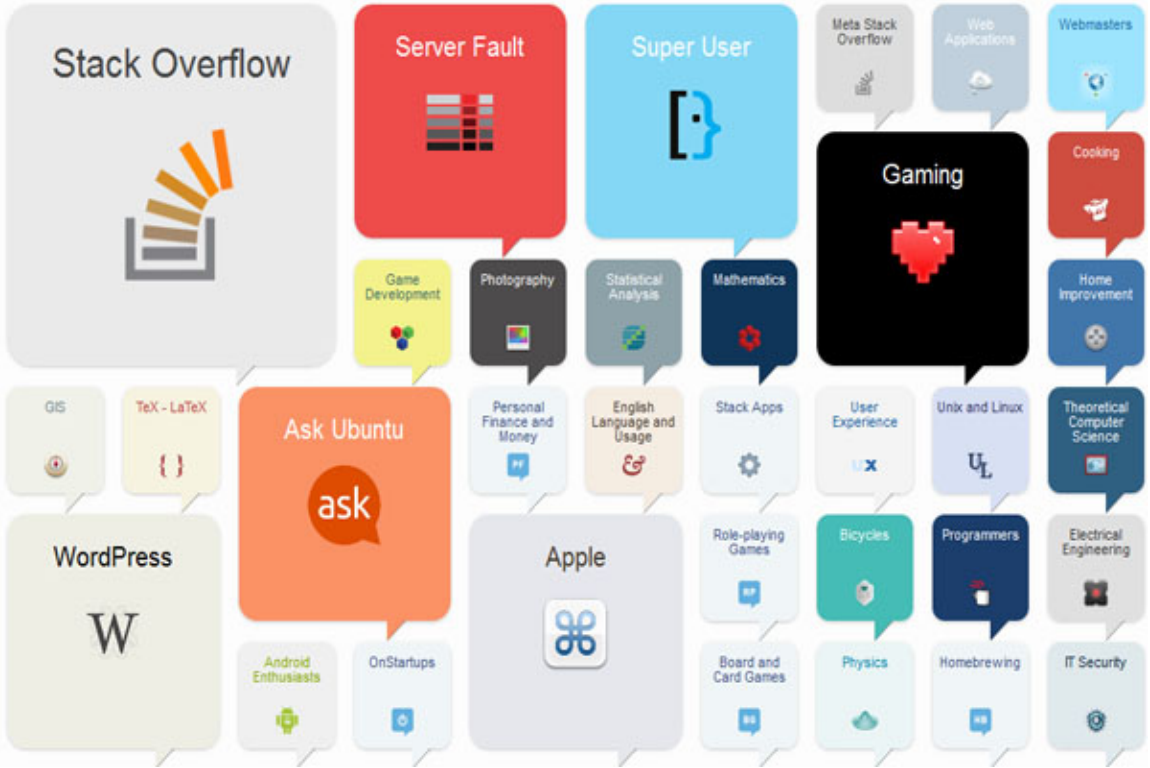
StackExchange

By Team3

Introduction

- Stack Exchange is a network of question-and-answer websites on topics in diverse fields, each site covering a specific topic, where questions, answers, and users are subject to a reputation award process



stackoverflow

Users Meta Data

Problem Statement

Goal is to predict "LastAccessDate" from the Stack OverFlow Meta Data of Users profile. And try to "minimize the error"

Source Data

- https://ia800107.us.archive.org/view_archive.php?archive=/27/items/stackexchange/meta.stackoverflow.com.7z
- Users.xml 2020-12-07 01:12:31 617425520

* Id (It is the Id of the user with respect to specific site/domain)
* Reputation (It is a way to measure user expertise)
* CreationDate (when the user is created)
* DisplayName
* LastAccessDate (Datetime user last loaded a page; updated every 30 min at most)
* WebsiteUrl
* Location
* AboutMe
* Views (Number of times the profile is viewed by X number of profilers)
* UpVotes (How many upvotes the user has cast)
* DownVotes (How many downvotes the user has cast)
* ProfileImageUrl
* AccountId (User's Stack Exchange Network profile ID)

Approach

EDA

- Data Conversion from XML to Pandas DataFrame
- Data Cleansing & Data Analysis w.r.to each feature
- New Features derivation from Create and LastAccess Date features

Build Model

- Used Pretrained Stanford's GloVe word embedding for the text to feature vector
- Applied normalization on independent features apart from text vectors
- Used LSTM model to predict months duration

Tune Model

- To achieve generalisation, tuned hypermeters of LSTM model
- Such as Kernal-initializer, Dropout Ratio, Activation Function, Optimizers, Loss Functions, Metrics w.r.to test loss and validation loss data points

Apply Prediction

- Gets the predicted months duration for all the test data points
- And compared the model predicted results with GroudTruth values
- Choosen best model based on least validation loss

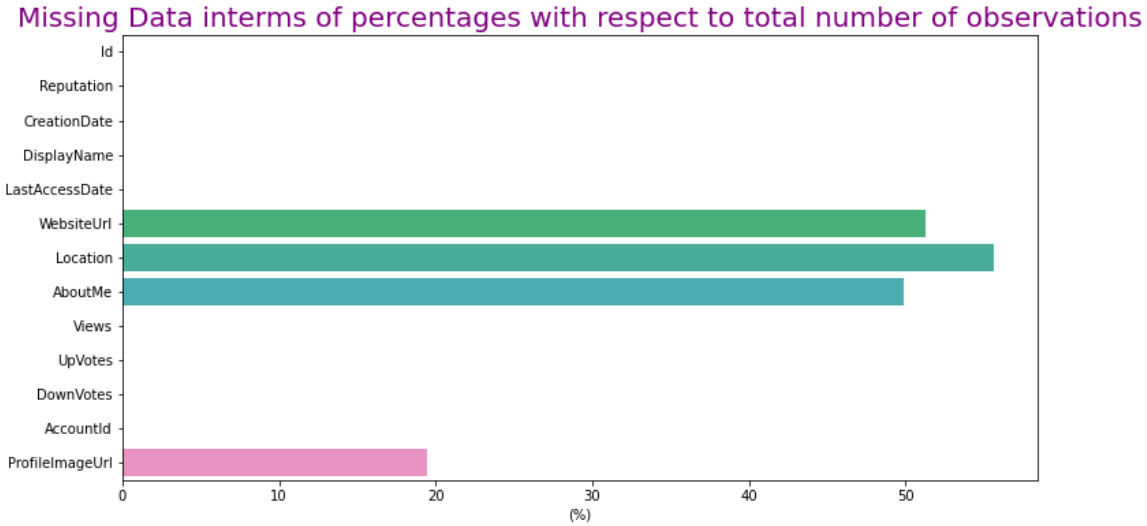
EDA – Data Conversion & Cleansing

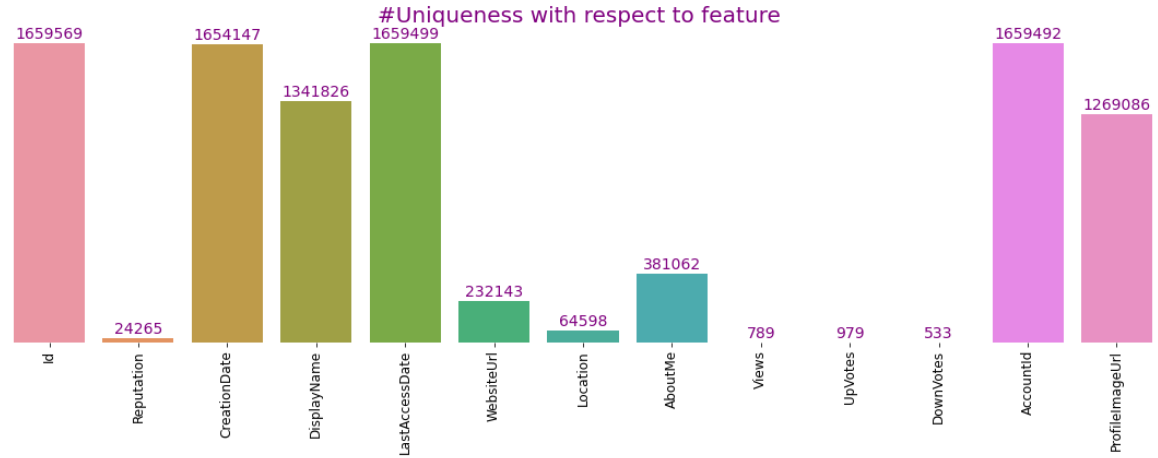
```
<?xml version="1.0" encoding="utf-8"?>
<users>
  <row Id="-1" Reputation="1" CreationDate="2008-07-31T00:00:00.000" DisplayName="Community" LastAccessDate="2014-04-17T00:17:22.260"
WebsiteUrl="http://meta.stackexchange.com/" Location="on the server farm" AboutMe="&lt;p&gt;Hi, I'm not really a person.&lt;/
p&gt;&#xA;&#xA;&lt;p&gt;I'm a background process that helps keep this site clean!&lt;/p&gt;&#xA;&#xA;&lt;p&gt;I do things like&lt;/
p&gt;&#xA;&#xA;&lt;ul&gt;&#xA;&lt;li&gt;Randomly poke old unanswered questions every hour so they get some attention&lt;/
li&gt;&#xA;&lt;li&gt;Own community questions and answers so nobody gets unnecessary reputation from them&lt;/li&gt;&#xA;&lt;li&gt;Own
downvotes on spam/evil posts that get permanently deleted&lt;/li&gt;&#xA;&lt;li&gt;Own suggested edits from anonymous users&lt;/
li&gt;&#xA;&lt;li&gt;&lt;a href=&quot;http://meta.stackexchange.com/a/92006&quot;&gt;Remove abandoned questions&lt;/a&gt;&lt;/
li&gt;&#xA;&lt;/ul&gt;&#xA;" Views="7537" UpVotes="24132" DownVotes="3669" AccountId="-1" />
  <row Id="1" Reputation="60145" CreationDate="2008-07-31T14:22:31.000" DisplayName="Jeff Atwood"
LastAccessDate="2020-10-19T18:40:56.053" WebsiteUrl="http://www.codinghorror.com/blog/" Location="El Cerrito, CA"
AboutMe="&lt;p&gt;&lt;a href=&quot;http://www.codinghorror.com/blog/archives/001169.html&quot; rel=&quot;nofollow&quot;&gt;Stack
Overflow Valued Associate #00001&lt;/a&gt;&lt;/p&gt;&#xA;&#xA;&lt;p&gt;Wondering how our software development process works? &lt;a
href=&quot;http://www.youtube.com/watch?v=08xQLGWTSag&quot; rel=&quot;nofollow&quot;&gt;Take a look!&lt;/a&gt;&lt;/
p&gt;&#xA;&#xA;&lt;p&gt;Find me &lt;a href=&quot;http://twitter.com/codinghorror&quot; rel=&quot;nofollow&quot;&gt;on twitter&lt;/a&gt;,
or &lt;a href=&quot;http://www.codinghorror.com/blog&quot; rel=&quot;nofollow&quot;&gt;read my blog&lt;/a&gt;. Don't say I didn't warn
you &lt;em&gt;because I totally did&lt;/em&gt;.&lt;/p&gt;&#xA;&#xA;&lt;p&gt;However, &lt;a href=&quot;http://www.codinghorror.com/blog/
2012/02/farewell-stack-exchange.html&quot; rel=&quot;nofollow&quot;&gt;I no longer work at Stack Exchange, Inc&lt;/a&gt;. I'll miss you
all. Well, &lt;em&gt;some&lt;/em&gt; of you, anyway. :)&lt;/p&gt;&#xA;" Views="5548" UpVotes="194" DownVotes="20"
ProfileImageUrl="https://www.gravatar.com/avatar/51d623f33f8b83095db84ff35e15dbe8?s=128&amp;amp;d=identicon&amp;amp;r=PG"
AccountId="1" />
  <row Id="2" Reputation="5752" CreationDate="2008-07-31T14:22:31.000" DisplayName="Geoff Dalgas"
LastAccessDate="2020-08-27T14:55:29.363" WebsiteUrl="http://stackoverflow.com" Location="Corvallis, OR" AboutMe="&lt;p&gt;Developer on
the Stack Overflow team.  Find me on&lt;/p&gt;&#xA;&#xA;&lt;p&gt;&lt;a href=&quot;http://www.twitter.com/SuperDalgas&quot;
rel=&quot;nofollow noreferrer&quot;&gt;Twitter&lt;/a&gt;&#xA;&lt;br&gt;&lt;br&gt;&#xA;&lt;a href=&quot;http://blog.stackoverflow.com/
2009/05/welcome-stack-overflow-valued-associate-00003/&quot;&gt;Stack Overflow Valued Associate #00003&lt;/a&gt;&lt;/p&gt;&#xA;"
Views="991" UpVotes="76" DownVotes="15" ProfileImageUrl="https://i.stack.imgur.com/nDllk.png?s=256&amp;g=1" AccountId="2" />
```

Id	Reputation	CreationDate	DisplayName	LastAccessDate	WebsiteUrl	Location	AboutMe	Views	UpVotes	DownVotes	AccountId	ProfileImageUrl
-1	1	2008-07-31T00:00:00.000	Community	2014-04-17T00:17:22.260	http://meta.stackexchange.com/	on the server farm	Hi, I'm not really a person. \n\n I'm ...	7537	24132	3669	-1	NaN
1	60145	2008-07-31T14:22:31.000	Jeff Atwood	2020-10-19T18:40:56.053	http://www.codinghorror.com/blog/	El Cerrito, CA	<a href="http://www.codinghorror.com/blog/a...	5548	194	20	1	https://www.gravatar.com/avatar/51d623f33f8b83...
2	5752	2008-07-31T14:22:31.000	Geoff Dalgas	2020-08-27T14:55:29.363	http://stackoverflow.com	Corvallis, OR	Developer on the Stack Overflow team. Find...	991	76	15	2	https://i.stack.imgur.com/nDllk.png?s=256&g=1
3	15256	2008-07-31T14:22:31.000	Jarrod Dixon	2020-09-14T17:27:57.783	http://jarroddixon.com	Raleigh, NC, United States	<a href="http://blog.stackoverflow.com/2009...	1108	353	7	3	https://i.stack.imgur.com/2mwFl.jpg?s=256&g=1
4	32216	2008-07-31T14:22:31.000	Joel Spolsky	2020-06-07T14:41:33.910	https://joelonsoftware.com/	New York, NY	In 2000 I co-founded Fog Creek Software, wh...	9152	39	9	4	https://i.stack.imgur.com/C5gBG.jpg?s=128&g=1

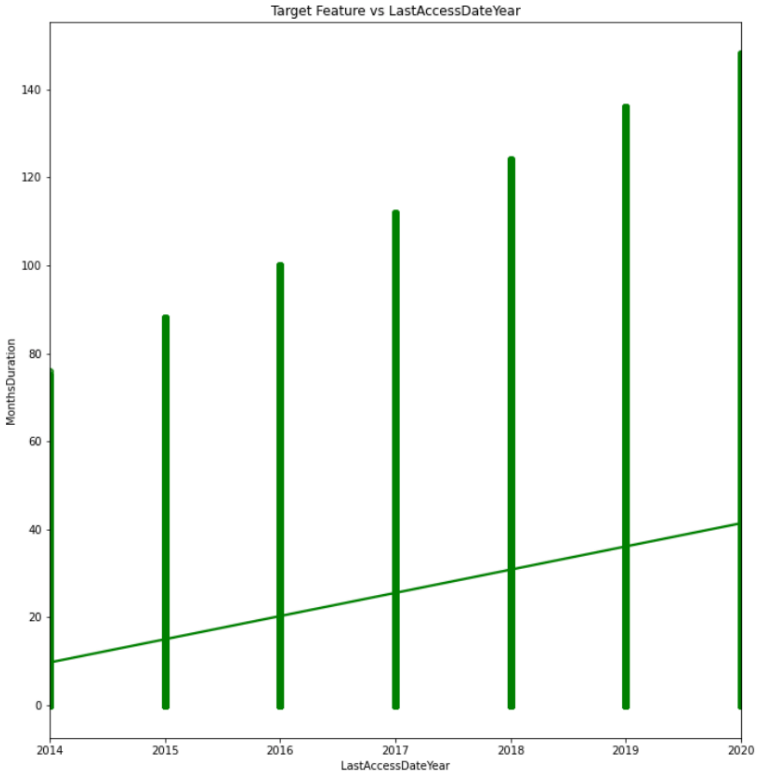
```
after cleaning the text (AboutMe feature) - total records will be  (383225,) out of (1659569,)
```

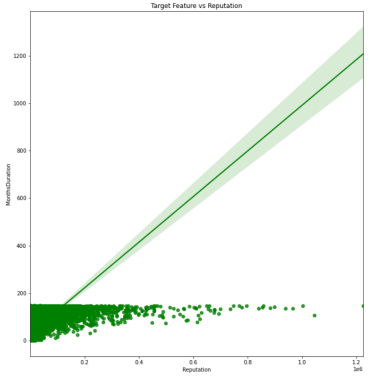
EDA – Missing Data & Uniqueness

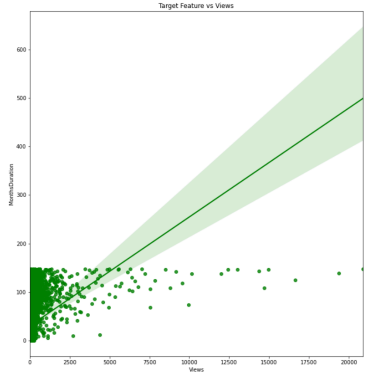


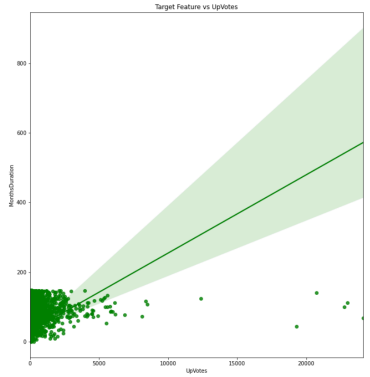


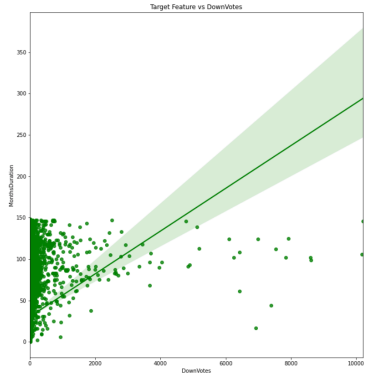
EDA on Numeric Data Types

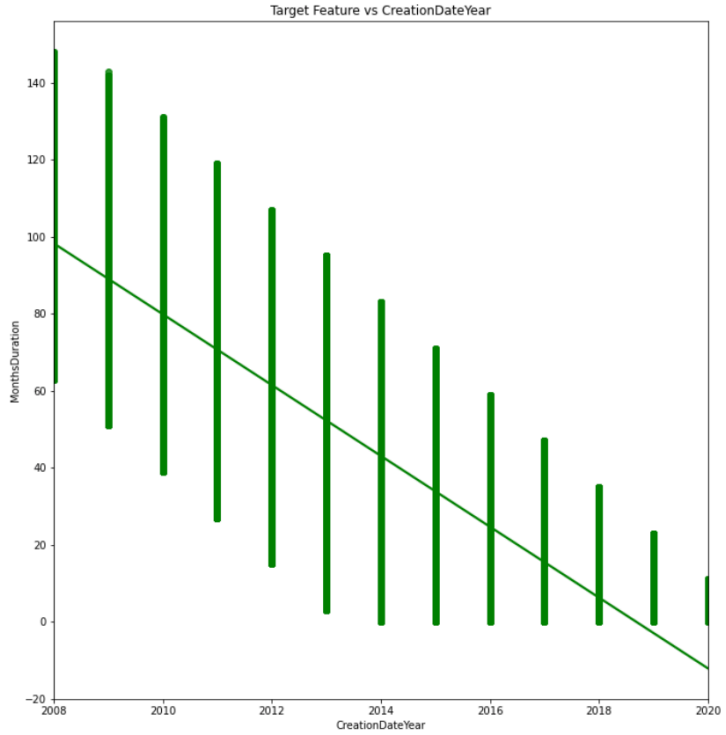






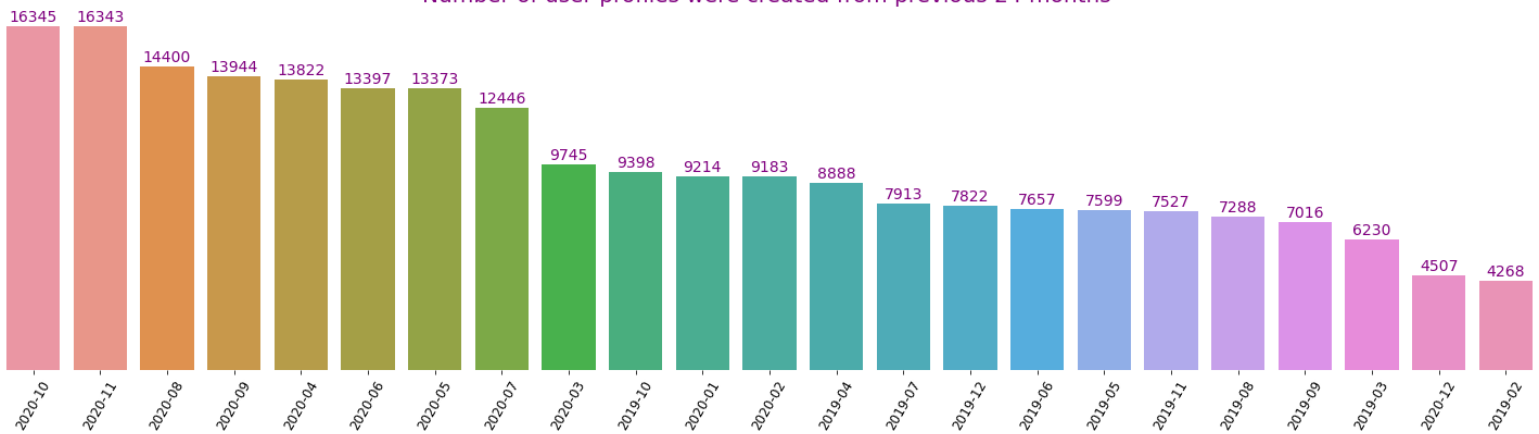




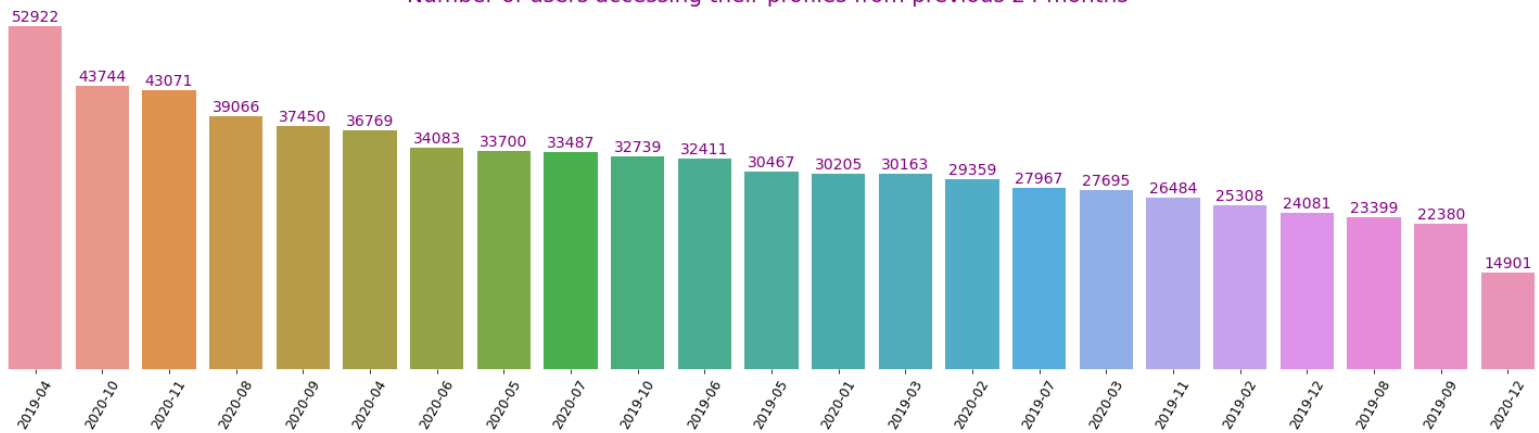


EDA on Date Types

Number of user profiles were created from previous 24 months

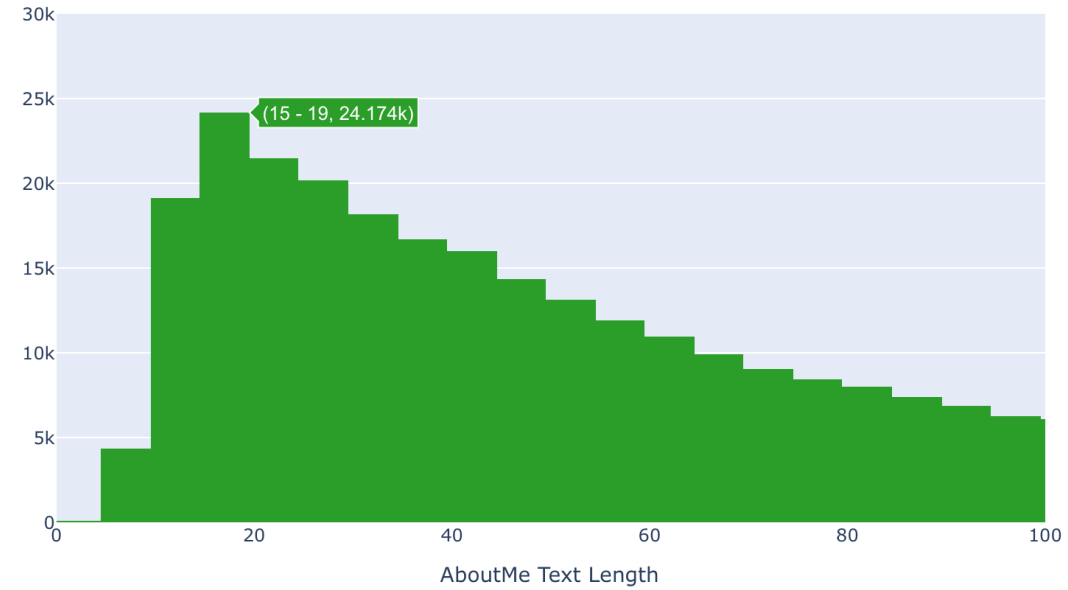


Number of users accessing their profiles from previous 24 months



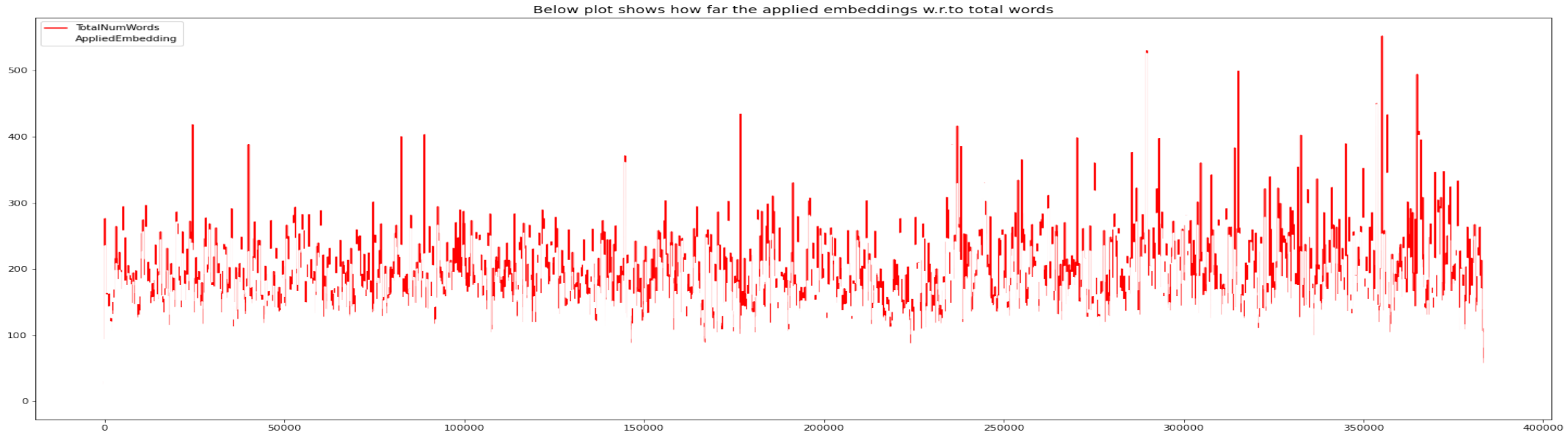
EDA on Text Type

Frequency of AboutMe feature text length



Word Embedding on text feature

v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	...	v_49	Reputation	CreationDate	LastAccessDate	Views	UpVotes	DownVotes	MonthsDuration	CreationDateYear	LastAccessDateYea
0.267495	-0.046916	0.235946	-0.141691	0.100916	-0.158346	-0.394266	-0.193191	0.019417	0.197517	...	0.211529	60145	2008-07-31 14:22:31	2020-10-19 18:40:56.053	5548	194	20	146	2008
0.352862	0.181632	0.522008	0.388044	0.192042	-0.210707	-0.671004	0.093214	-0.028968	0.020160	...	-0.021741	5752	2008-07-31 14:22:31	2020-08-27 14:55:29.363	991	76	15	144	2008
0.157551	-0.124406	-0.306622	0.009339	-0.092997	0.365333	-0.352843	0.056237	0.088811	0.078312	...	0.199671	15256	2008-07-31 14:22:31	2020-09-14 17:27:57.783	1108	353	7	145	2008
0.101777	-0.047044	0.124040	0.023608	-0.036157	-0.100952	-0.497196	-0.176642	-0.083046	-0.079067	...	0.166363	32216	2008-07-31 14:22:31	2020-06-07 14:41:33.910	9152	39	9	142	2008
0.331277	-0.008430	0.224194	0.017008	0.098048	-0.184746	-0.456291	-0.242567	-0.049400	0.079222	...	0.215173	49406	2008-07-31 14:22:31	2020-03-30 18:50:41.280	148	0	0	139	2008



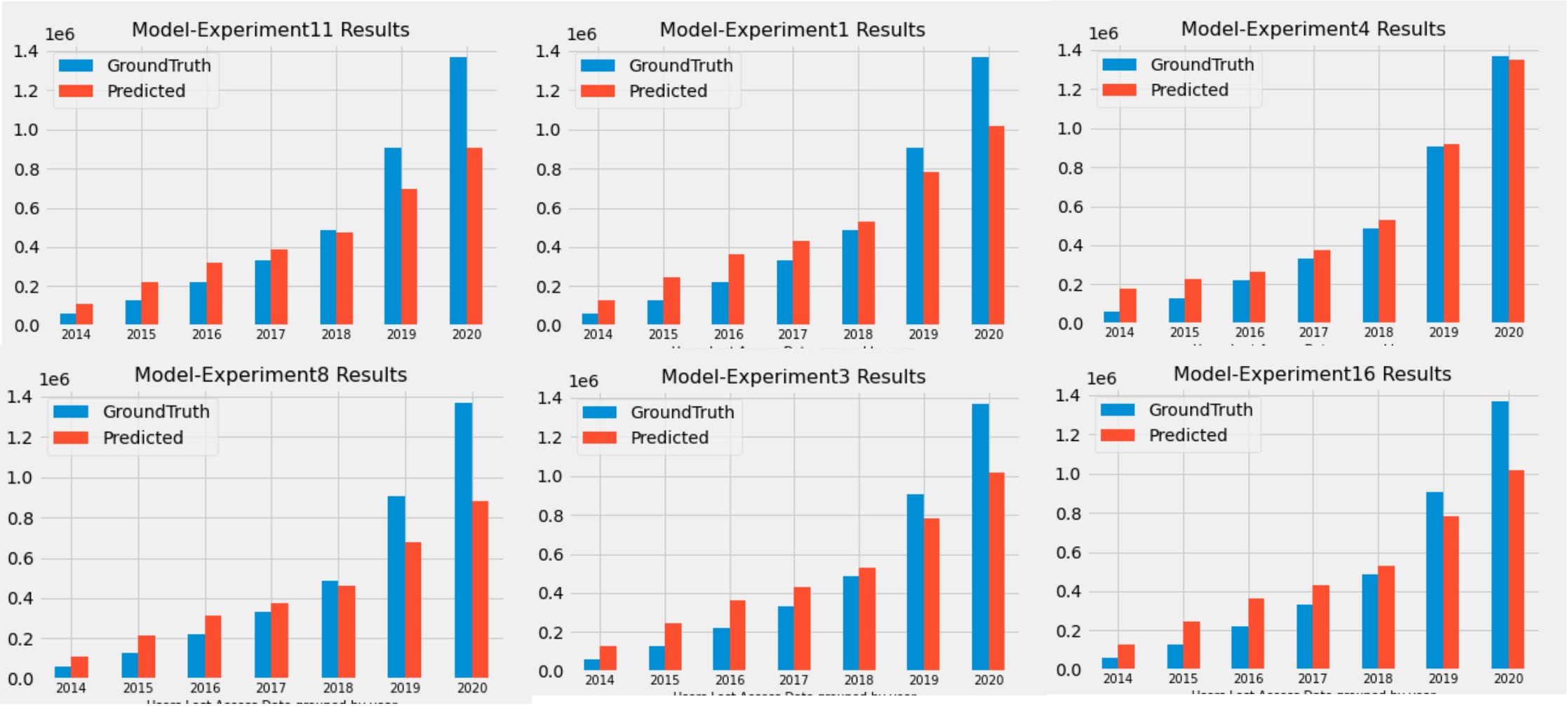
Approach to Model

- Applied Min-Max & Z-score Normalization Techniques on Independent Numeric Features
- Derived MonthsDuration Feature which is our Dependent Feature that needs to predict
- Split done w.r.to Train Test and Validation Data Sets
- Used LSTM with 2 hidden layers and performed Hyper parameter tuning
- Chosen the best parameters based on least validation loss
- Predicted the Dependent Feature on top of test data points
- Converted the derived MonthsDuration feature to Date format

Model Tuning Results

activation	dropout	kernel_initializer	optimizer	loss	metrics	Batch_size	epochs	Test_loss	Val_loss	optimal_epoch	normalization
relu	0.2	normal	Adam	mse	RootMeanSquaredError	64	30	1255.761963	1243.241821	2	z-score
relu	0.2	normal	Adam	mse	RootMeanSquaredError	64	30	1353.406860	1243.245117	1	min-max
relu	0.2	normal	RMSprop	mse	RootMeanSquaredError	64	30	1255.790894	1243.252563	2	z-score
relu	0.2	normal	RMSprop	mse	RootMeanSquaredError	64	30	407.281677	217.227722	6	min-max
relu	0.2	normal	SGD	mse	RootMeanSquaredError	64	30	NaN	NaN	1	z-score
relu	0.2	normal	SGD	mse	RootMeanSquaredError	64	30	1262.088745	1243.853516	2	min-max
relu	0.2	normal	Adam	mae	MeanAbsoluteError	64	30	29.592804	29.388159	2	z-score
relu	0.2	normal	Adam	mae	MeanAbsoluteError	64	30	30.272444	29.397850	1	min-max
relu	0.2	normal	RMSprop	mae	MeanAbsoluteError	64	30	30.194130	29.388554	1	z-score
relu	0.2	normal	RMSprop	mae	MeanAbsoluteError	64	30	30.186201	29.402666	1	min-max
relu	0.2	normal	SGD	mae	MeanAbsoluteError	64	30	29.593204	29.391979	2	z-score
relu	0.2	normal	SGD	mae	MeanAbsoluteError	64	30	30.036415	29.388124	1	min-max
relu	0.2	uniform	Adam	mae	MeanAbsoluteError	64	30	30.326859	29.391895	1	z-score
relu	0.2	uniform	Adam	mae	MeanAbsoluteError	64	30	29.593670	29.388048	2	min-max
relu	0.2	uniform	Adam	mse	RootMeanSquaredError	64	30	1359.548340	1243.238281	1	z-score
relu	0.2	uniform	Adam	mse	RootMeanSquaredError	64	30	1255.750488	1243.237427	4	min-max

Model Tuning Results



Model Predicted Users Last Access Date

User Id	LastAccessDate	PredictedLastAccessDate
326171	2018-05-18 07:44:35.990	2018-05-13 14:18:18
236495	2018-04-30 15:03:21.393	2018-04-25 08:51:49
226455	2020-10-15 11:22:45.943	2020-10-14 13:30:51
192955	2019-06-27 06:34:54.540	2019-05-28 07:14:35
55278	2019-03-07 10:07:04.760	2019-03-01 14:42:13
358688	2020-02-04 14:09:56.430	2020-01-13 23:57:23
29281	2020-04-11 19:43:15.213	2020-04-08 18:23:40
323327	2020-07-30 15:57:39.523	2020-07-14 07:29:35
218062	2016-03-11 03:02:45.810	2016-02-13 00:30:35
104375	2015-12-15 06:50:51.873	2015-12-11 20:46:08
317143	2019-09-22 07:10:41.647	2019-09-09 09:37:07
102182	2020-03-09 12:35:20.500	2020-02-23 00:06:03
240896	2018-03-20 07:52:46.330	2018-02-28 18:05:51
48606	2016-03-16 13:22:12.077	2016-03-09 21:22:09
290792	2019-09-18 05:56:50.580	2019-09-01 09:54:29
297723	2020-11-01 12:45:19.347	2020-10-06 13:38:12
87922	2019-07-18 14:26:25.137	2019-06-19 16:18:07
249211	2020-07-21 08:24:28.620	2020-07-03 16:57:03
356078	2020-01-30 12:55:23.060	2020-01-03 12:38:17

Conclusion

Able to minimize the test & validation loss

Questions

Link to the notebook

https://www.kaggle.com/soujanyag/stackoverflow-userlastaccessdateprediction


THANK YOU
ASANTE
ASANTE
OBRIGADO
MATUR NUWUN
NIRRINGRAZZJAK
GRAZIE
RAIBH MAITH AGAT
KIA ORA
MULTUMESC
MOCHCHAKKERAM
OBRIGADO
SPASIBO
SALAMAT
RAIBH MAITH AGAT
MOCHCHAKKERAM
TERMA KASIH
MAMAVA
MATONDO
GRAZIE
KIITOS
SPASIBO
MERCI
KIITOS
MOCHCHAKKERAM
CHOKRANE
MATONDO
UA TSAUG RAU KOJ
MERCI
GRAZIE
OBRIGADO
CAM ON BAN
OBRIGADO
MAAKE
GRAZIE
DANKON
ARIGATO
SPASIBO
KIITOS
RAIBH MAITH AGAT
MERCI
MULTUMESC
CHOKRANE
MOCHCHAKKERAM
MATONDO
MULTUMESC
CHOKRANE
MAAKE
VINAKA
JUSPAXAR
MOCHCHAKKERAM
ASANTE
MOCHCHAKKERAM
ARIGATO
KIITOS
DANKON
MULTUMESC
NIRRINGRAZZJAK
MULTUMESC
UA TSAUG RAU KOJ
OBRIGADO
DANK JE
MOCHCHAKKERAM
DANKON
SPASIBO
MOCHCHAKKERAM
TERMA KASIH
WELALIN
KIA ORA
SALAMAT
NIRRINGRAZZJAK
MATUR NUWUN
MOCHCHAKKERAM