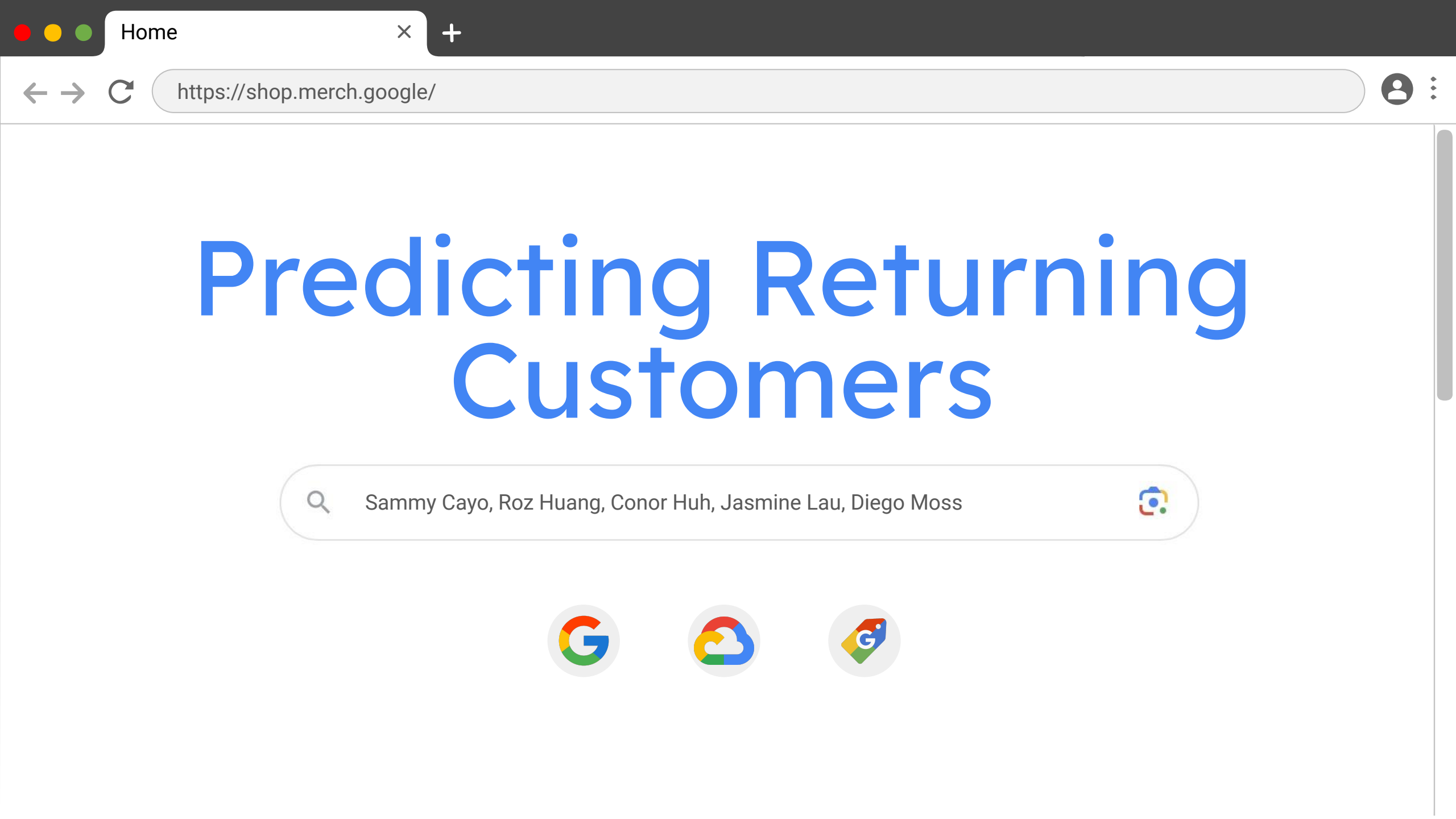
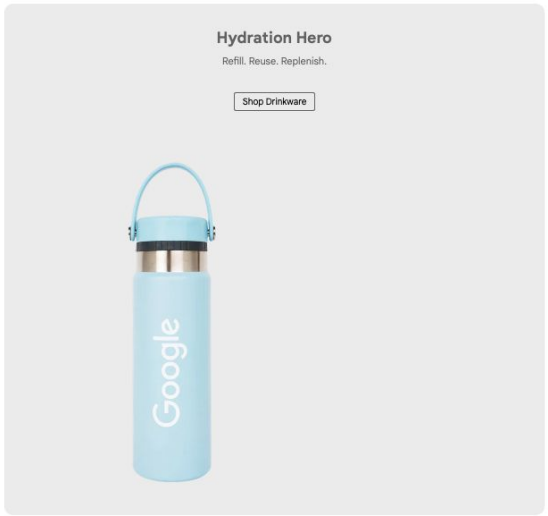

Home
https://shop.merch.google/
Predicting Returning Customers
Sammy Cayo, Roz Huang, Conor Huh, Jasmine Lau, Diego Moss

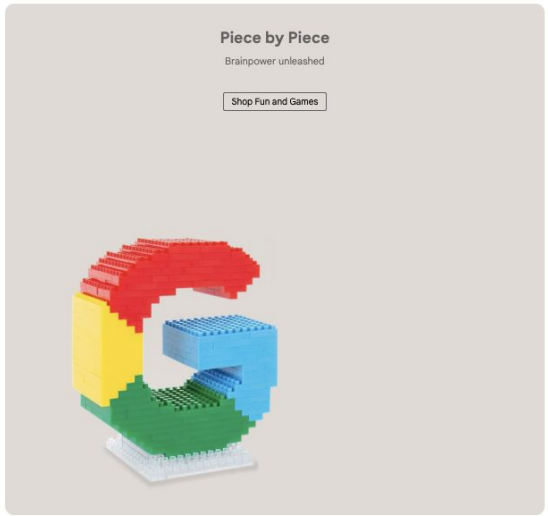
Home | Merch | Motivation

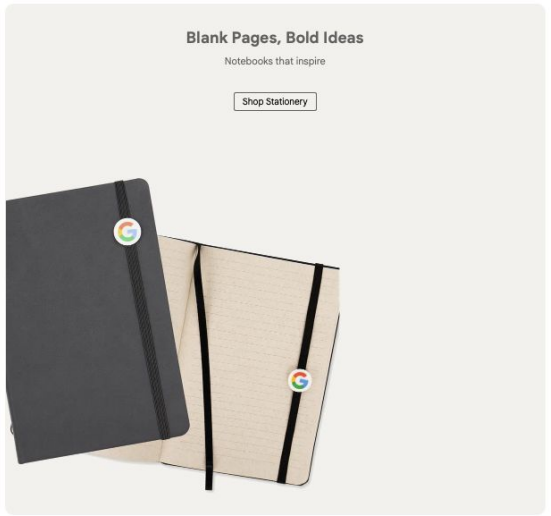
https://shop.merch.google/

Motivation

Popular on the Google Merch Shop







Goal:

- User Retention Prediction

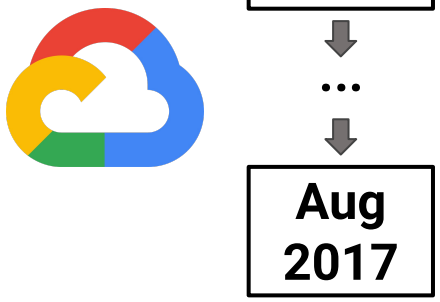
Impact:

- Higher conversion rate of ad spend

Challenge:

- Implement ML to predict customer return to Google's online store

BigQuery



Data: Aug 1, 2016 - Aug 1 ,2017

- 366 files
- ~35 GB of data

Content:

- Traffic Source Data
- Content Data
- Transnational Data

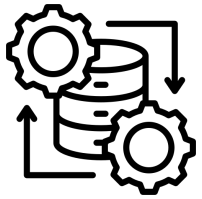
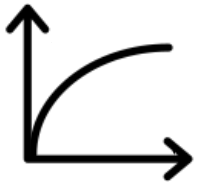
Difficulties:

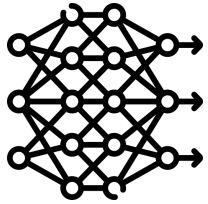
- Kaggle → BigQuery client
- Outdated documentation
- SQL → Python
- API call + query + download = 6+ hrs.

Home | Big Query | Project Outline

Project Outline

Data Querying → Data Processing → Baseline Model → Advanced Model → Output



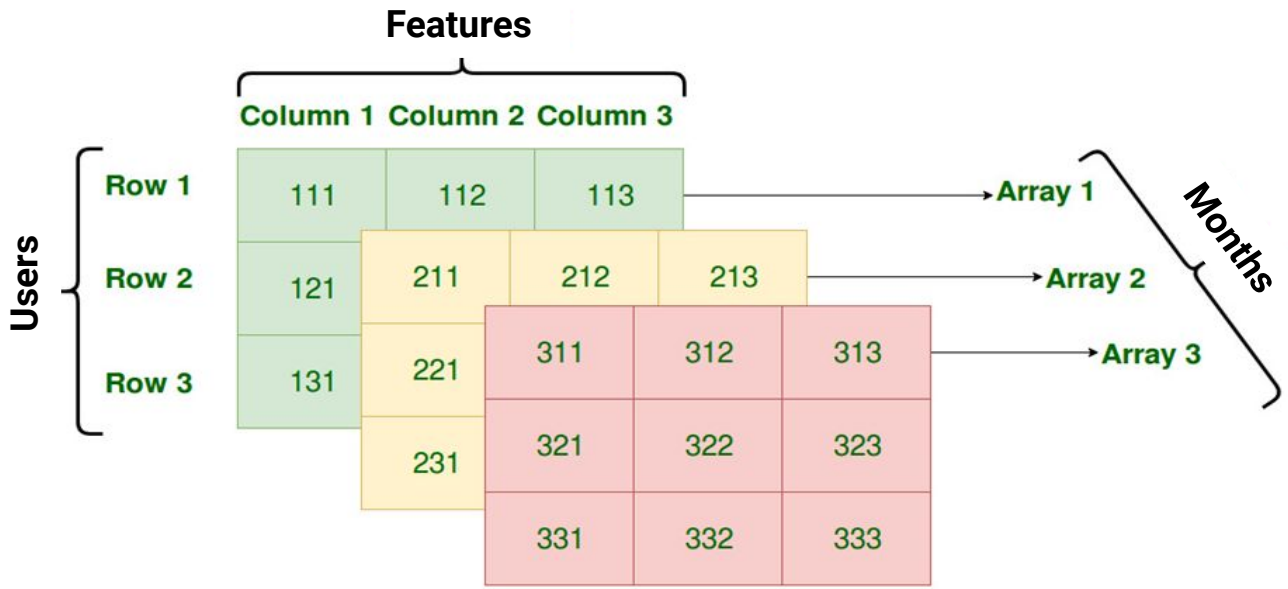
Data

Data Set Overview:

- 718,161 unique users
- 566,477 unique users with session data
- Data from 13 months (366 days)

Missing Values:

- Data was complete except for pieces of session data
  - ie. If there were no add-to-carts during session
- Filled missing data with 0's



Feature Engineering

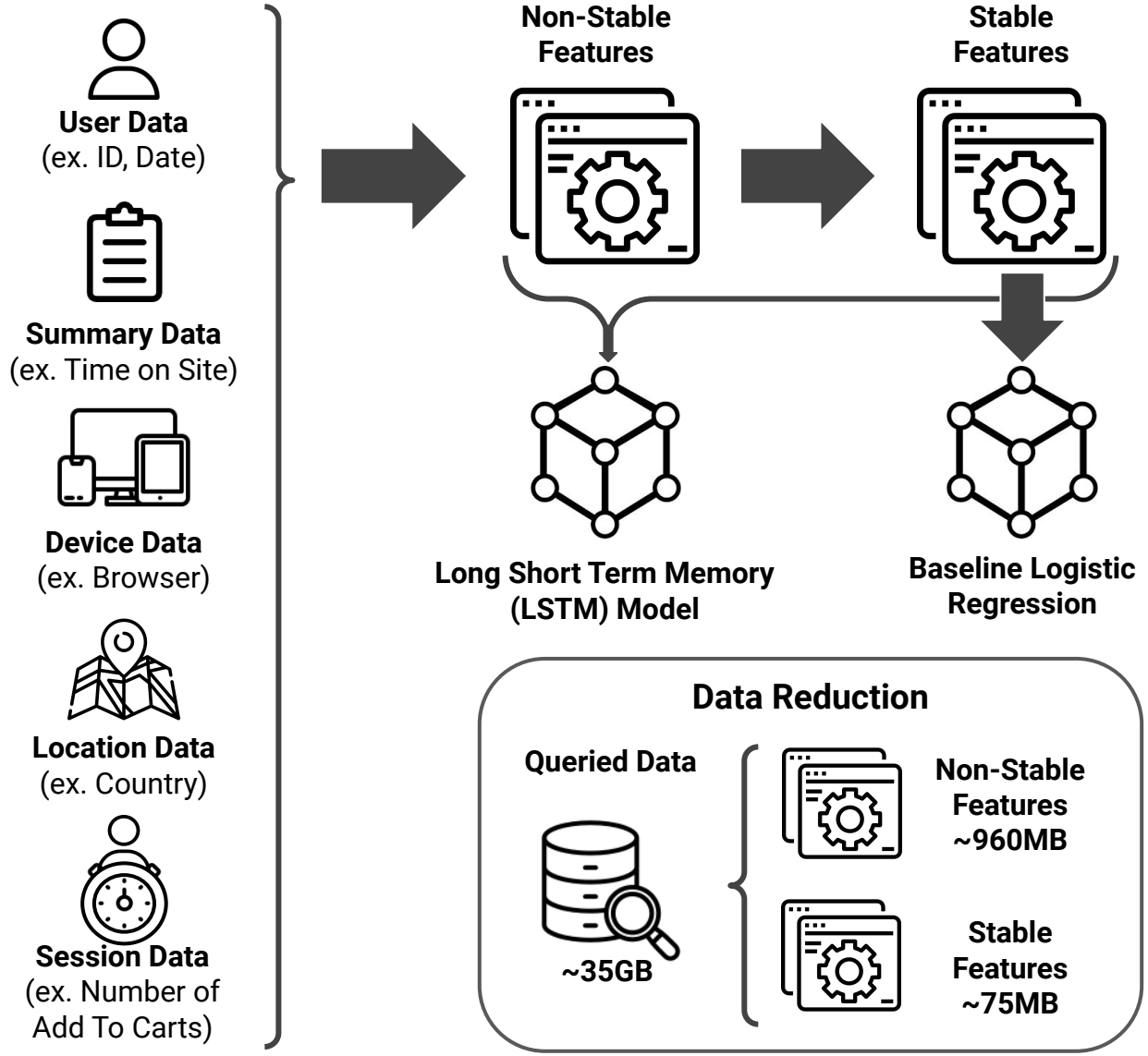
Two Sets of Features:

Non-Stable Features:

- For each user, their respective features were organized by month.
- Non-numeric features such as a users' location were **label encoded** using **sklearn**

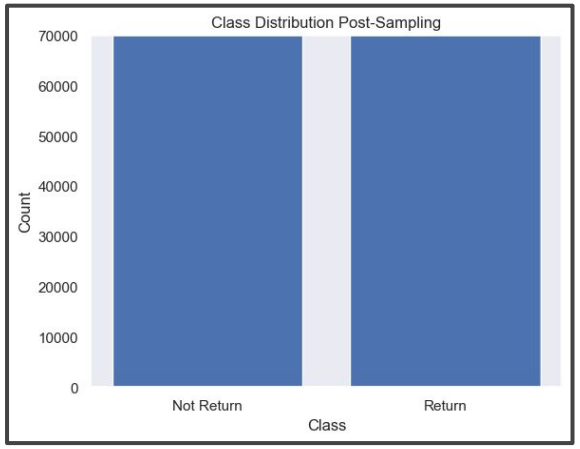
Stable Features:

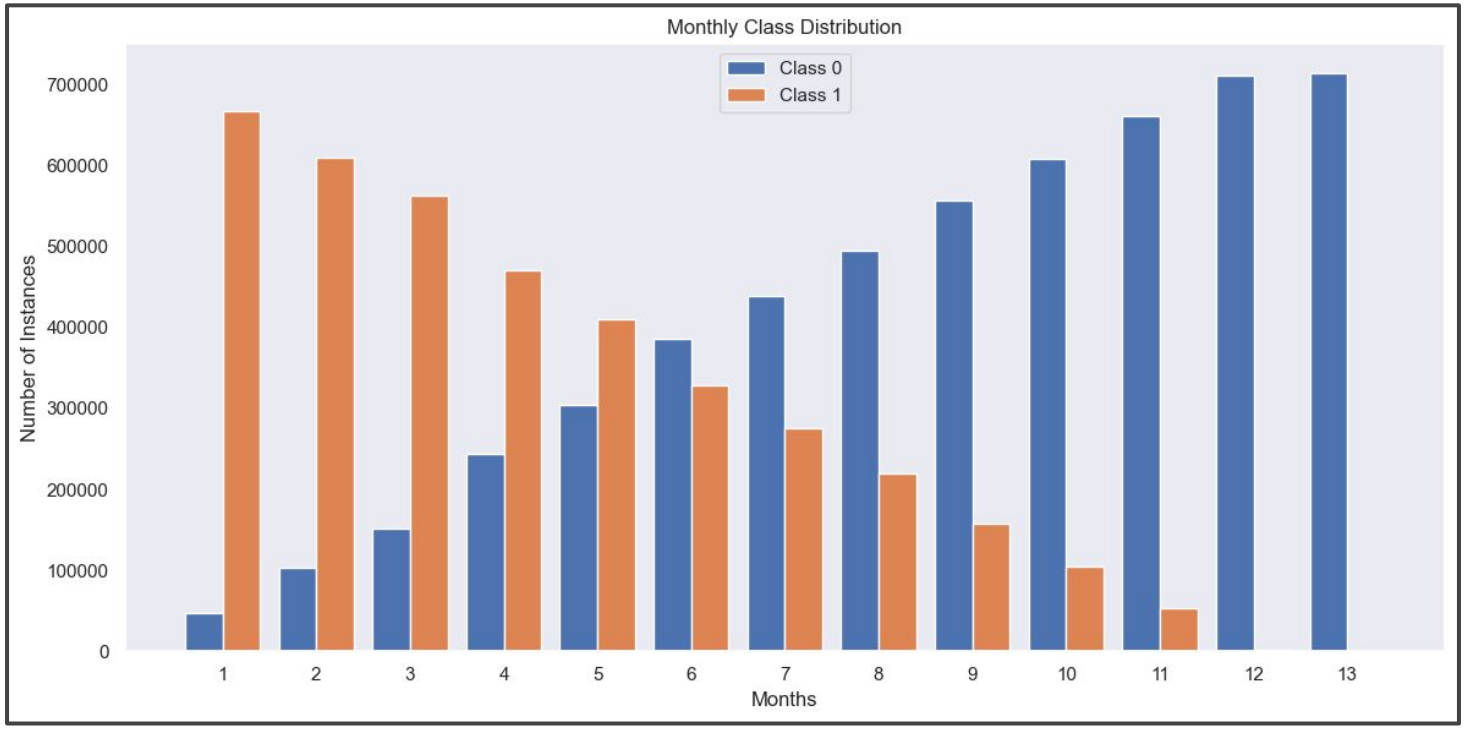
- Non Stable Features for each user were averaged across all 13 months.



** Baseline Logit model used only stable features; full description of features used in appendix*

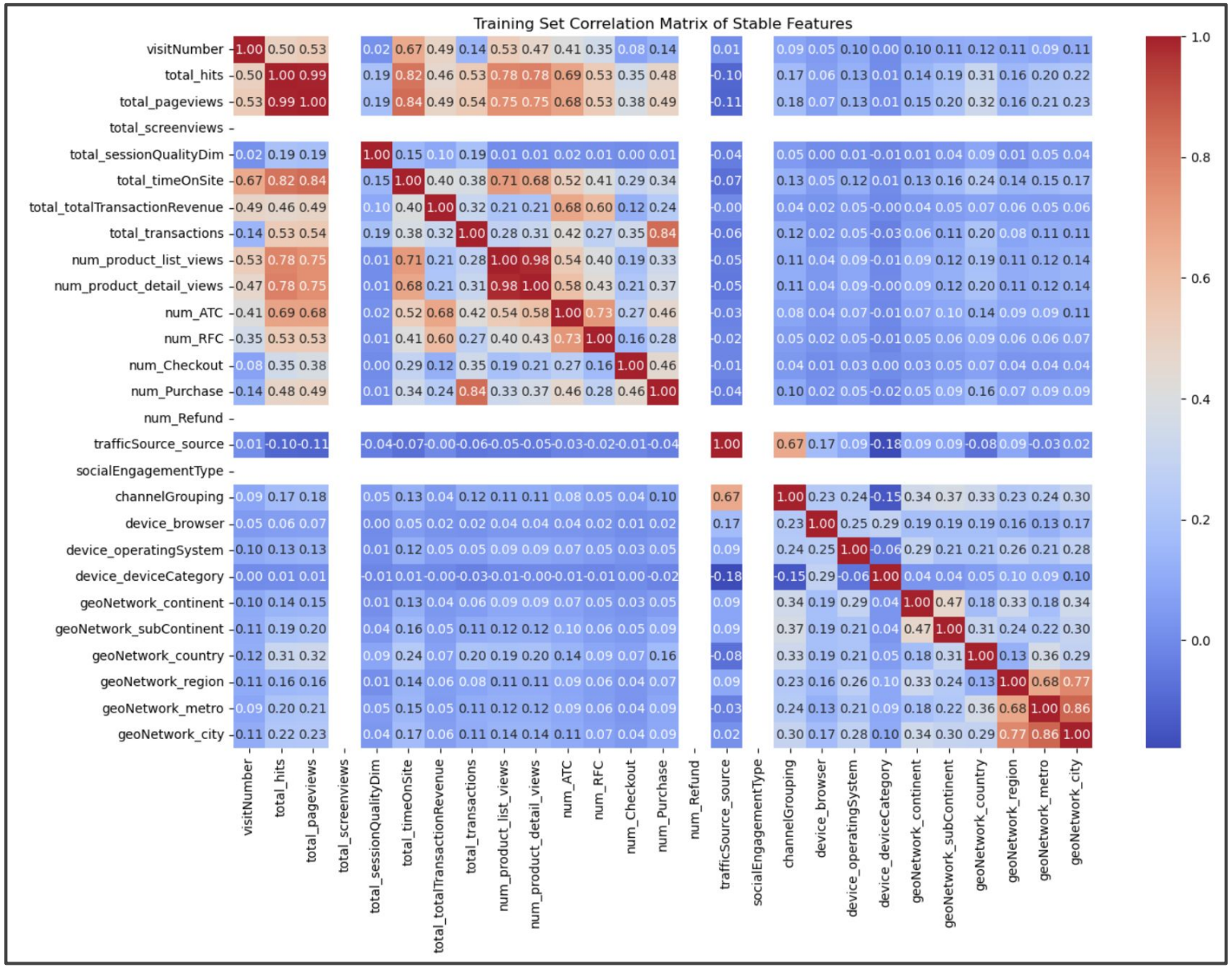
Data Imbalance



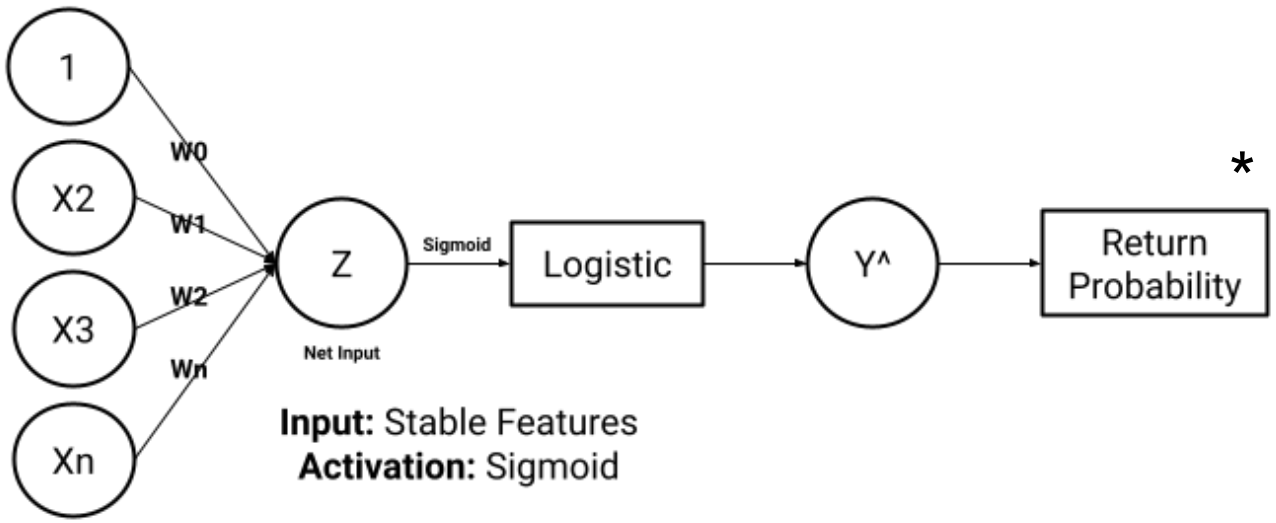


Home | Data Imbalance | EDA

EDA



Baseline Model



Input: Stable Features
Activation: Sigmoid

Layer	Output Shape	Param #
dense (Dense)	(None, 1)	28

Total Parameters: 86

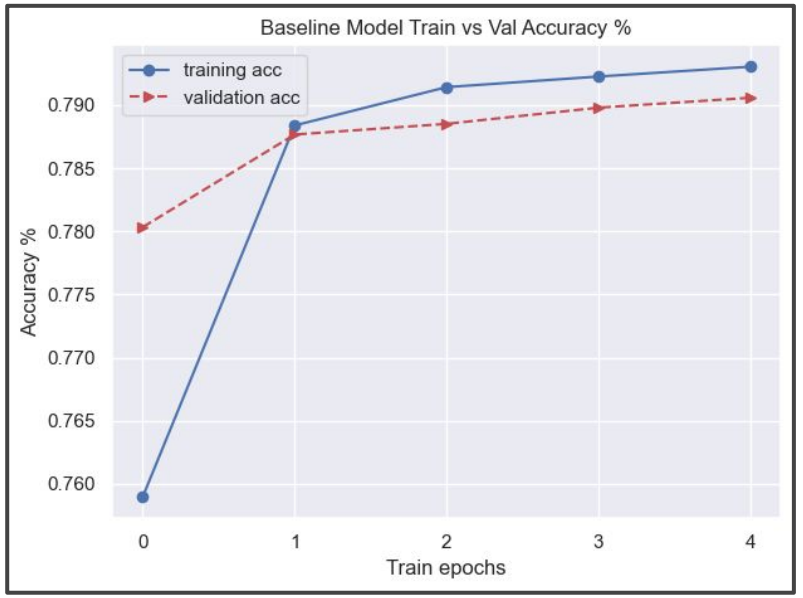
Trainable Parameters: 28

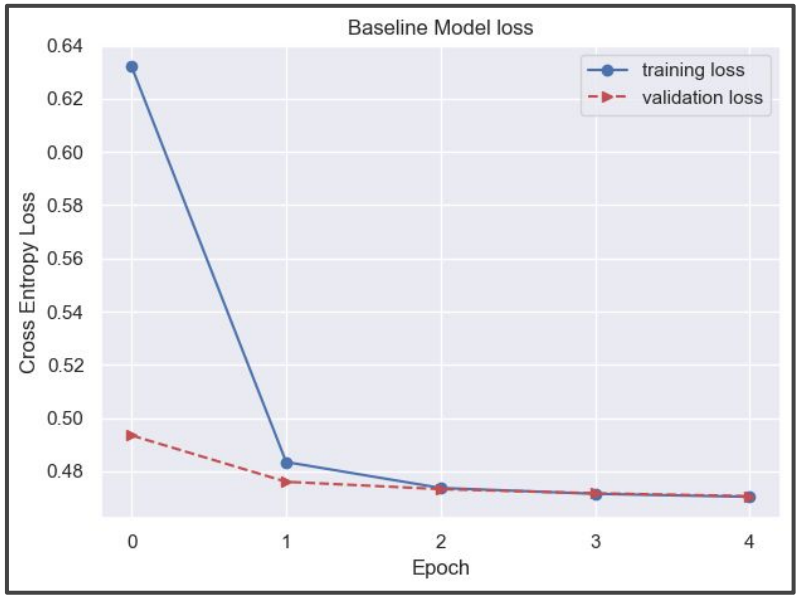
Non-Trainable Parameters: 0

Optimizer Parameters: 58

* ML for Biologists

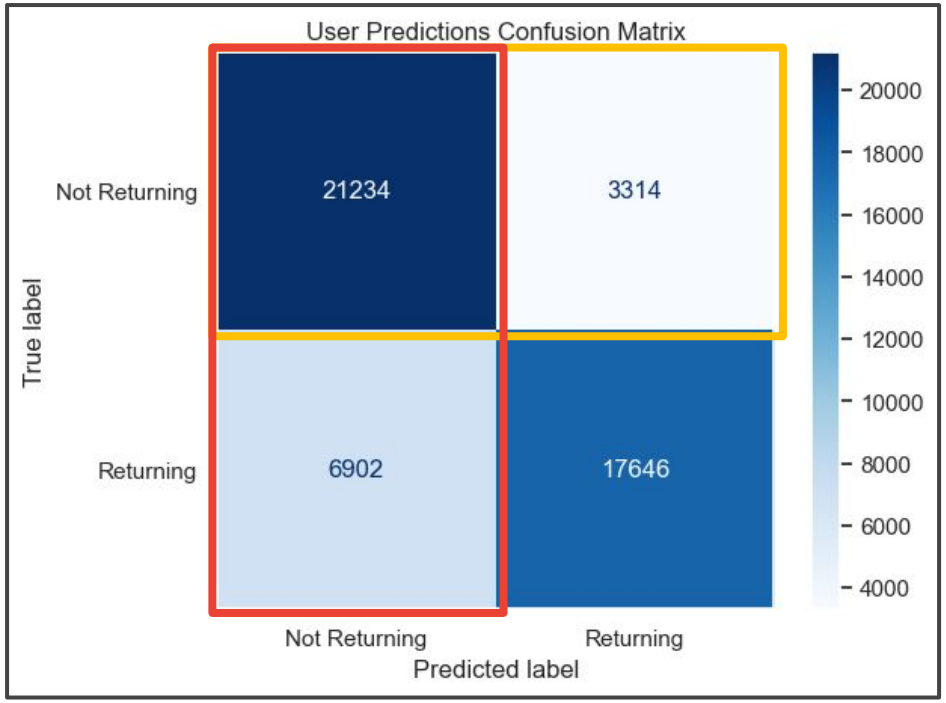
Baseline Model Performance

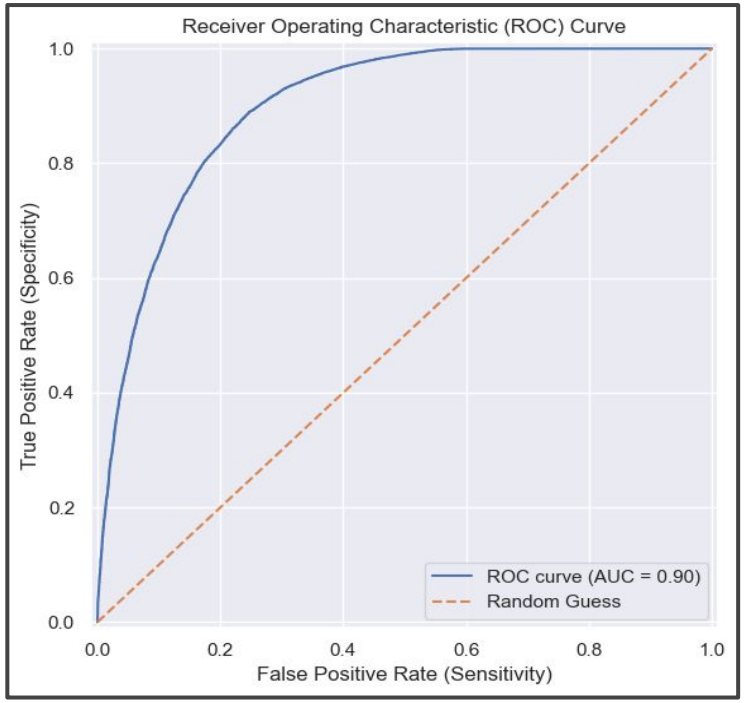




Model	Loss	Accuracy	Precision	Recall
Baseline	0.470315	0.793532	0.83964	0.725654

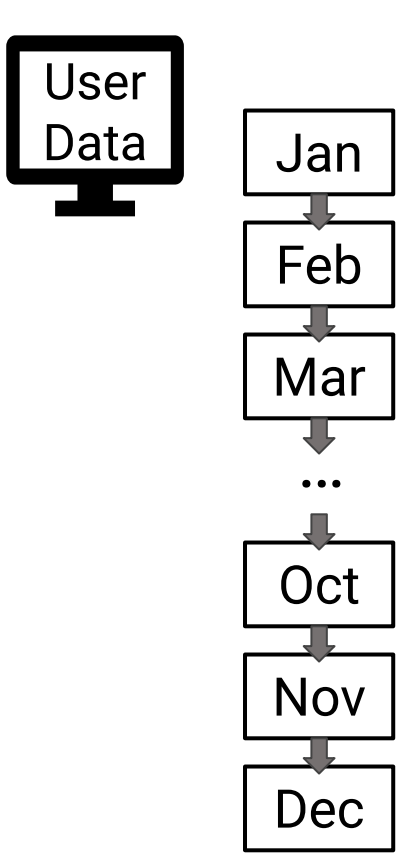
Baseline Model Performance

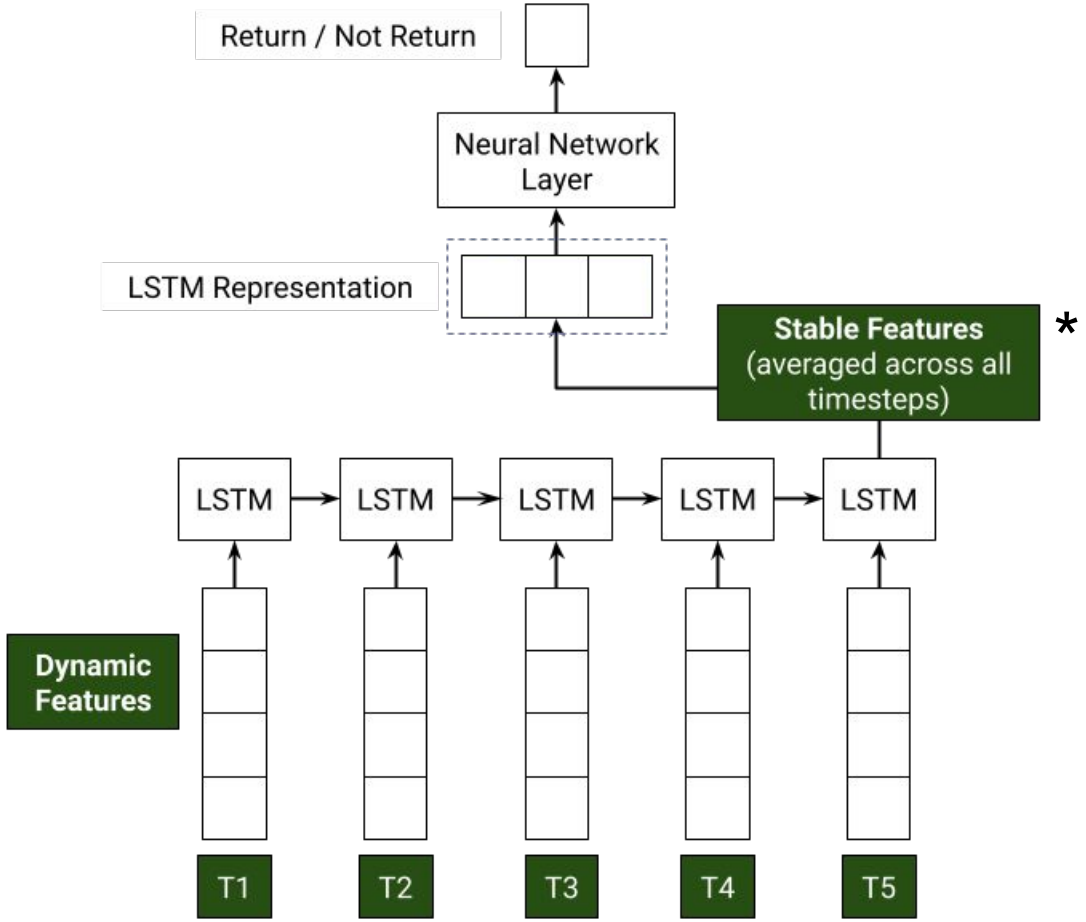




Recall: TP/ (TP + FN)
Precision: TP/ (TP + FP)

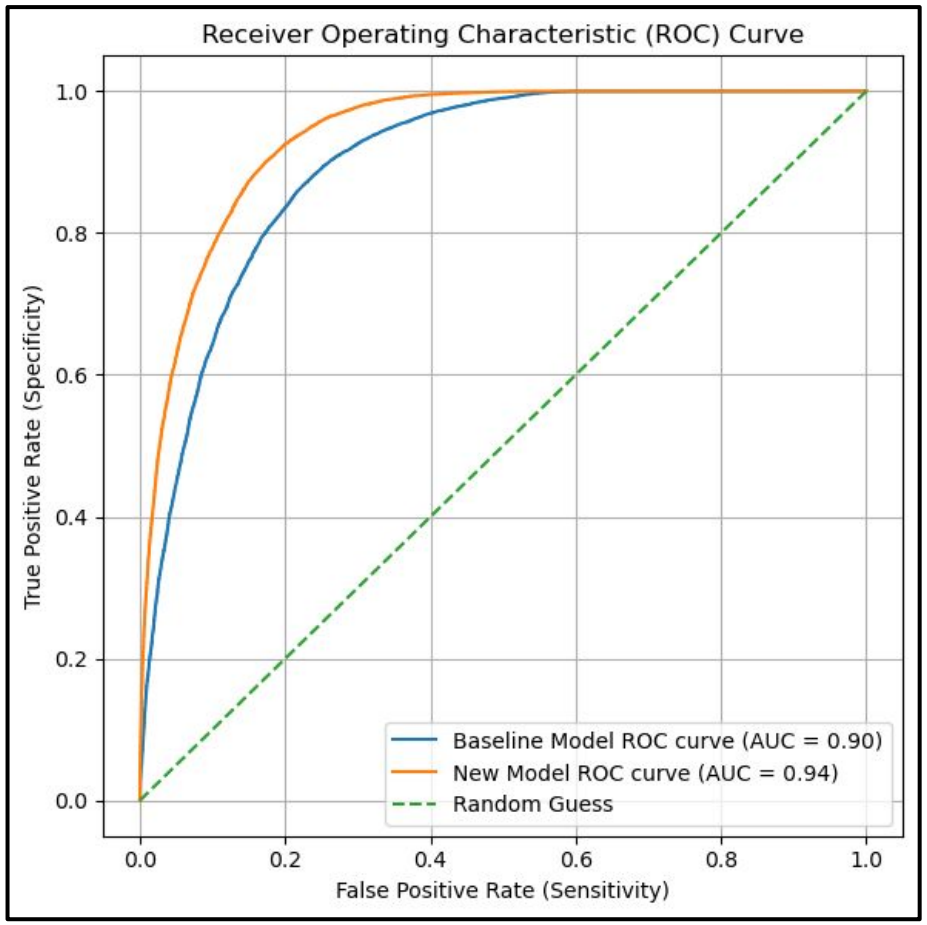
LSTM Model Architecture





* Wilcox, K., & Wang, L. (2022)

LSTM ROC Model Performance



AUC from .90 - .94

Recall from .73 - .93

Model Results and Hyperparameter Tuning

drop out rate	dense neurons	lr	train loss	train recall	val loss	val recall
0.2	None	0.001	0.270952	0.913334	0.268425	0.915902
0.2	None	0.0001	0.386963	0.914186	0.385042	0.915159
0.2	None	0.001	0.239962	0.913222	0.241114	0.916109
0.2	None	0.0001	0.381681	0.904108	0.380113	0.904007
0.2	50	0.001	0.30718	0.939715	0.307181	0.935233
0.2	50	0.0001	0.312531	0.930838	0.31023	0.932425
0.2	**50**	**0.001**	**0.266768**	**0.957386**	**0.270468**	**0.951425**
0.2	50	0.0001	0.306068	0.922156	0.305149	0.922181
0.5	None	0.001	0.460965	0.934927	0.463083	0.928872
0.5	None	0.0001	0.381211	0.895817	0.378881	0.89558
0.5	None	0.001	0.299457	0.928521	0.297285	0.930979
0.5	None	0.0001	0.379716	0.883869	0.377461	0.884015
0.5	50	0.001	0.298759	0.939338	0.298797	0.937794
0.5	50	0.0001	0.341267	0.914255	0.338925	0.916894

Final Model Test Recall:

93.72%

Final Model Parameter Count:

50,255 params (~197 KB)

Prediction Time (Inference):

~2 ms for one prediction for a user

Conclusions

Conclusion:

- Identify high-potential customers
  - Focus marketing efforts
  - ↑ User engagement + ↑ User loyalty

Application:

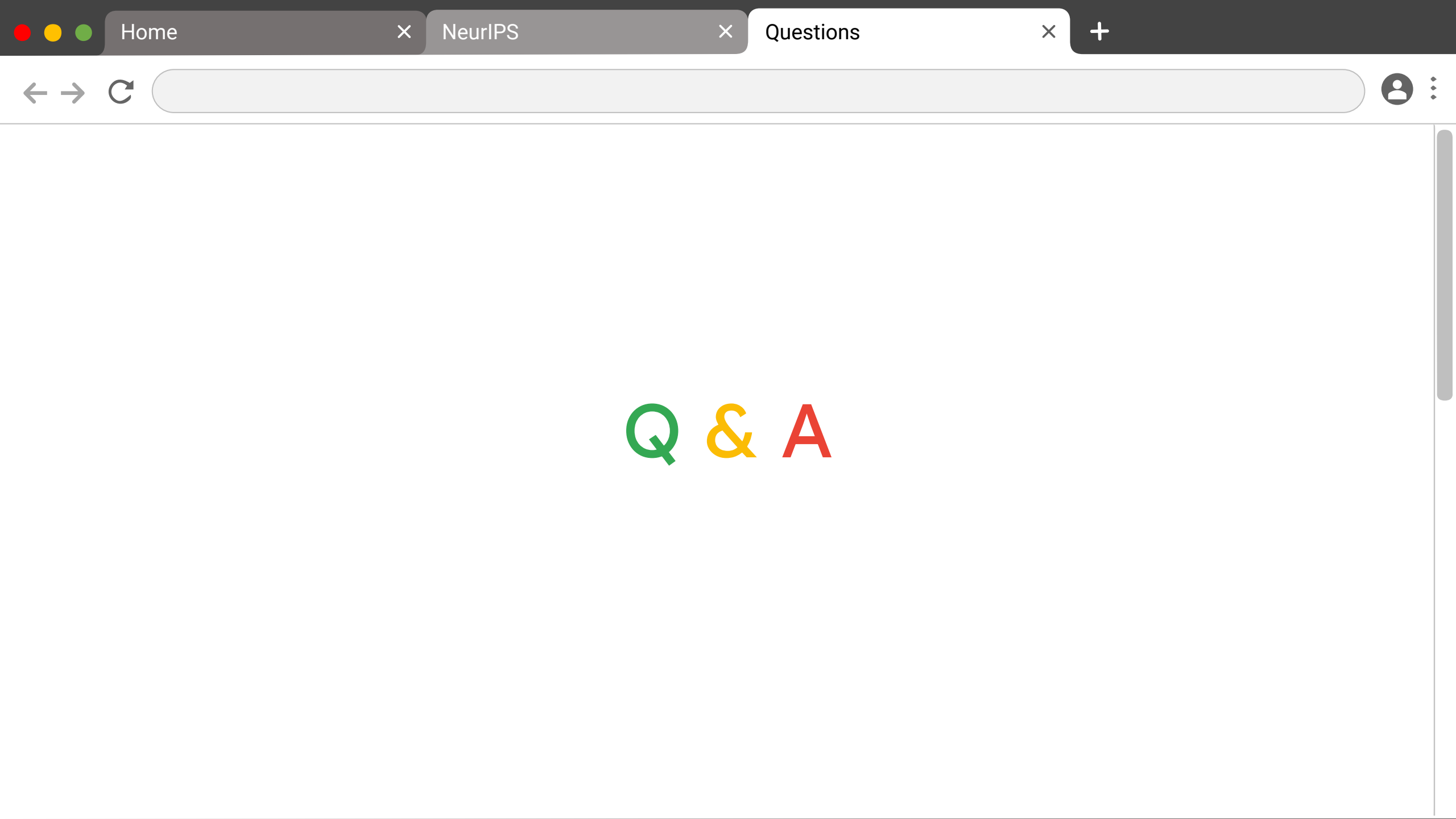
- Managers run model periodically to:
  - Monitor trends
  - Optimize resources
  - Evaluate performance

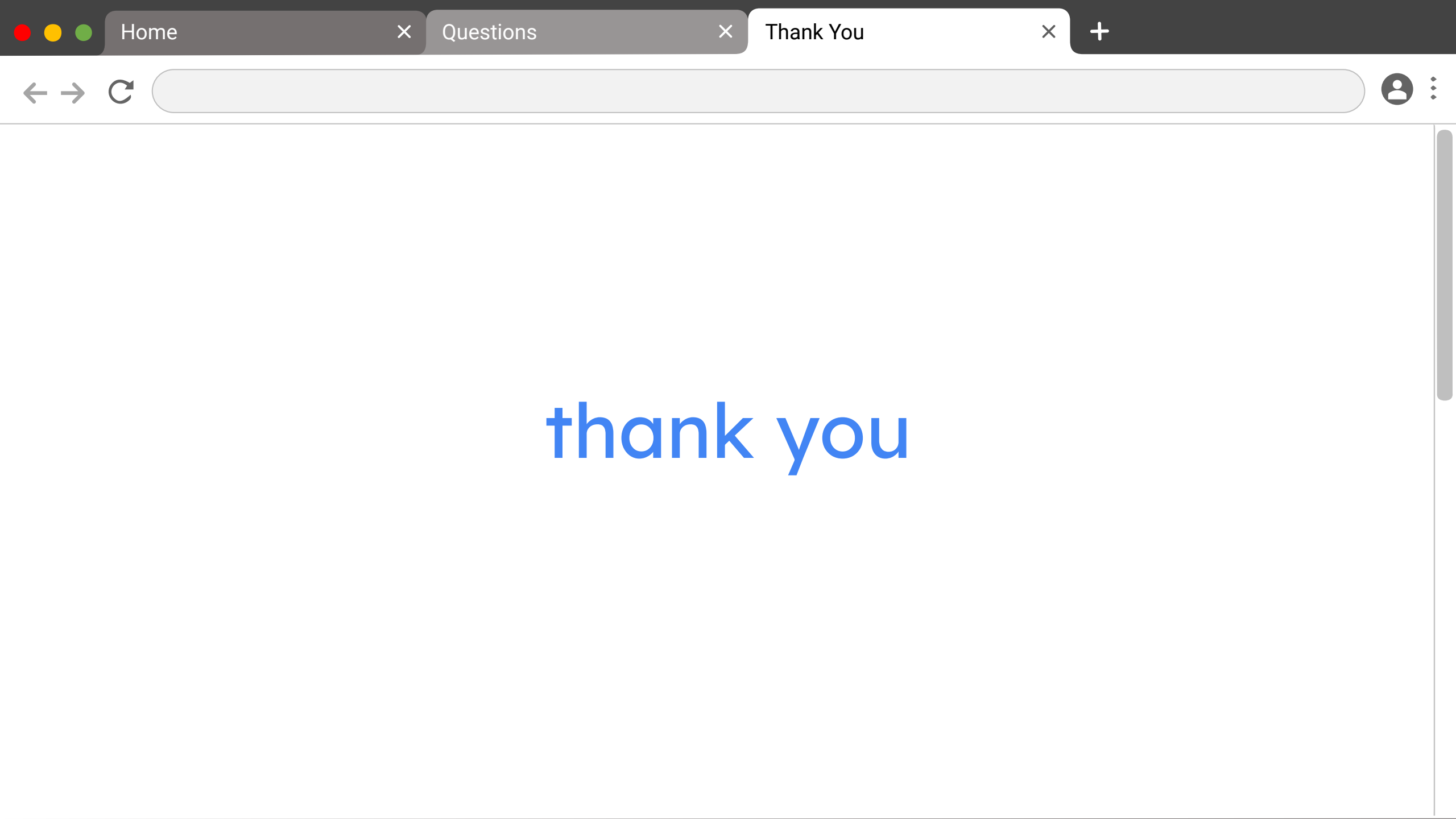
Improvement Avenues:

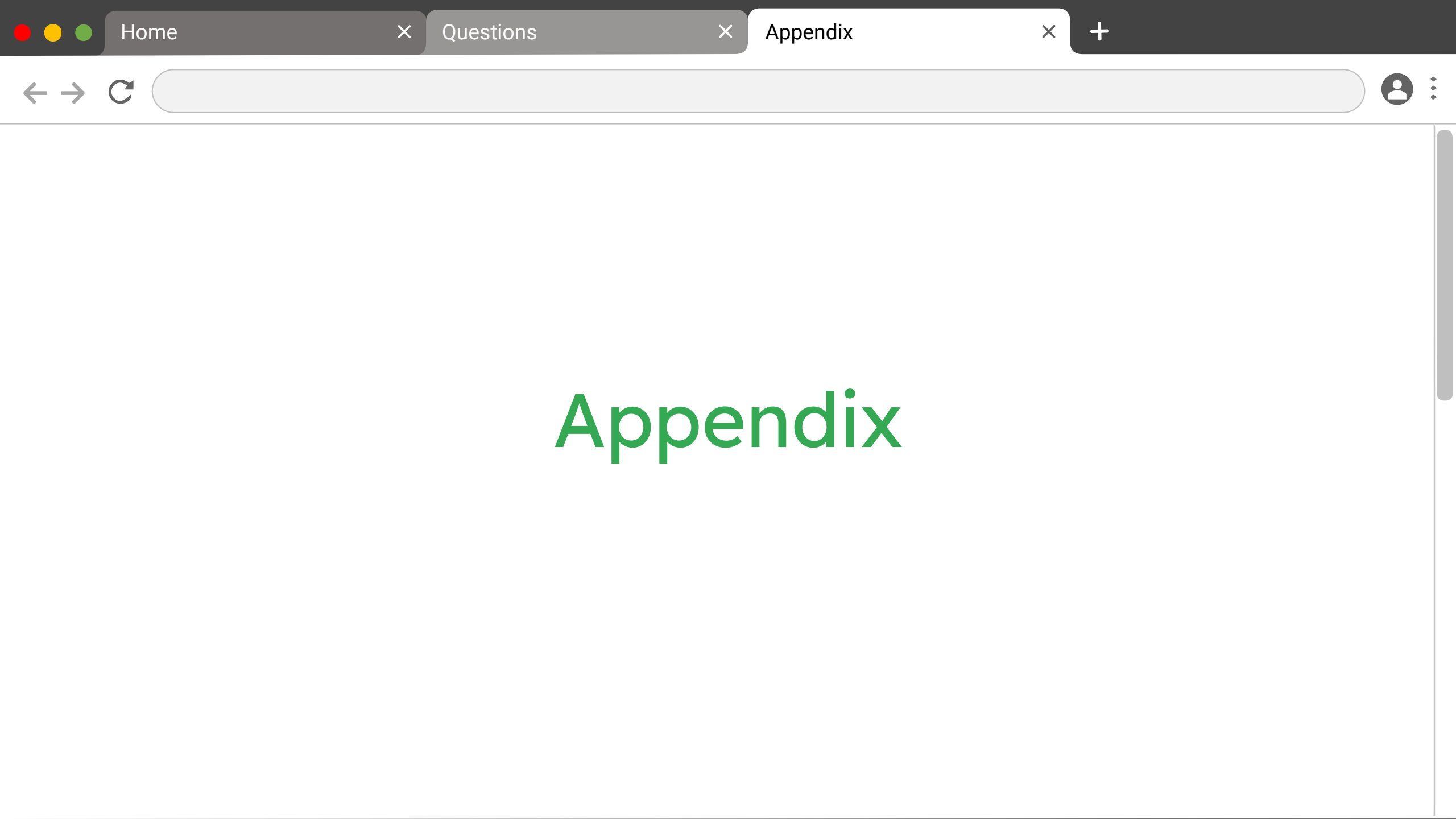
- Play with hyperparameter sets
- ↑ Feature engineering
- Method of filling NaN's
- Compute metrics
  - ie. Session quality
- Work with stable features
- Improve ML fairness
  - ie. Geographical location

NeurIPS Checklist

1. Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? YES
2. Have you read the ethics review guidelines and ensured that your paper conforms to them? YES
3. Did you discuss any potential negative societal impacts of your work?
   a. **Our project has no direct paths to negative outcomes when used as intended**
4. Did you describe the limitations of your work? YES
5. If you are including theoretical results...? NA
6. Did you include the code, data, and instructions needed to reproduce results? YES
7. Did you specify all the training details? YES
8. Did you report error bars? NA
9. Did you include the amount of compute and the type of resources used? NO
10. If your work uses existing assets, did you cite the creators? YES
11. Did you mention the license of the assets? YES
12. Did you include any new assets either in the supplemental material or as a URL? NA
13. Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
14. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
15. If you used crowdsourcing or conducted research with human subjects...? NA


Home
NeurIPS
Questions
Q & A


Home
Questions
Thank You
thank you


Home
Questions
Appendix
Appendix

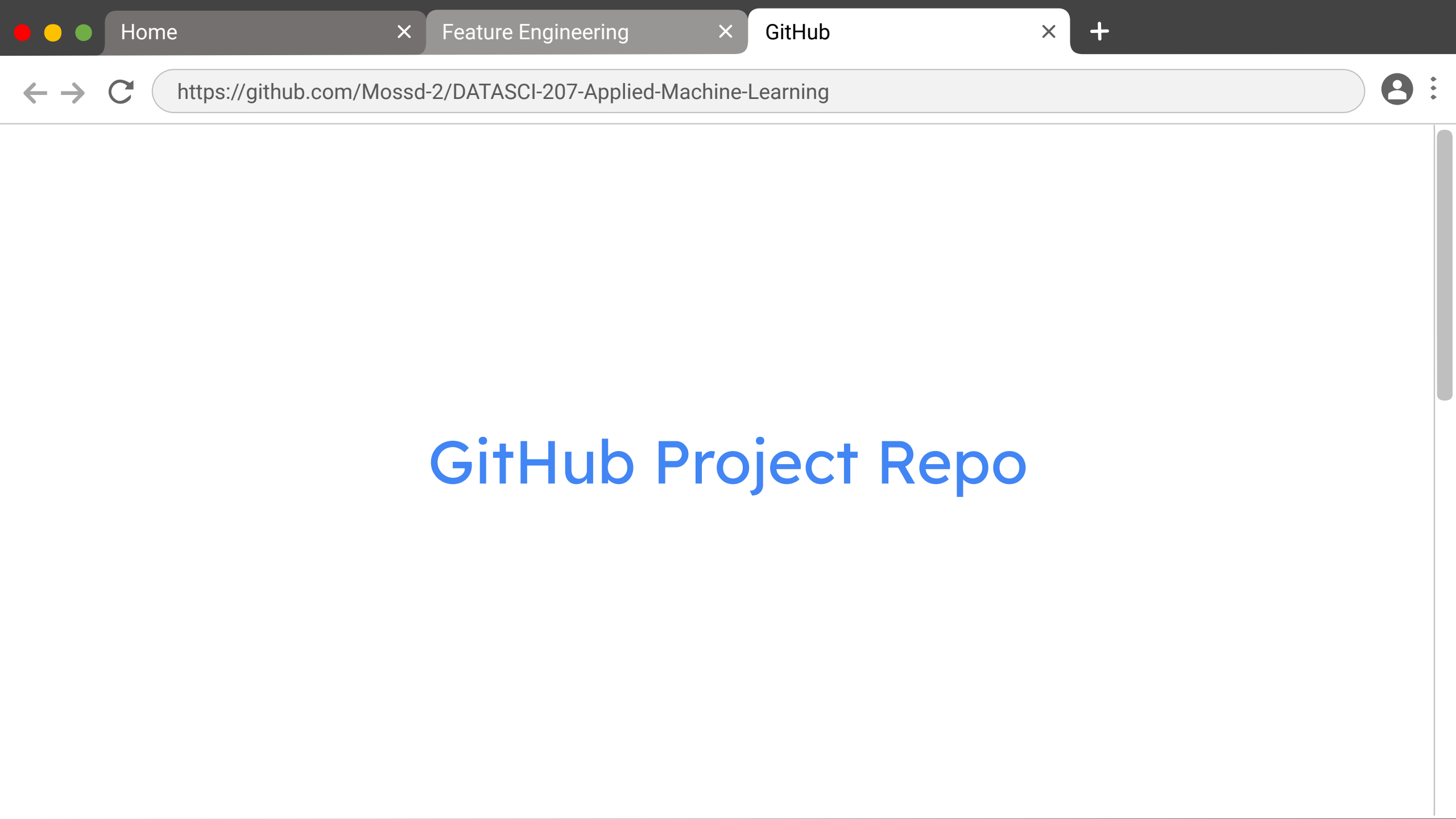
Contributions

	Jasmine Lau	Diego Moss	Roz Huang	Conor Huh	Sammy Cayo
Data Querying	X			X	
EDA			X		
Data Cleaning / Splitting		X	X	X	
Feature Engineering				X	
Modeling		X			X
Presentation Slides	X	X	X	X	X

Feature Engineering (Details)

Feature	Description
fullVisitorId	The unique visitor ID.
visitNumber	The session number for this user. If this is the first session, then this is set to 1.
date	The date of the session in YYYYMMDD format.
total_hits	Total number of hits within the session.
total_pageviews	Total number of pageviews within the session. (desktop only field)
total_screenviews	Total number of screenviews within the session. (mobile only field)
total_sessionQualityDim	An estimate of how close a particular session was to transacting, ranging from 1 to 100.
total_timeOnSite	Total time of the session expressed in seconds.
total_totalTransactionRevenue	Total transaction revenue
total_transactions	Total number of ecommerce transactions within the session.
trafficSource_source	Traffic Source from which the session originated.
socialEngagementType	Engagement type, either "Socially Engaged" or "Not Socially Engaged".
channelGrouping	The Default Channel Group associated with an end user's session for this View.
device_browser	The browser used (e.g., "Chrome" or "Firefox").

Feature	Description
device_operatingSystem	Device
device_deviceCategory	The type of device (Mobile, Tablet, Desktop).
geoNetwork_continent	The continent from which sessions originated, based on IP address.
geoNetwork_subContinent	The sub-continent from which sessions originated, based on IP address of the visitor.
geoNetwork_country	The country from which sessions originated, based on IP address.
geoNetwork_region	The region from which sessions originate, derived from IP addresses.
geoNetwork_metro	The Designated Market Area (DMA) from which sessions originate.
geoNetwork_city	Users' city, derived from their IP addresses or Geographical IDs.
num_product_list_views	Number of times a user views a product list
num_product_detail_views	Number of times a user views a product detail page
num_ATC	Number of Add To Carts
num_RFC	Number of Removes from Cart
num_checkout	Number of Checkouts
num_purchase	Number of Purchases [0,1]
num_refunds	Number of Refunds [0,1]

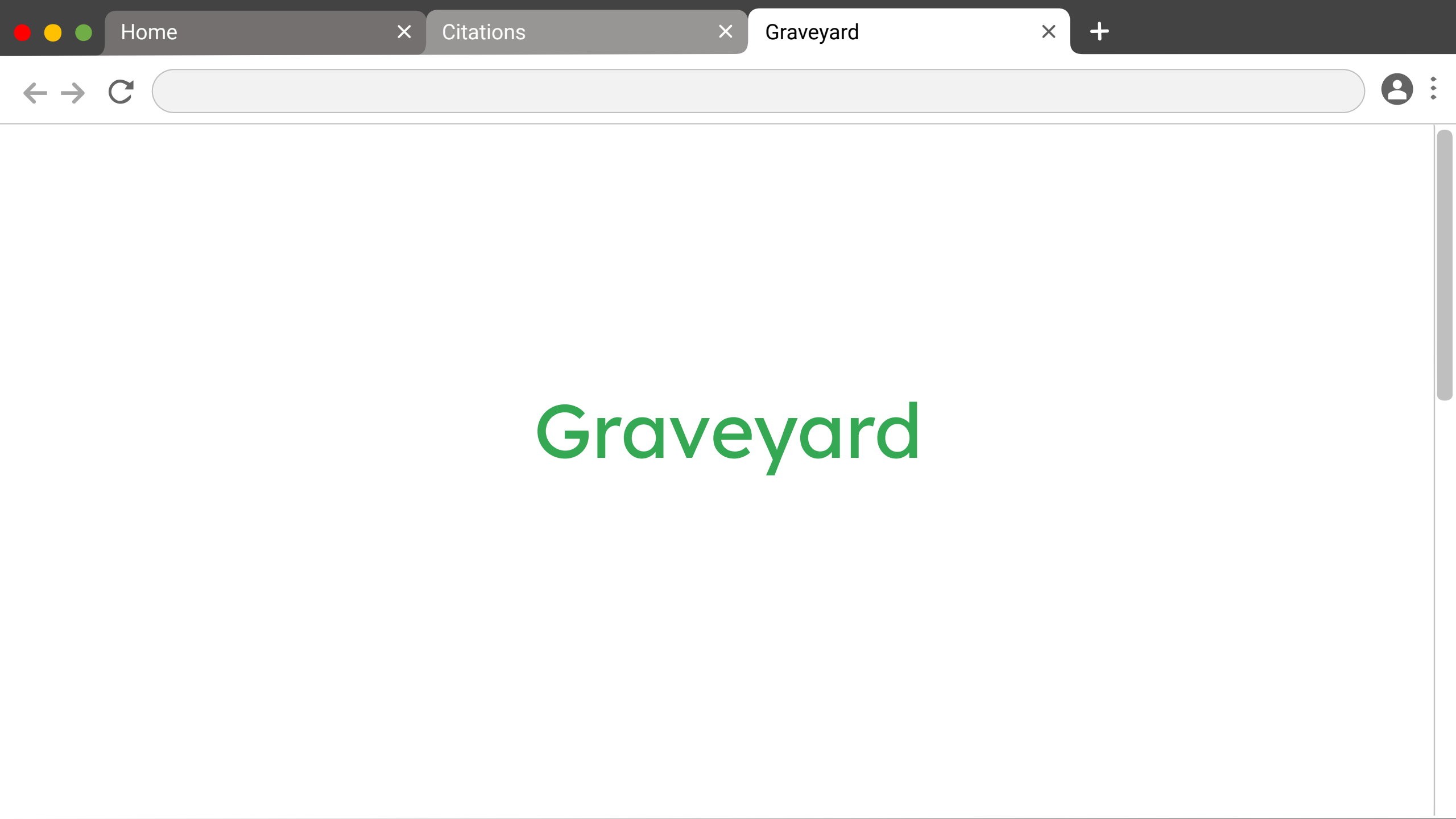

Home
Feature Engineering
GitHub
https://github.com/Mossd-2/DATASCI-207-Applied-Machine-Learning
GitHub Project Repo

Works Cited

1. Data from: https://bigquery.cloud.google.com/table/bigquery-public-data:google_analytics_sample.ga_sessions_20170801
   a. *License: CC0: Public Domain*

2. "Logistic Regression, Artificial Neural Networks, and Linear Separability." *Logistic Regression, Artificial Neural Networks, and Linear Separability – Machine Learning for Biologists*, carpentries-incubator.github.io/ml4bio-workshop/05-logit-ann/index.html. Accessed 5 Aug. 2024.

3. Wilcox, K., & Wang, L. (2022). Jointly Modeling Participant-Level Data and Summary Statistics for Treatment Differences. Multivariate Behavioral Research, 57, 175 - 176. https://doi.org/10.1080/00273171.2022.2030204.


Home
Citations
Graveyard
Graveyard

Credits.

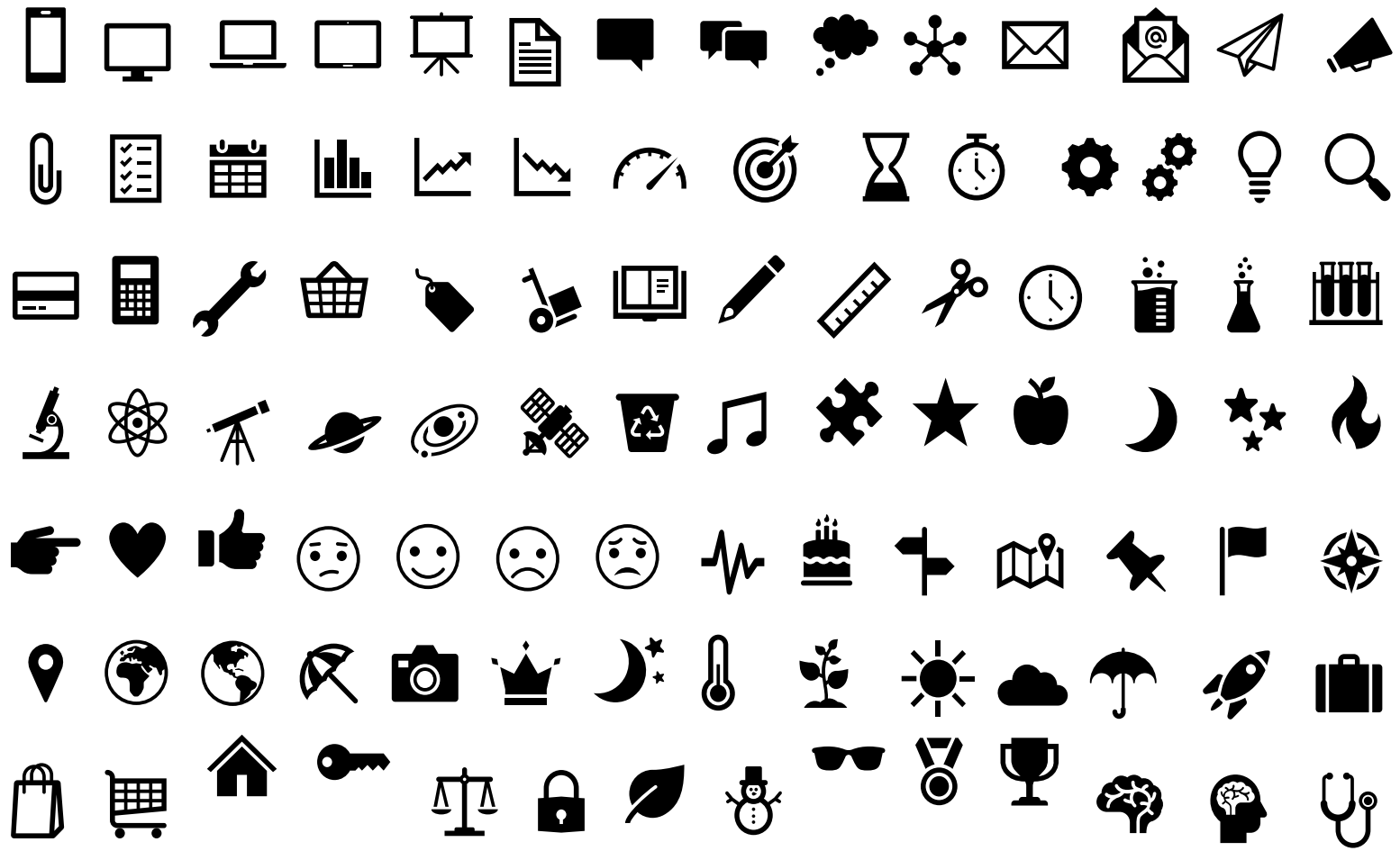
Presentation Template: SlidesMania

Images: Unsplash

Fonts used in this presentation: Roboto and Lexend Deca

Need help editing this template? Check out this video

Editable Icons



Free themes and templates for
Google Slides or **PowerPoint**