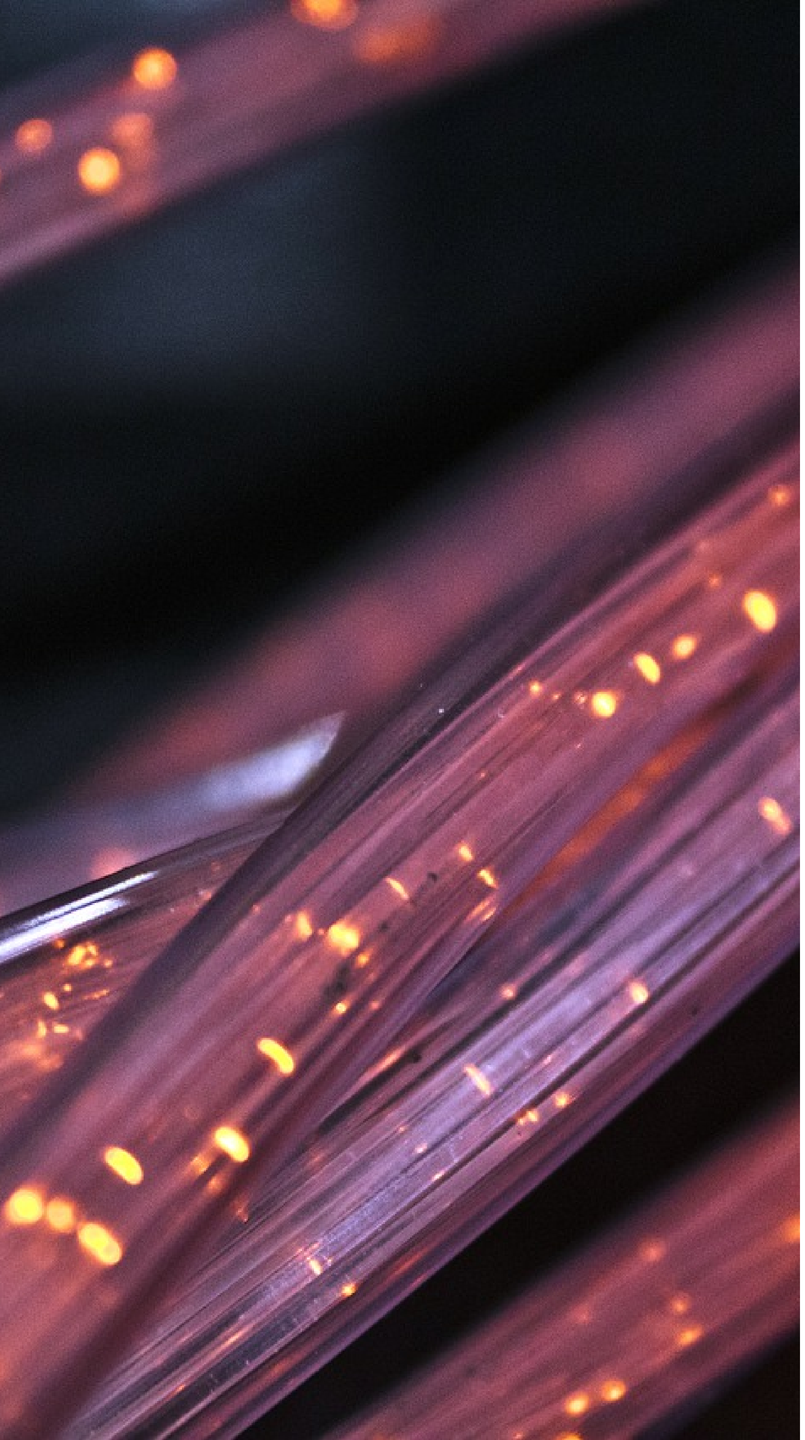


Deep Learning Alpha

Project Update #4

Sergio Rio & Jason Wheeler

MADE WITH beautiful.ai

Our Work

- ## What's the Significance?

  User or customer journey identification and analysis are valuable to any business in general. It allows them to identify customer behaviors, preferences, and trends in order to improve customer satisfaction, retention and revenue growth.

- ## Our Tasks:

  1. Clustering of customer journey based on numerical and categorical feature types.
  2. Classification of course specialization buyers given sequences of URLs visited over time.

Task 1: Clustering

- It's an unsupervised learning analysis and modeling problem where no predefined customer groups or categories were provided.

- Significant data cleaning, wrangling, and patching efforts. Inexistent data dictionary (meaning) and lack of documented associated business rules.

- It deals with a variety of feature types (numerical, timestamps, and most of them categorical)

Task 1: Clustering

- The standard k-means algorithm isn't directly applicable to categorical data, for various reasons. The sample space for categorical data is discrete and doesn't have a natural origin. A Euclidean distance function on such a space isn't really meaningful.
- There's are two relevant variations of k-means: k-modes and k-prototypes. Both were introduced in the paper by Zhexue Huang, and are suitable for categorical data. The latter uses a distance measure that mixes the Hamming distance for categorical features and the Euclidean distance for numeric features.

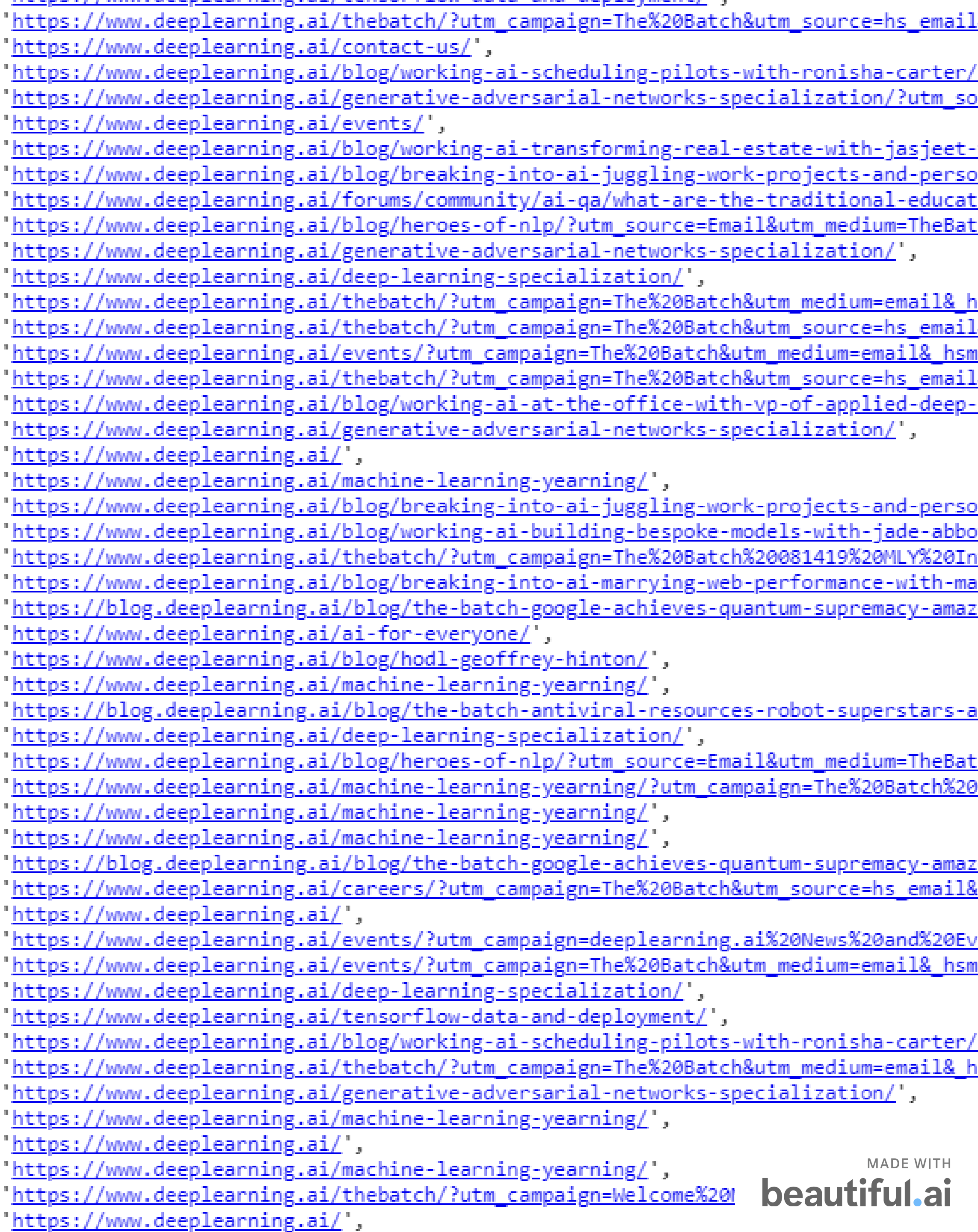
Clustering Demo

```
<Link href={'#!/' + name
)
)}
</div>
<Preview code={code} evalInContext=
div>
showCode ? (
<div>
<Editor code={code} onChange=
<button type="button" className=
Hide code
</button>
</div>
) : (
<button type="button" className=
Show code
</button>
)}
</div>
```

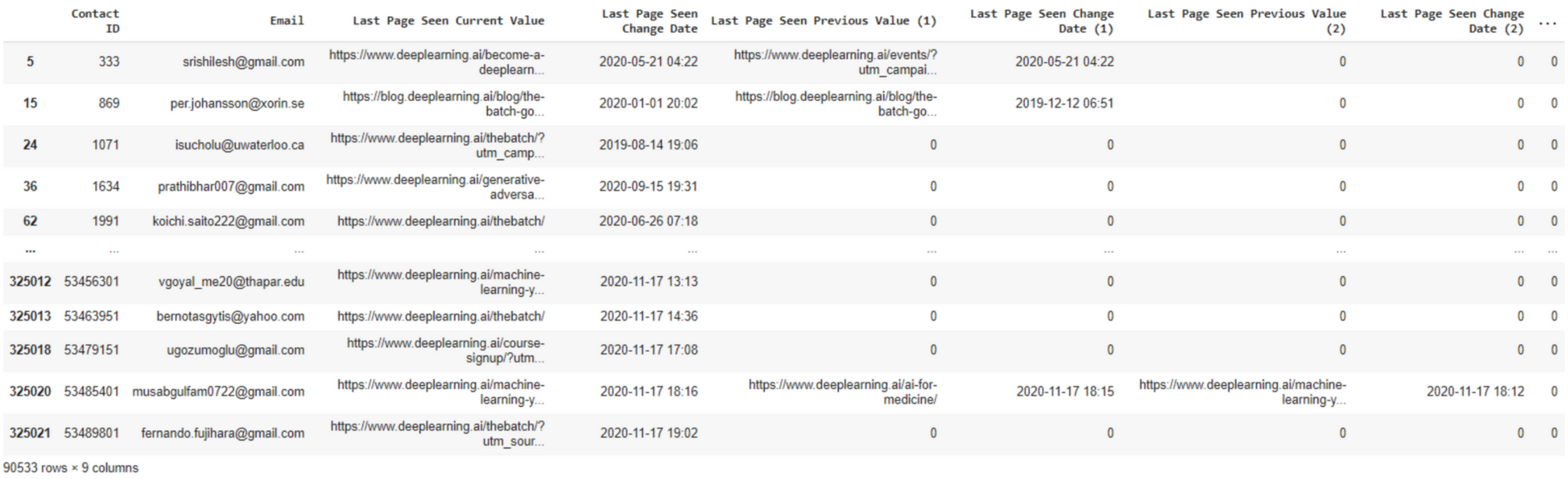
Task 2: Classification

Data Preprocessing

- Steps:
  1. Drop samples without pageview history
  2. Drop extraneous columns/features
  3. Analyze URLs for groupings into categories based on page content and URL substring associations
  4. Map categorical encodings to URLs
  5. Collate sequence data by user --> [num_sessions, IP Country, URL0, URL1, URL2, URL3]

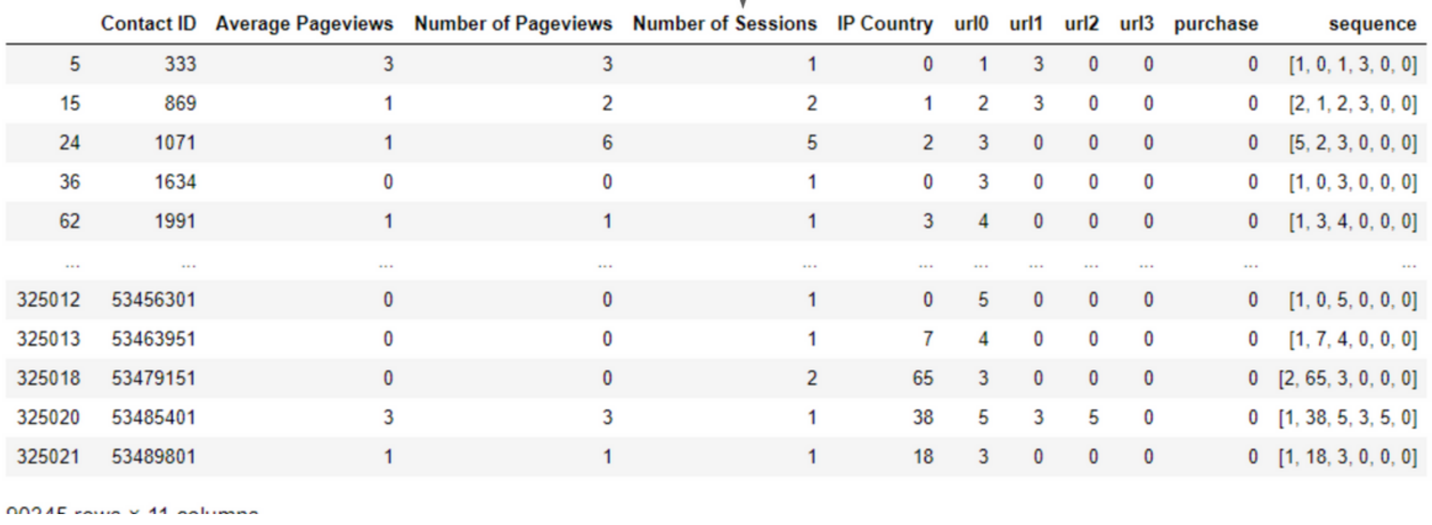


```
'https://www.deeplearning.ai/tensorflow-data-and-deployment/',
'https://www.deeplearning.ai/thebatch/?utm_campaign=The%20Batch&utm_source=hs_email
'https://www.deeplearning.ai/contact-us/',
'https://www.deeplearning.ai/blog/working-ai-scheduling-pilots-with-ronisha-carter/
'https://www.deeplearning.ai/generative-adversarial-networks-specialization/?utm_so
'https://www.deeplearning.ai/events/',
'https://www.deeplearning.ai/blog/working-ai-transforming-real-estate-with-jasjeet-
'https://www.deeplearning.ai/blog/breaking-into-ai-juggling-work-projects-and-perso
'https://www.deeplearning.ai/forums/community/ai-qa/what-are-the-traditional-educat
'https://www.deeplearning.ai/blog/heroes-of-nlp/?utm_source=Email&utm_medium=TheBat
'https://www.deeplearning.ai/generative-adversarial-networks-specialization/',
'https://www.deeplearning.ai/deep-learning-specialization/',
'https://www.deeplearning.ai/thebatch/?utm_campaign=The%20Batch&utm_medium=email&_h
'https://www.deeplearning.ai/thebatch/?utm_campaign=The%20Batch&utm_source=hs_email
'https://www.deeplearning.ai/events/?utm_campaign=The%20Batch&utm_medium=email&_hsm
'https://www.deeplearning.ai/thebatch/?utm_campaign=The%20Batch&utm_source=hs_email
'https://www.deeplearning.ai/blog/working-ai-at-the-office-with-vp-of-applied-deep-
'https://www.deeplearning.ai/generative-adversarial-networks-specialization/',
'https://www.deeplearning.ai/',
'https://www.deeplearning.ai/machine-learning-yearning/',
'https://www.deeplearning.ai/blog/breaking-into-ai-juggling-work-projects-and-perso
'https://www.deeplearning.ai/blog/working-ai-building-bespoke-models-with-jade-abbo
'https://www.deeplearning.ai/thebatch/?utm_campaign=The%20Batch%2008149%20MLY%20In
'https://www.deeplearning.ai/blog/breaking-into-ai-marrying-web-performance-with-ma
'https://blog.deeplearning.ai/blog/the-batch-google-achieves-quantum-supremacy-amaz
'https://www.deeplearning.ai/ai-for-everyone/',
'https://www.deeplearning.ai/blog/hodl-geoffrey-hinton/',
'https://www.deeplearning.ai/machine-learning-yearning/',
'https://blog.deeplearning.ai/blog/the-batch-antiviral-resources-robot-superstars-a
'https://www.deeplearning.ai/deep-learning-specialization/',
'https://www.deeplearning.ai/blog/heroes-of-nlp/?utm_source=Email&utm_medium=TheBat
'https://www.deeplearning.ai/machine-learning-yearning/?utm_campaign=The%20Batch%20
'https://www.deeplearning.ai/machine-learning-yearning/',
'https://www.deeplearning.ai/machine-learning-yearning/',
'https://blog.deeplearning.ai/blog/the-batch-google-achieves-quantum-supremacy-amaz
'https://www.deeplearning.ai/careers/?utm_campaign=The%20Batch&utm_source=hs_email&
'https://www.deeplearning.ai/',
'https://www.deeplearning.ai/events/?utm_campaign=deeplearning.ai%20News%20and%20Ev
'https://www.deeplearning.ai/events/?utm_campaign=The%20Batch&utm_medium=email&_hsm
'https://www.deeplearning.ai/deep-learning-specialization/',
'https://www.deeplearning.ai/tensorflow-data-and-deployment/',
'https://www.deeplearning.ai/blog/working-ai-scheduling-pilots-with-ronisha-carter/
'https://www.deeplearning.ai/thebatch/?utm_campaign=The%20Batch&utm_medium=email&_h
'https://www.deeplearning.ai/generative-adversarial-networks-specialization/',
'https://www.deeplearning.ai/machine-learning-yearning/',
'https://www.deeplearning.ai/',
'https://www.deeplearning.ai/machine-learning-yearning/',
'https://www.deeplearning.ai/thebatch/?utm_campaign=Welcome%20M
'https://www.deeplearning.ai/',
```



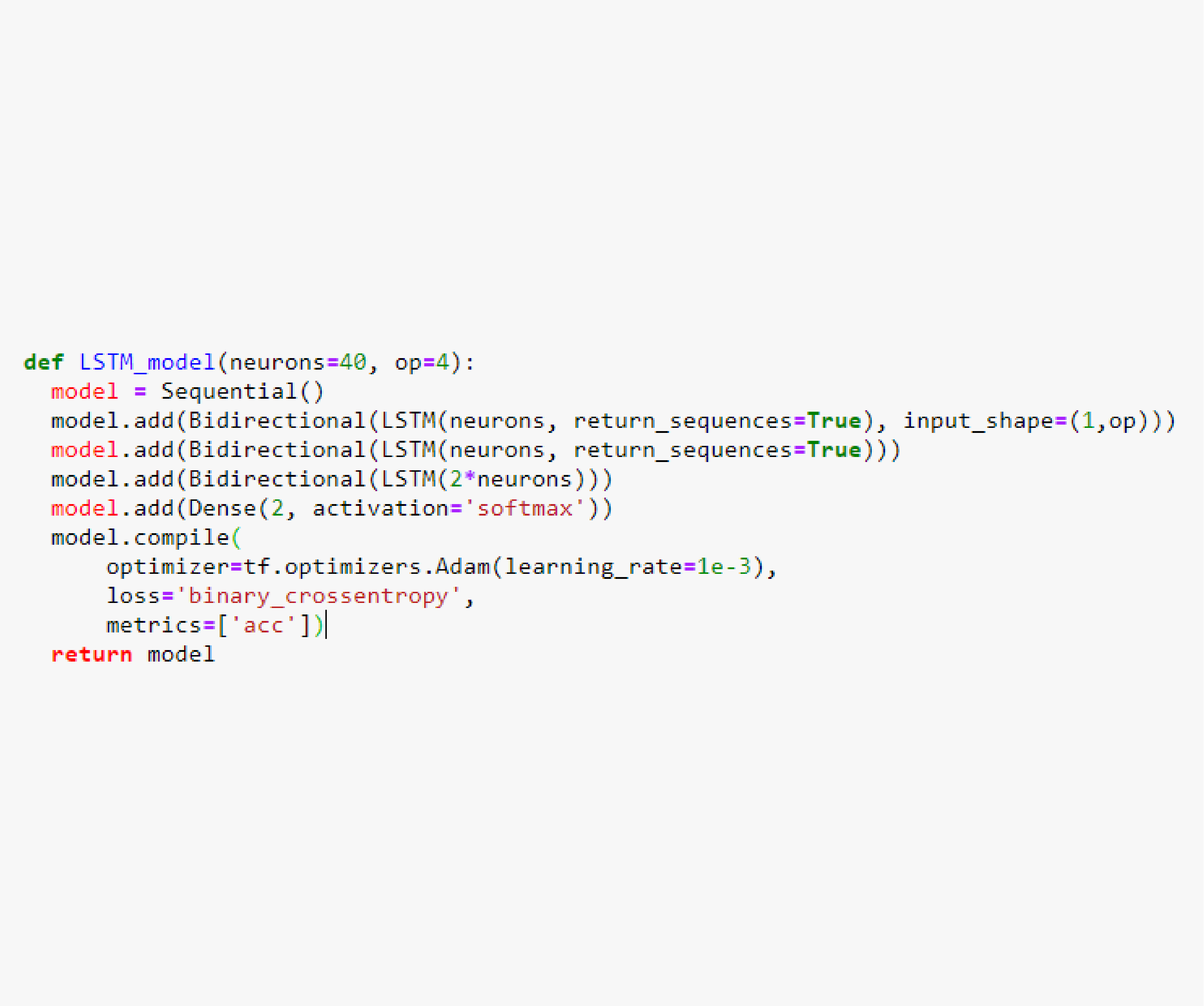
	Contact ID	Email	Last Page Seen Current Value	Last Page Seen Change Date	Last Page Seen Previous Value (1)	Last Page Seen Change Date (1)	Last Page Seen Previous Value (2)	Last Page Seen Change Date (2)	...
5	333	srishilesh@gmail.com	https://www.deeplearning.ai/become-a-deeplearn...	2020-05-21 04:22	https://www.deeplearning.ai/events/?utm_campai...	2020-05-21 04:22	0	0	0
15	869	per.johansson@xorin.se	https://blog.deeplearning.ai/blog/the-batch-go...	2020-01-01 20:02	https://blog.deeplearning.ai/blog/the-batch-go...	2019-12-12 06:51	0	0	0
24	1071	isucholu@uwaterloo.ca	https://www.deeplearning.ai/thebatch/?utm_camp...	2019-08-14 19:06	0	0	0	0	0
36	1634	prathibhar007@gmail.com	https://www.deeplearning.ai/generative-adversa...	2020-09-15 19:31	0	0	0	0	0
62	1991	koichi.saito222@gmail.com	https://www.deeplearning.ai/thebatch/	2020-06-26 07:18	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
325012	53456301	vgoyal_me20@thapar.edu	https://www.deeplearning.ai/machine-learning-y...	2020-11-17 13:13	0	0	0	0	0
325013	53463951	bernotasgytis@yahoo.com	https://www.deeplearning.ai/thebatch/	2020-11-17 14:36	0	0	0	0	0
325018	53479151	ugozumoglu@gmail.com	https://www.deeplearning.ai/course-signup/?utm...	2020-11-17 17:08	0	0	0	0	0
325020	53485401	musabgulfam0722@gmail.com	https://www.deeplearning.ai/machine-learning-y...	2020-11-17 18:16	https://www.deeplearning.ai/ai-for-medicine/	2020-11-17 18:15	https://www.deeplearning.ai/machine-learning-y...	2020-11-17 18:12	0
325021	53489801	fernando.fujihara@gmail.com	https://www.deeplearning.ai/thebatch/?utm_sour...	2020-11-17 19:02	0	0	0	0	0

90533 rows × 9 columns

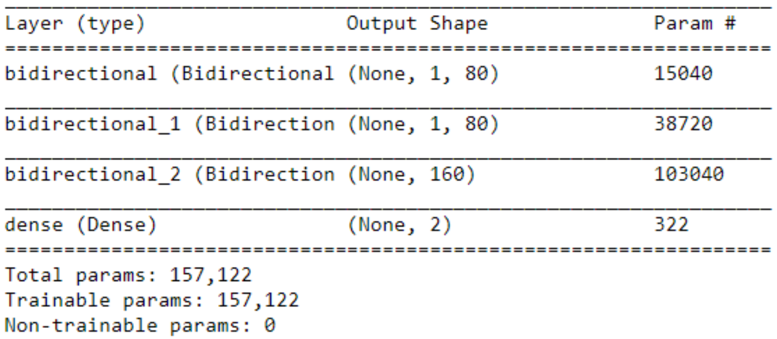


	Contact ID	Average Pageviews	Number of Pageviews	Number of Sessions	IP Country	url0	url1	url2	url3	purchase	sequence
5	333	3	3	1	0	1	3	0	0	0	[1, 0, 1, 3, 0, 0]
15	869	1	2	2	1	2	3	0	0	0	[2, 1, 2, 3, 0, 0]
24	1071	1	6	5	2	3	0	0	0	0	[5, 2, 3, 0, 0, 0]
36	1634	0	0	1	0	3	0	0	0	0	[1, 0, 3, 0, 0, 0]
62	1991	1	1	1	3	4	0	0	0	0	[1, 3, 4, 0, 0, 0]
...	...	...	...	...	...	...	...	...	...	...	...
325012	53456301	0	0	1	0	5	0	0	0	0	[1, 0, 5, 0, 0, 0]
325013	53463951	0	0	1	7	4	0	0	0	0	[1, 7, 4, 0, 0, 0]
325018	53479151	0	0	2	65	3	0	0	0	0	[2, 65, 3, 0, 0, 0]
325020	53485401	3	3	1	38	5	3	5	0	0	[1, 38, 5, 3, 5, 0]
325021	53489801	1	1	1	18	3	0	0	0	0	[1, 18, 3, 0, 0, 0]

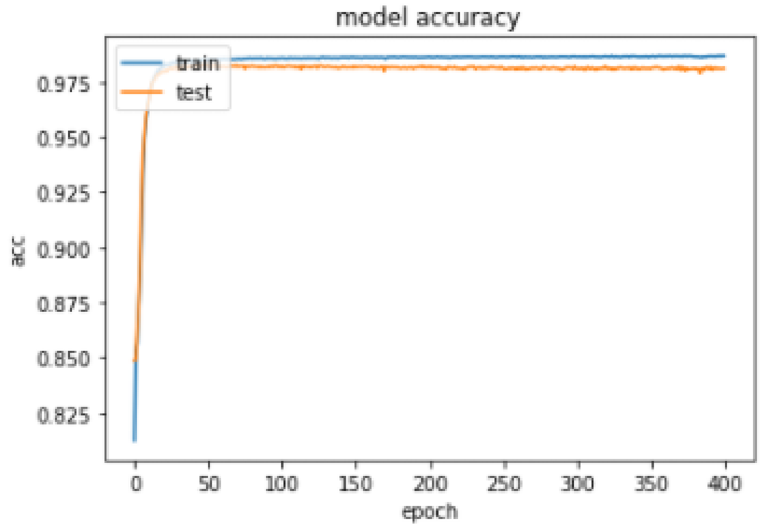
90245 rows × 11 columns

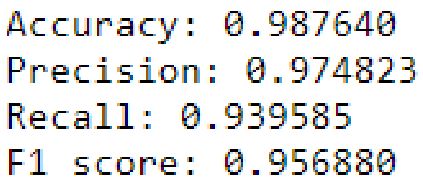


```python
def LSTM_model(neurons=40, op=4):
  model = Sequential()
  model.add(Bidirectional(LSTM(neurons, return_sequences=True), input_shape=(1,op)))
  model.add(Bidirectional(LSTM(neurons, return_sequences=True)))
  model.add(Bidirectional(LSTM(2*neurons)))
  model.add(Dense(2, activation='softmax'))
  model.compile(
      optimizer=tf.optimizers.Adam(learning_rate=1e-3),
      loss='binary_crossentropy',
      metrics=['acc'])
  return model
```

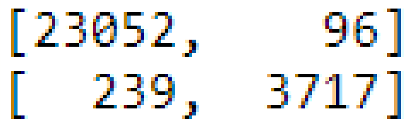


```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
bidirectional (Bidirectional (None, 1, 80)             15040
_________________________________________________________________
bidirectional_1 (Bidirection (None, 1, 80)             38720
_________________________________________________________________
bidirectional_2 (Bidirection (None, 160)               103040
_________________________________________________________________
dense (Dense)                (None, 2)                 322
=================================================================
Total params: 157,122
Trainable params: 157,122
Non-trainable params: 0
```

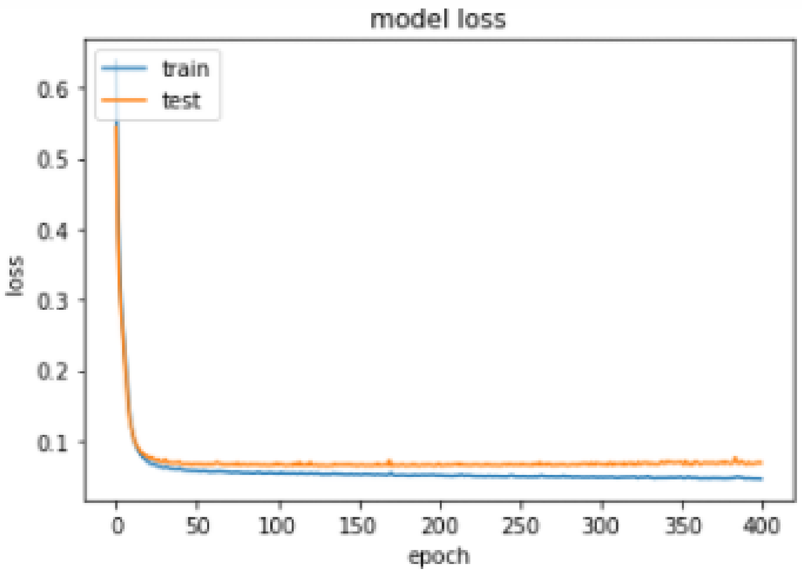


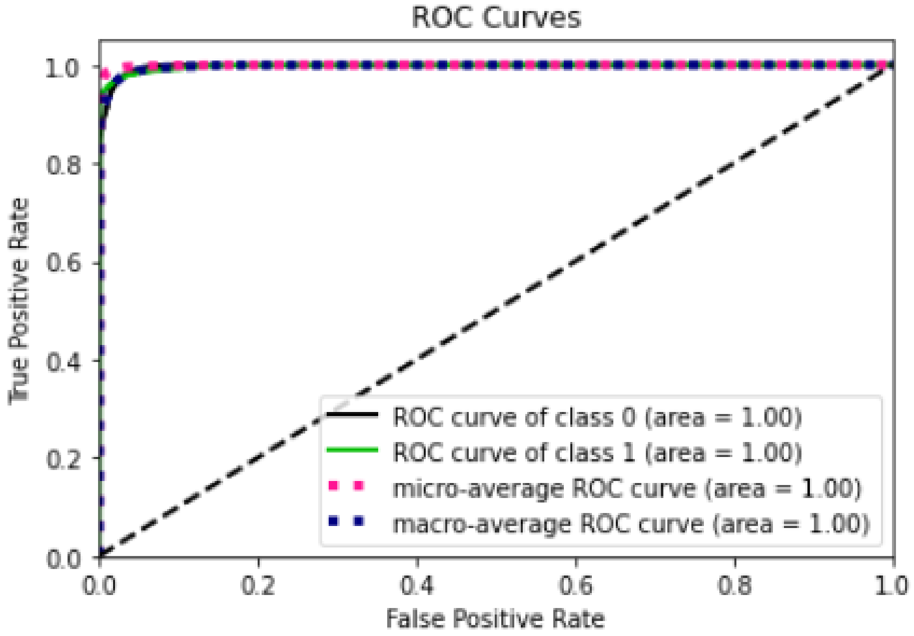


```
Accuracy: 0.987640
Precision: 0.974823
Recall: 0.939585
F1 score: 0.956880
```



```
[23052,    96]
[  239,  3717]
```





Next Steps / System Design

- Explore cluster affinity/dissimilarities between groups of interest #1 and #2 to objectively assess potential incentives/nudges
- Create data cleaning/wrangling pipeline in python (automation)
- Save models in binary format and create two APIs/web services to expose them through AWS7.

- Test sequence classification model on different targets.
- Train classification model on varying sequence lengths and categorical encoding schemes.
- Classify users by cluster assignment given pageview sequences.

Time Plan

● 01/31 - 02/06

- Model development and deployment
- Debugging
- Improve project repository

● 02/07 - 02/13

- Finalize project presentation
- Debugging