KPMG 1C: Driving Donations
AI Studio Final Presentation

Break Through Tech Virtual
12/05/2024

Presentation Agenda

- Introductions
- AI Studio Project Overview
- Data Understanding & Data Preparation
- Modeling & Evaluation
- Final Thoughts
- Q&A

Introductions

Meet Our Team!

Nour Khalifa
LSU

Angela Sidhu
Georgetown University

Janya Bhaskar
University of Colorado Boulder

Dasha Dziadzinets
Rice University

Mmesoma Ezeudu

AI Studio Project Overview

Our Project Partners

KPMG: A Big Four accounting and professional services firm

Mission: To inspire confidence and empower change

Goal: Derive meaningful insights from the data and provide C5LA with practical, actionable recommendations to drive business value.

C5LA: A 501(c)(3) nonprofit

Mission: Aid high-potential teens from under-resourced communities by inspiring them to pursue success and preparing them for leadership roles through programming on leadership, mentorship, and professional development

Goal: To analyze donor and donation trends to better understand their donors and aid future outreach

Given the donation history of individual households and corporations to a non-profit organization, use machine learning to identify individuals likely to donate again.

Predicting Future Donations to C5LA

Goal: Develop a binary classification model that predicts whether individual households and corporations are likely to make a future donation to C5LA.

Business Need: Nonprofits rely on donations from households and corporations to fund their programs. However, donor attrition is a common challenge. Predicting the likelihood of future donations enables C5LA to:

- Improve donor retention by increasing engagement with donors who are likely to donate again.
- Optimize resource allocation for outreach efforts.

Problem Statement

Problem: Using previous donation data and donor attributes, predict whether a donor will make another donation within 12 months.

Output: Binary prediction (1 = likely to donate, 0 = unlikely to donate).

Scope: Inputs include donation history, socioeconomic and demographic information, and outputs include binary predictions.
Assumptions: Predictions are based on a past donor behavior, so the model is based off that select data.

Data Understanding & Data Preparation

Data Requirements

Data Sources: Three data files from the KPMG team

Donation History: Donor ID, donation amounts, donation dates.

Sociodemographic Data: Income, education, employment, region (zip code and MSA).

Assumptions: The model's predictions are based on past donor behavior, using historical donation data and sociodemographic information to forecast future donations.

Data overview

Number of Rows: 3122

Donation Amounts:

- Range from $1 to $5,000,029
- Had two negative outliers for -$200 and -$75 and multiple donations of $0
- Negative outliers are likely people accidentally donating more than intended

Dates:

- Span from Q1-2014 (1/1) to Q3-2023 (8/28)
- Sociodemographic data is available from 2019-2021 only

Donation Account Type:

- 2659 Households
- 463 Corporate

Data Preprocessing

Handling Outliers:

- Negative donations and $0 donations will be cleaned.

Missing Data:

- Identify and handle missing values in sociodemographic data and donation history.

Data Merging:

- Combine donor data with sociodemographic and geographic information.

Feature Engineering

Donation History Features:

- Number of donations, average donation amount, and maximum donation.

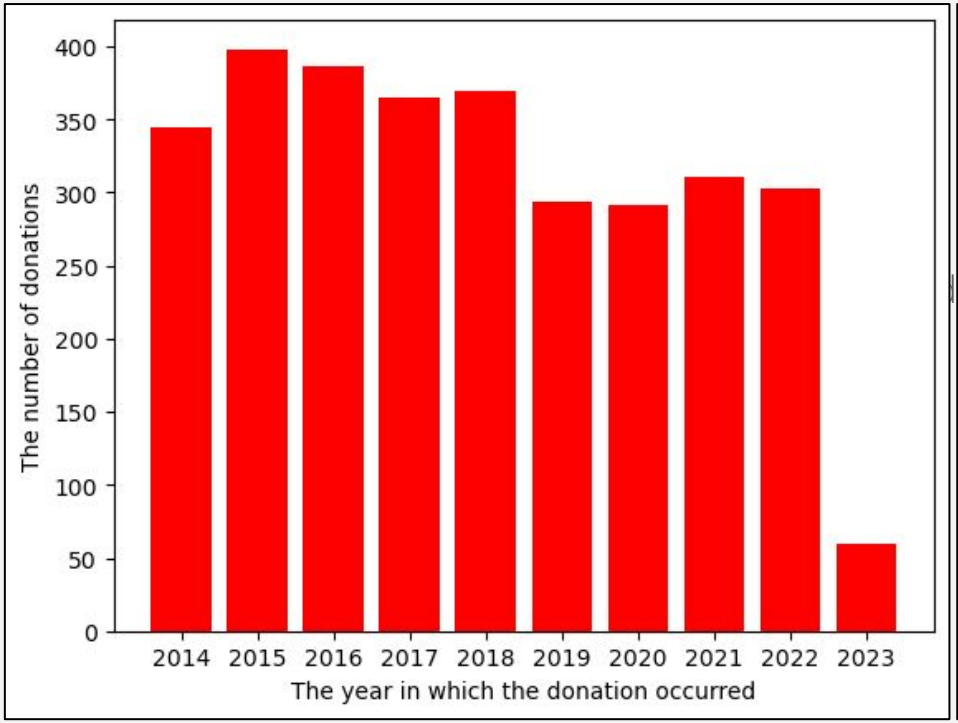
Sociodemographic Features:

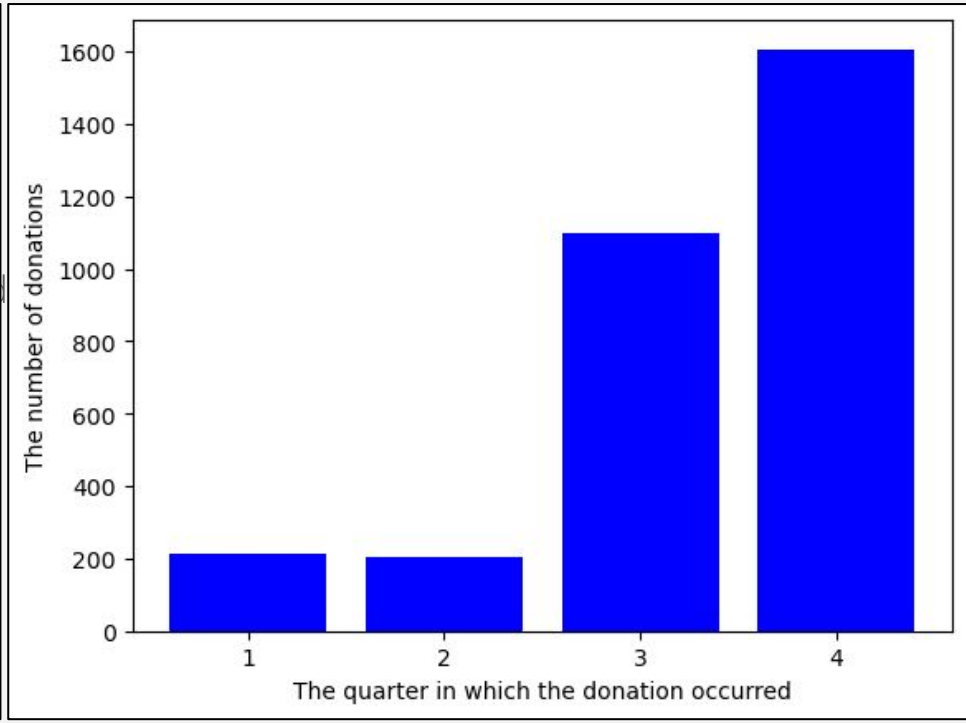
- Income, education, and employment status.
- Binary outcome (1 = likely to donate again, 0 = unlikely).

Feature Engineering:

- donation statistics (number of donations, average donation, max donation).

Visualizations





Modeling & Evaluation

Model Comparison

Model Name	Description	Pros	Cons
Linear Regression	A simple regression model that assumes a linear relationship between features and the target.	Simple and interpretable	Assumes linear relationships, which may limit accuracy for complex data patterns
Random Forest	An ensemble model that combines multiple decision trees to improve prediction accuracy.	Helps capture nonlinear relationships	Can be harder to interpret compared to linear models
Gradient Boosting Machine (GBM)	An iterative ensemble model that builds trees sequentially to minimize errors.	Can capture more intricate relationships	May require more data, time, tuning, and energy

Key Terms

Accuracy: Percentage of correctly predicted donations.

Precision: The proportion of predicted positive cases that are actually positive.

Recall: The proportion of actual positive cases that are correctly predicted.

F1-score: A balance between precision and recall, ensuring both false positives and false negatives are minimized.

Confusion Matrix: A table used to evaluate the performance of a classification model.

Label Creation & Feature Extraction

Label Creation:

- Training Set: Donors who donated before the model_date (2021-01-01).
- Evaluation Set: Donors who donated after the model_date.

Features:

- Donation Statistics: Number of donations, average donation, max donation.
- Top 3 Demographic Features: Based on importance from preliminary model training

Model Training and Evaluation

Model: Random Forest Classifier

- Balanced Dataset: Random oversampling to address class imbalance.
- Evaluation: Stratified K-Fold Cross-Validation (5 folds) with metrics:

  - Accuracy, F1 Score, Precision, Recall, Confusion Matrix.

Model Results

```
Top 3 Demographic Features: ['Poor Family', 'HH Income', 'Minority']
Accuracy: 0.9600997506234414
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.93      0.96       206
           1       0.93      0.99      0.96       195

    accuracy                           0.96       401
   macro avg       0.96      0.96      0.96       401
weighted avg       0.96      0.96      0.96       401

Confusion Matrix:
[[192  14]
 [  2 193]]

Percentage of Donors Predicted to Repeat Again: 51.6209476309227
```

Model Evaluation

```
Top 3 Demographic Features: ['HH Income', 'Poor', '$125-150k']
Accuracy: 0.9812206572769953
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98       213
           1       0.96      1.00      0.98       213

    accuracy                           0.98       426
   macro avg       0.98      0.98      0.98       426
weighted avg       0.98      0.98      0.98       426

F1 Score: 0.9622
Precision: 0.9281
Recall: 0.9991
Confusion Matrix:
[[205   8]
 [  0 213]]

Percentage of Donors Predicted to Repeat Again: 68.29268292682927
```

Final Thoughts

Next Steps and Future Work

- **Model Refinement:**
  - Hyperparameter tuning to improve model accuracy and reduce overfitting.
  - Incorporating additional features such as engagement data (e.g., event attendance, communication with the nonprofit).
- **Alternative Models:**
  - Exploring other algorithms like gradient boost or logistic regression for comparison.
- **Business Impact:**
  - **Targeted Outreach:** Focus on high-potential donors.
  - **Improved Retention:** Engage donors at risk of attrition based on model predictions.

What We Learned

- **Working on a large project:**
  - Importance of understanding the problem
    - What is the input/output?
  - Combining datasets
  - Working through code step-by-step
- **Importance of understanding your data:**
  - What do the outliers indicate
  - How to handle missing values depending on the situation
  - How to best use the data to achieve the project goal
- **Team collaboration:**
  - How to handle conflicting schedules
  - Distributing labor


Questions?