Houston Market Targeting Release Notes

Anas Junedahmed Puthawala

Overview

- Developed a predictive model to identify Houston-area residents most likely to enroll in Texas Health’s HSA / FSA programs. The model scores and ranks 2,500 prospects, enabling the outreach team to prioritize efficiency in outreach.

How to use the score?

- Each prospect in the Houston file has an enrollment probability score between 0 and 1.
  - Numbers closer to 1 Being that they are highly likely to enroll in **EITHER** of the programs
  - Numbers closer to 0 Being that they are unlikely to enroll in **EITHER** of the programs
- Important Note: This model does not dictate which of the two plans the participant has a propensity for, additional care and modeling should be taken to prioritize and differentiate between the two plans.
  - A secondary model can be utilized to differentiate between the two programs.

Model Validation & Metrics

- Validated the model using 5-fold cross validation on historical enrollment records from across Texas.
  - Split the dataset in various different splits, trained the model(s), examined the metrics, rinse and repeat multiple times to ensure consistency.
- I experimented with a wide set of models, from tree-based to statistical linear models. Found that Logistic Regression was ideal for our use-case due to interpretability, and performance.

Metric	Value
AUC	0.74
Recall	66.0%
F2 Score	0.66
Precision	66.7%

- Recall
  - Of all the people who would actually enroll, what percentage did we correctly identify?
- Precision
  - Of the people that the model predicted would enroll, what percentage actually enrolled?
- F2 Score
  - Variant of F1 score, putting more emphasis on Recall
- AUC
  - Measures how well the model ranks people, do likely enrollers end up at the top of the list?

Model Impact

- Recall (66%) - Model successfully captured 66% of people who were going to enroll. If there were 100 actual enrollers hidden in the data, the model successfully identified 66%.

- Precision (66.7%) - How efficient the model can be. If we are doing outreach to 100 people that the model flag, on average around 67 of those should be actual enrollers, and 33 will not.

- F1 and F2 scores are means to optimize and get a holistic view of combining metrics such as Precision and Recall, and in the case of F2 giving more weight towards Recall, as is important in our use-case.

Metric	Value
AUC	0.74
Recall	66.0%
F2 Score	0.66
Precision	66.7%

Model Impact

- Gain @ X% - Cumulative percentage of all possible outcomes. As ‘X’ increases, Gain will increase too (if you look at 100% data, your Gain should be 100%).

- Lift @ X% - The ratio of how much better the model performance is than random. As ‘X’ increases, Lift usually decreases because the model’s ability to filter the best cases become diluted as you include lower-probability records.

% Of List Contacted	% Of Enrollers Captured	Lift vs Random
Top 10% model scores	17.7%	1.77x
Top 20% model scores	32.9%	1.64x
Top 30% model scores	45.9%	1.53x

Model Impact

- By only contacting the top 30% of the list as per the model predictions, we can successfully capture nearly half of all potential enrollers.
  - Ignore 70% of potential noise, more honed in and focused outreach effort for top predicted enrollees

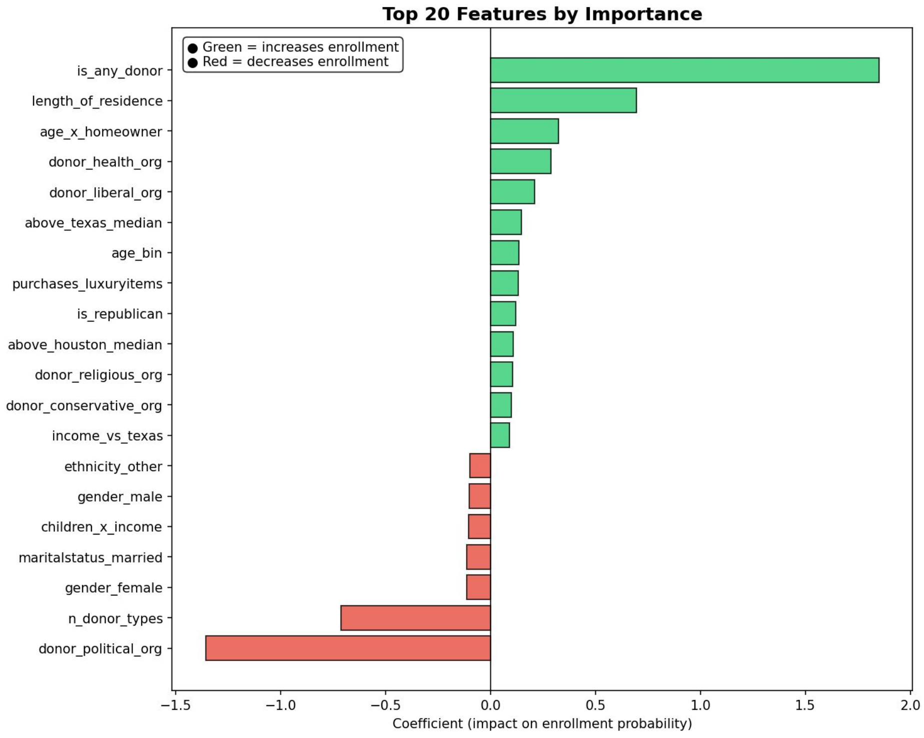
- The model's prioritization is 77% more effective than random sampling in the top 10% of outreach alone.

% Of List Contacted	% Of Enrollers Captured	Lift vs Random
Top 10% model scores	17.7%	1.77x
Top 20% model scores	32.9%	1.64x
Top 30% model scores	45.9%	1.53x

Model Interpretability

- Features such as length of residence, age, if they are a donor tend to drive the individual to be marked *more likely* to be a enrollee by the model.

- Features such as the gender (female vs male), marital status, and the political organization tend to drive the individual to be marked *less likely* to be an enrollee by the model

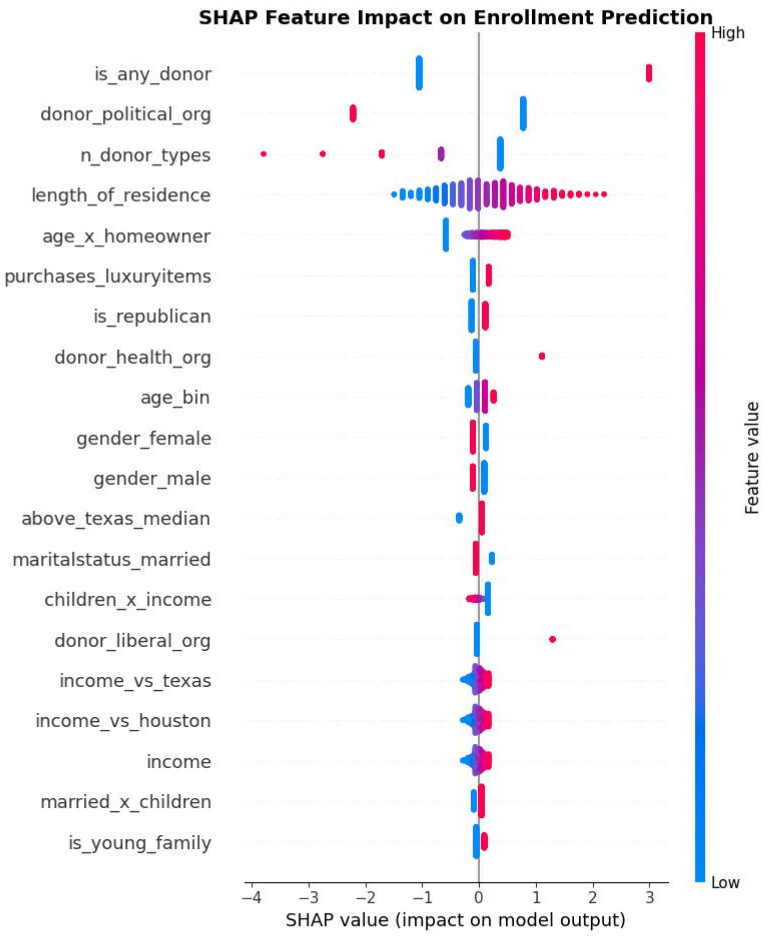


Model Interpretability

- While the coefficients don't really tell us much about the actual *values* themselves that are causing the prediction to be driven a certain way, the SHAP values do.

- For example, as the age increases, it has a slight positive impact on the model to predict that they will be an enrollee

- Those that make above the median salary in Texas (feature; `above_texas_median`) have a somewhat higher impact to drive the model to predict that they will be an enrollee



Assumptions and limitations

1. Houston behavior mirrors rest of Texas

2. Outreach capacity is not constrained– I optimized for Recall in this case. Assuming the cost of extra contact is low, relative to missing a potential enrollee

3. No prior Houston enrollment data – We cannot validate the immediate performance. We'd need to set up some infrastructure such that we receive feedback from outreach crew on who ended up enrolling and who didn't. We can utilize this to fine-tune the model and re-train.

4. If the underlying distribution of the data begins to drift drastically, we could re-train the model. This is a time-in-point snapshot for which the model was trained on. We can use something like Mlflow or Arize to integrate model monitoring and have drift detection :)

Appendix

Thanks for taking the time to read through the release notes and my code submission.

If I had some more time, i’d integrate model monitoring and versioning using something like Arize / MLflow. This enables us to make adjustments and upload new versions of the model which we can swap out and use in production with ease.