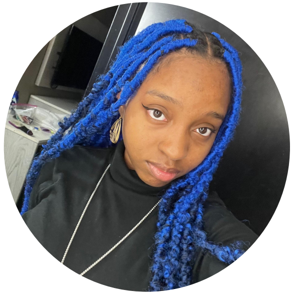
KPMG 1A: Driving Donations

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 team

Yi Tong

AI Studio TA

Morgan Abbitt

Challenge Advisor
Manager

Sam X. Tan

Challenge Advisor
Senior Specialist

Table of Contents

01

Background on Project Focus

About C5LA

501(c)(3) Charitable Nonprofit Organization

Committed to empowering underrepresented youth through leadership and education.

Key Programs & Initiatives Youth Leadership

Push students out of comfort zone with fun outdoor activities and foster higher education and initiative in the community

- Summer camp & hiking experience
- College/career access and success program

Our goal was simple: to help C5LA increase donations by identifying individuals likely to donate again.

Business Impact

Retention

Improved donor retention strategies

Engagement

Targeted Marketing and Outreach Campaigns

Analysis

Review of past decade's success in growing donations

Our Approach

Sprint 1: Data Exploration

Explore data and possible features

Sprint 2: Data Preprocessing

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

Sprint 3: Data Analysis and Modeling

- Select techniques for ML model
- Validate and enhance model

Sprint 4: Finalize Product

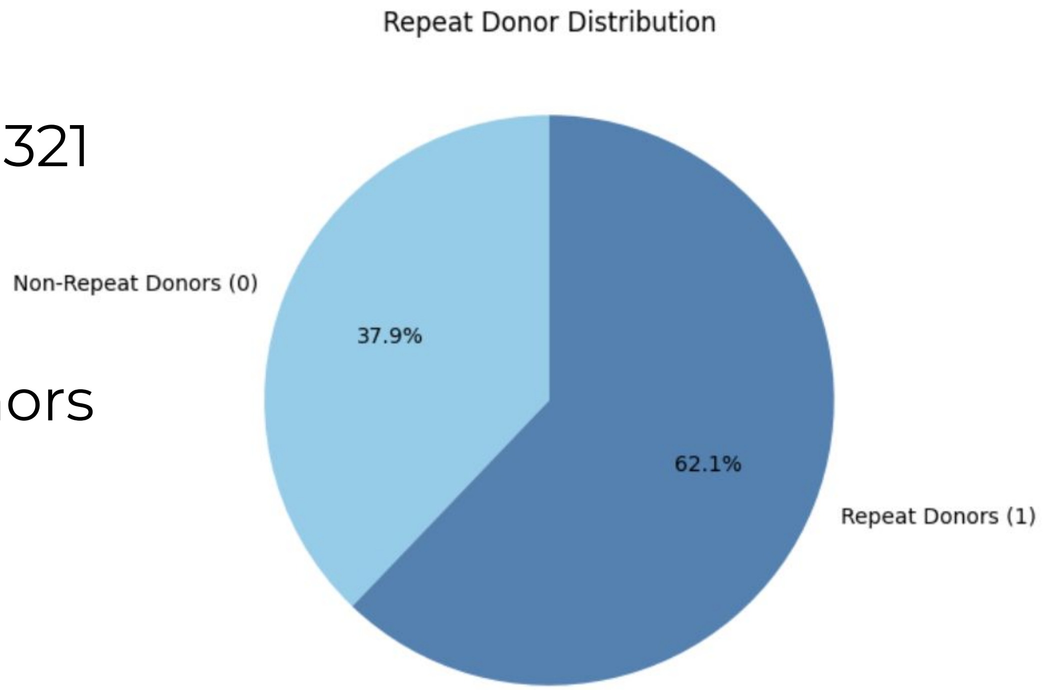
02

Data Preparation and Analysis

Summary of Data Collected

3122 donations, where 1321 are unique donors

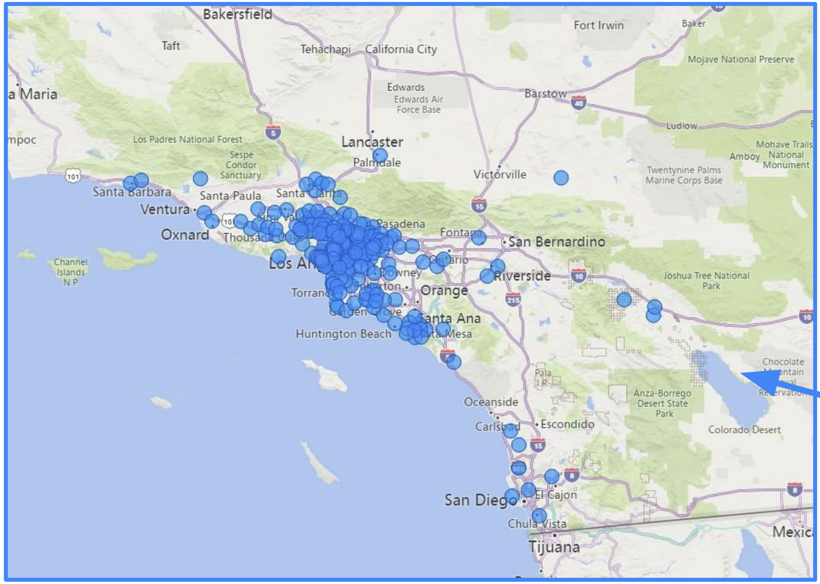
- 821 repeat donors
- 500 non-repeat donors

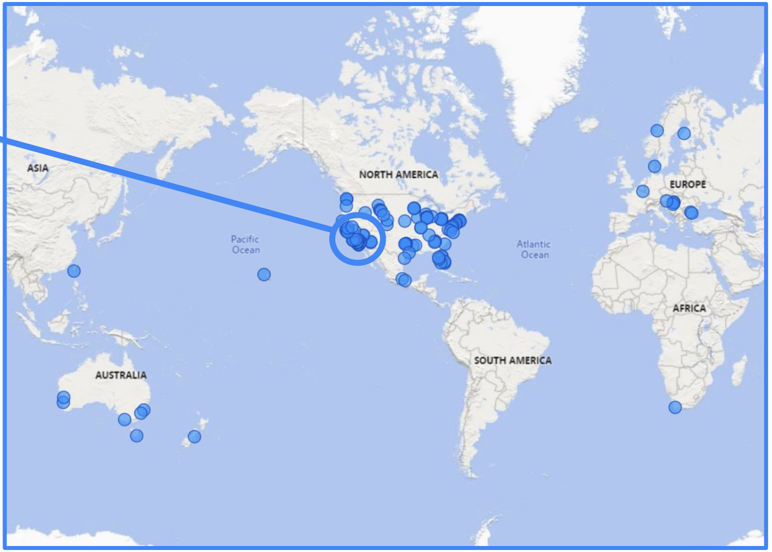


Summary of Data Collected

- Donor data included masked account ID, donation amount, account type, and donation date.
- Combined with a secondary dataset mapped to ZCTAs, creating 80 features on zip code demographics, such as average education years, unemployment rates, and health insurance coverage.

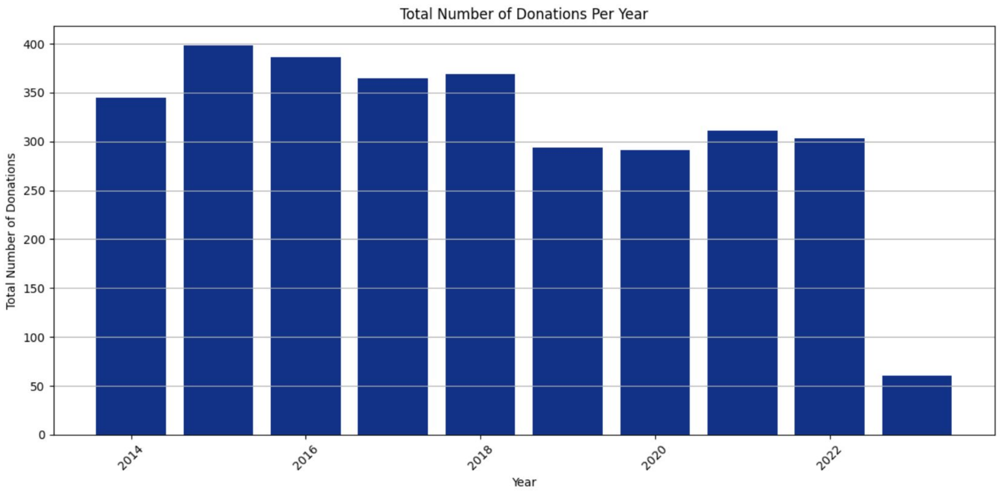
Where are the Donations From?

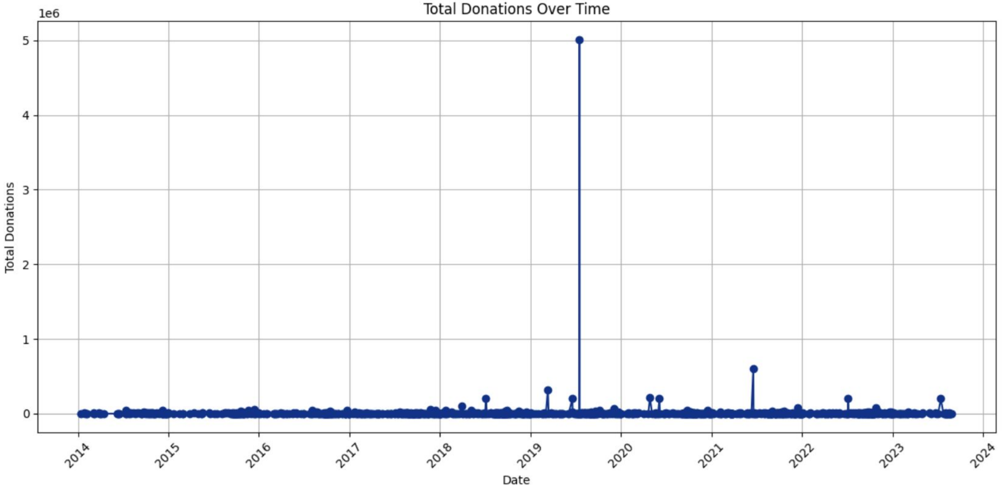




Donations Timeline

3122 donations collected between 2014 and 2021





Label creation

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

Masked Account ID	Maked Pr	Stage	Account T	Billing Zip	Fiscal Peri	Close Dat	Amount
0001	None	Closed W	Househol	90069	Q3-2022	7/5/2022	200000
0001	None	Closed W	Househol	90069	Q3-2023	7/14/202	200000
0001	None	Closed W	Househol	90069	Q3-2019	7/17/201	5000028.
0001	None	Closed W	Househol	90069	Q1-2019	3/12/201	320000
0001	None	Closed W	Househol	90069	Q2-2020	4/27/202	220000
0001	None	Closed W	Househol	90069	Q3-2018	7/5/2018	200000
0001	None	Closed W	Househol	90069	Q2-2019	6/19/201	200000
0001	None	Closed W	Househol	90069	Q2-2020	6/4/2020	200000
0001	None	Closed W	Househol	90069	Q1-2018	3/30/201	100000
0001	None	Closed W	Househol	90069	Q2-2018	5/8/2018	50000
0001	None	Closed W	Househol	90069	Q4-2019	10/7/201	40000
0001	None	Closed W	Househol	90069	Q1-2018	2/16/201	25000
0001	None	Commitm	Househol	90069	Q2-2021	6/18/202	200000
0001	None	Closed W	Househol	90069	Q2-2021	6/18/202	200000
0001	None	Closed W	Househol	90069	Q2-2021	6/18/202	200000

	Masked Account ID	Maked Primary Campaign	Stage	Account Type	Billing Zip/Postal Code	Fiscal Period	Close Date	Amount	ZCTA	Year	...	Labor Population	Armed Forces	Employed	Unemployed	Not in Labor Force	large_donation_flag	Close Year	Close Month	Close DayOfWeek	repeat_donor
0	1	Campaign One 2015	Closed Won	Household	90069	Q3-2022	2022-07-05	1.769909	90069.0	2021.0	...	19125.0	0.0	14080.0	1124.0	3921.0	1	2022	7	1	1

03

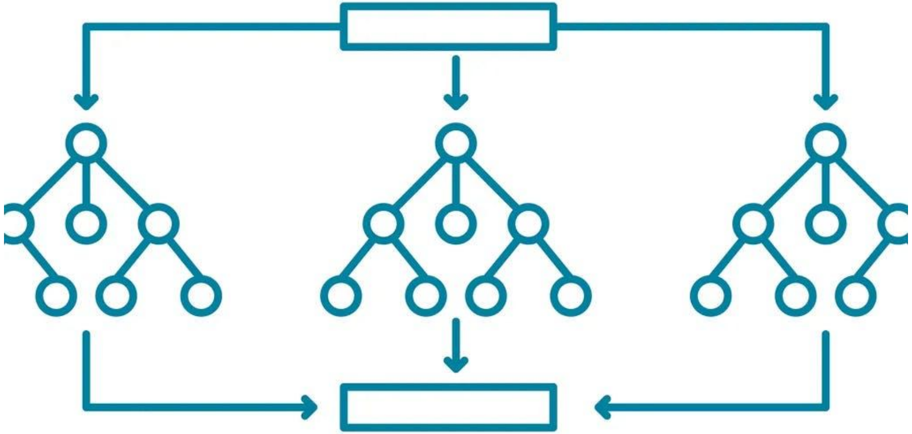
Building and Improving the Model

Overview of RFE

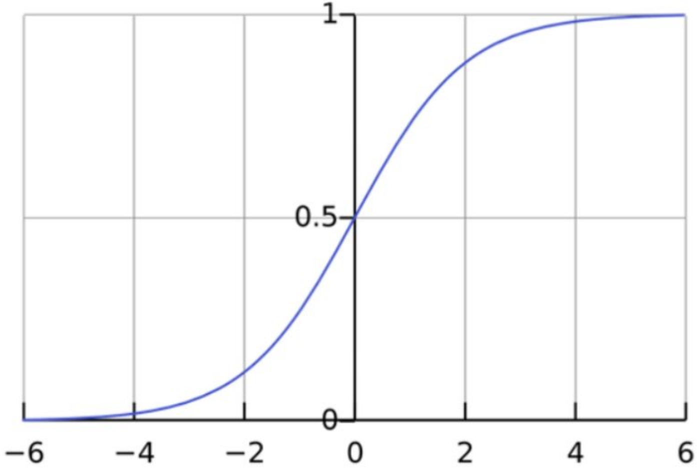
- Performed during training stage
- Popular approach to choosing features is selecting ones that have high correlation with label
- RFE (Recursive Feature Elimination) trains model multiple times to eliminate the weakest label each time

Modeling Components and Comparisons

'Billing Zip/Postal Code', 'ZCTA', 'Year', 'Minority %', 'Poor Family %', 'No Health Ins %', 'Unemployed %', 'Close Year', 'Close Month', 'Close DayOfWeek'



Decision Tree



Logistic Regression

Model Comparison

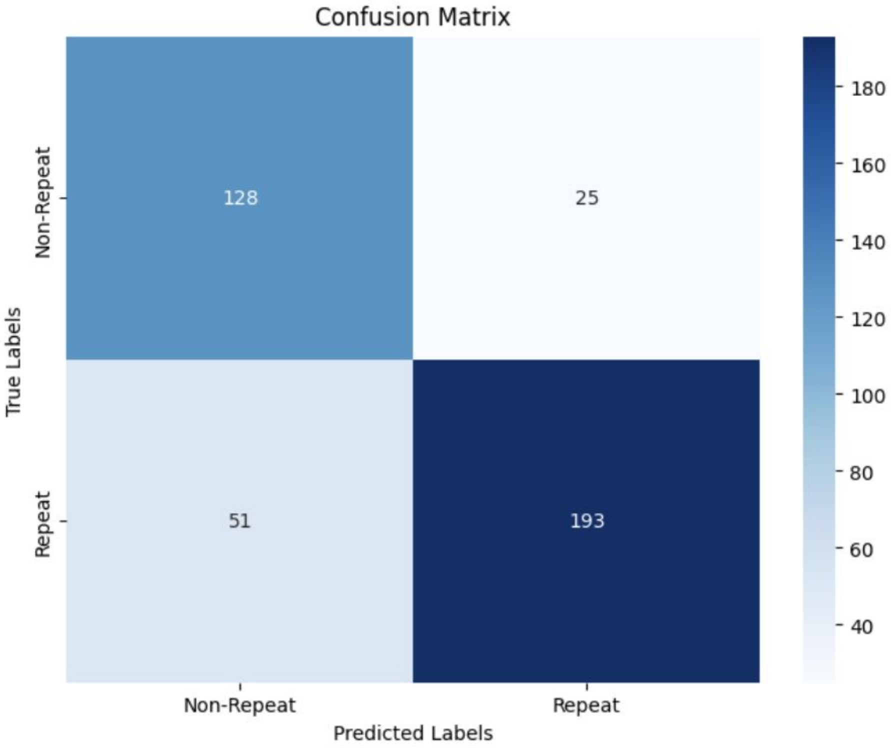
Model Name	Description	Results	Pros	Cons
RF Decision Tree	Creates many branches to evaluate if a donor will donate again	88% accuracy :) Highly predicted people would donate again	Handles large datasets well	Prone to overfitting
Logistic Regression	Does regression to classify beyond a threshold if someone will donate again	60% average accuracy...	Easy to set up Efficient training	Assumes linearity in relationships Bad for large datasets

Results

Classification Report:

	precision	recall	f1-score	support
0	$\boxed{0.72}$	$\boxed{0.84}$	0.77	153
1	$\boxed{0.89}$	$\boxed{0.79}$	0.84	244
accuracy			0.81	397
macro avg	0.80	0.81	0.80	397
weighted avg	0.82	0.81	0.81	397

AUC-ROC: 0.8812680810028929



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

Recommendations

- Donations are most likely to come from LA area, where the youth are impacted - C5LA could continue to grow its local presence
- Follow up with donors who donate large amounts - they are likely to donate again and with at least the same donation amount
- Donors from highly educated regions are most likely to donate again - follow up with donors from those Zip Codes

Resources we Leveraged

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

Bonus: Appendix

RandomizedSearchCV

best_estimator_: RandomForestClassifier

RandomForestClassifier

```
RandomForestClassifier(max_depth=40, min_samples_leaf=2, random_state=42)
```

Precision/Recall Data

	precision	recall	f1-score	support
0	0.72	0.84	0.77	153
1	0.89	0.79	0.84	244
accuracy			0.81	397
macro avg	0.80	0.81	0.80	397
weighted avg	0.82	0.81	0.81	397

Out of all the predictions where the model predicted non-repeat donors, **72%** were actually correct.

Precision/Recall Data

	precision	recall	f1-score	support
0	0.72	0.84	0.77	153
1	0.89	0.79	0.84	244
accuracy			0.81	397
macro avg	0.80	0.81	0.80	397
weighted avg	0.82	0.81	0.81	397

84% of the actual non-repeated donors were correctly identified by the model.

Precision/Recall Data

	precision	recall	f1-score	support
0	0.72	0.84	0.77	153
1	0.89	0.79	0.84	244
accuracy			0.81	397
macro avg	0.80	0.81	0.80	397
weighted avg	0.82	0.81	0.81	397

Out of all predictions labeled as repeat donors, 89% were correct.

Precision/Recall Data

	precision	recall	f1-score	support
0	0.72	0.84	0.77	153
1	0.89	0.79	0.84	244
accuracy			0.81	397
macro avg	0.80	0.81	0.80	397
weighted avg	0.82	0.81	0.81	397

The model successfully identified 79% of the actual repeat donor instances.

Questions?

Thank you everyone :)