	4
$\boxed{1}$	0.98	$\boxed{0.99}$	0.99	181
accuracy			0.97	185
macro avg	0.66	0.62	0.64	185
weighted avg	0.97	0.97	0.97	185

AUC-ROC: 0.951657458563536

The model successfully identified 99% of the actual repeat donor instances.

Questions?

Thank you everyone :)