Marketing Email Subject Evaluation

Problem

Having gathered data on marketing emails, Element 451 wants to find whether there is a relationship between marketing email subject line and email open rate.

Aim

Find out the **features** of the **Email Subject** that result in a **higher opening rate**

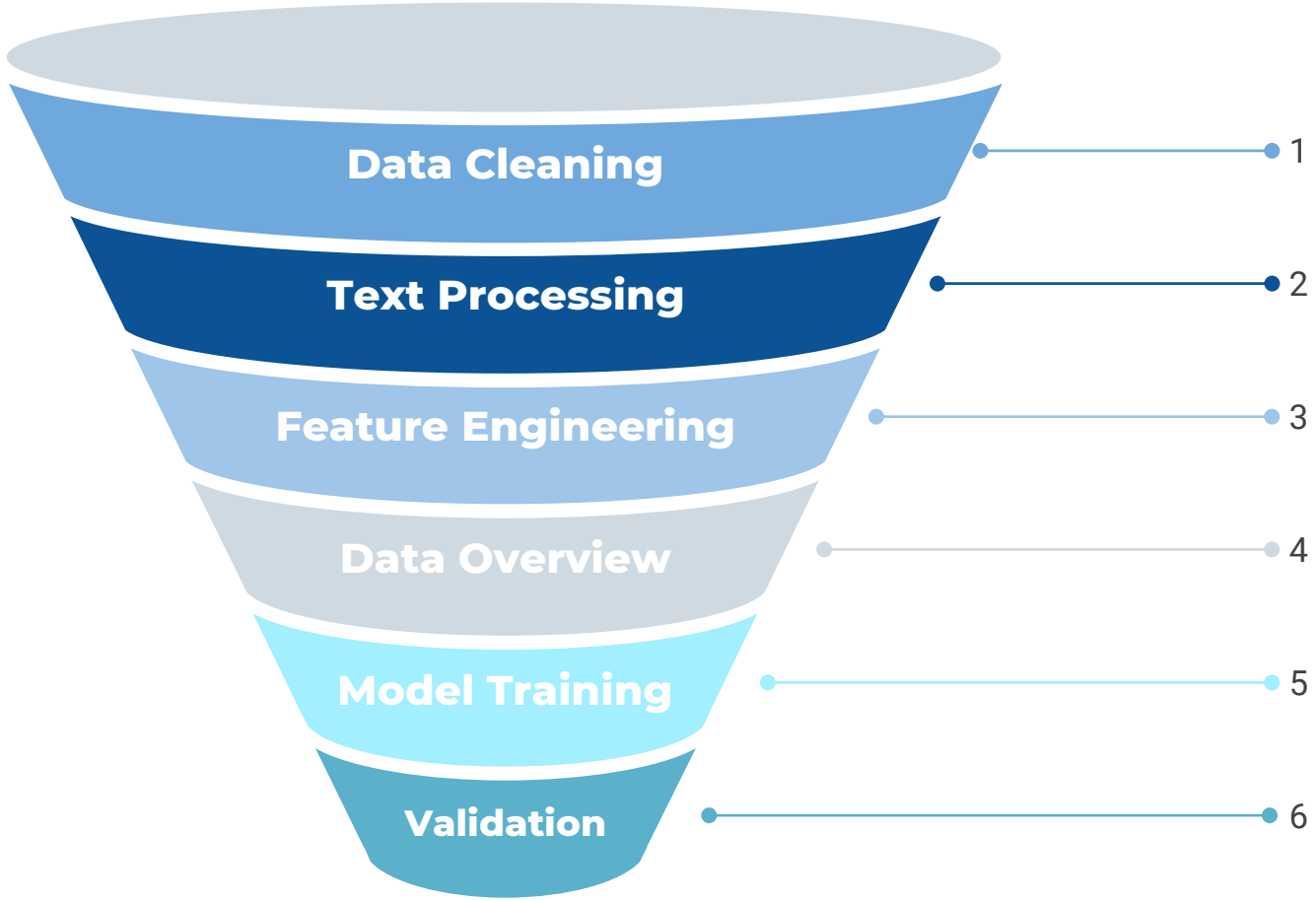
Brief

“The team created a pipeline which can clean, update useful features and predict using autotuned ML models.

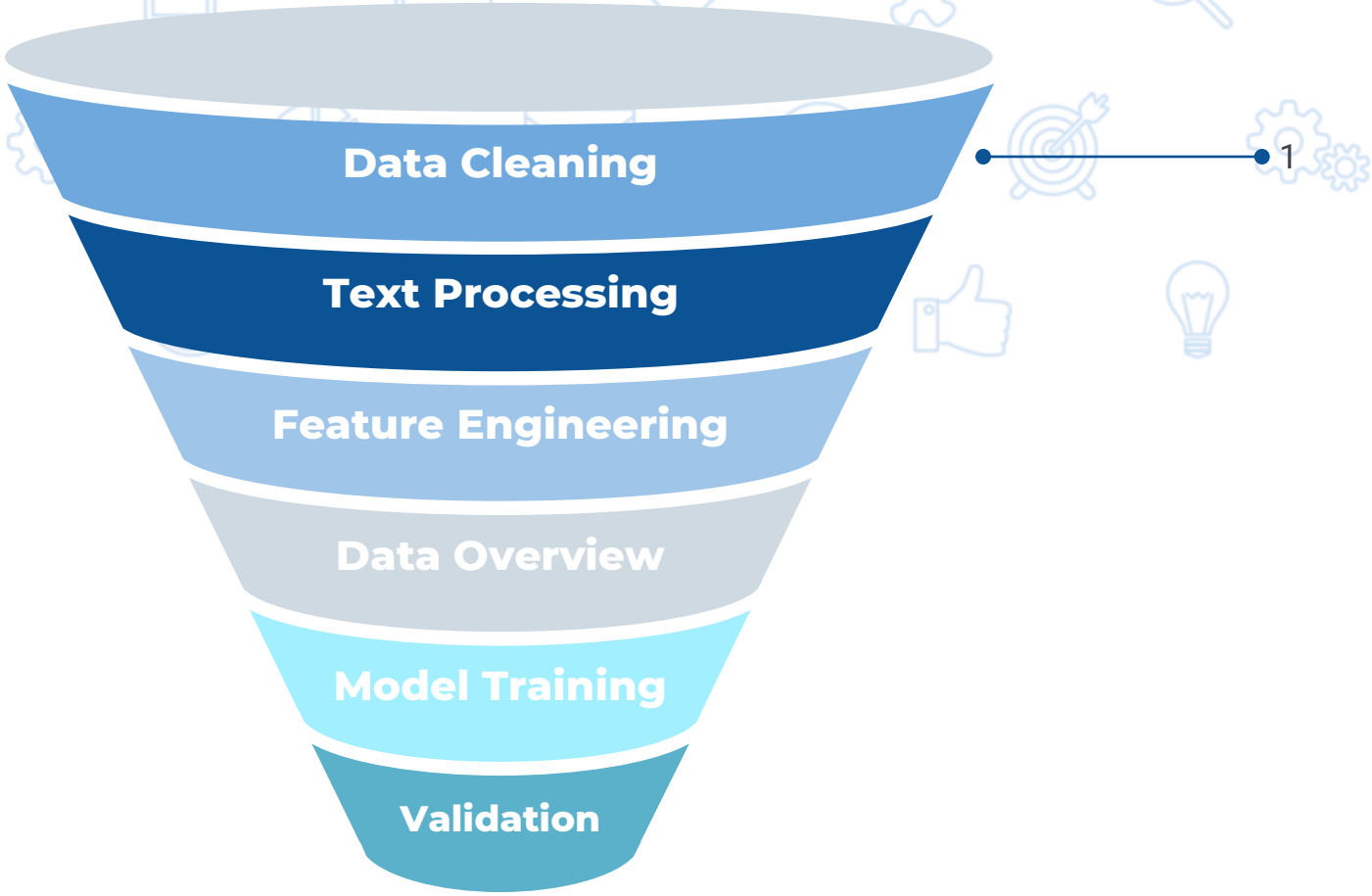
The final model can provide an approximate impact of the email, after you input email subject, number of recipients and the sending date”

0

Our Pipeline



1



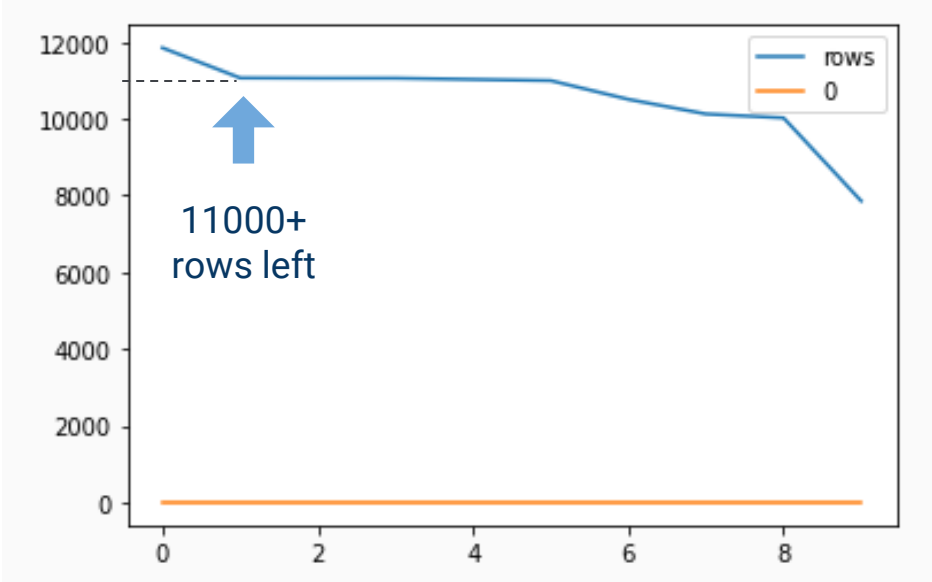
Data Cleaning

1

Data Cleaning

Step 1 / 4

Remove NAs



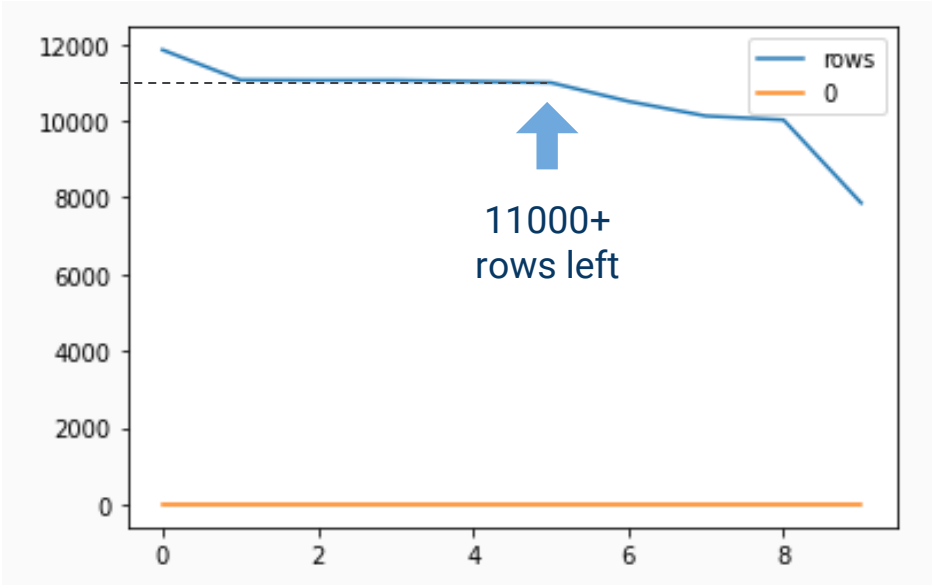
1

Data Cleaning

Step 2 / 4

Logic Check

- “opens” should be **greater** than “unsubscribed”
- “total” count should be **greater** than “unique” count
- “unique rate” should be **less** than 1



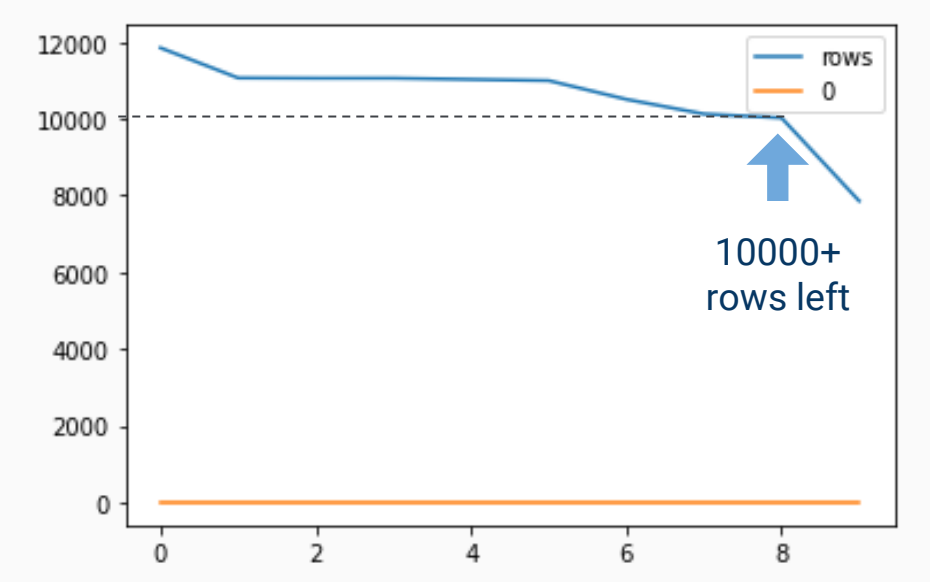
1

Data Cleaning

Step 3 / 4

Removing Precision Error

- CTV = unique clicked / unique opens
- unique open rate = unique opens / total delivered
- open rate = total opens / total delivered



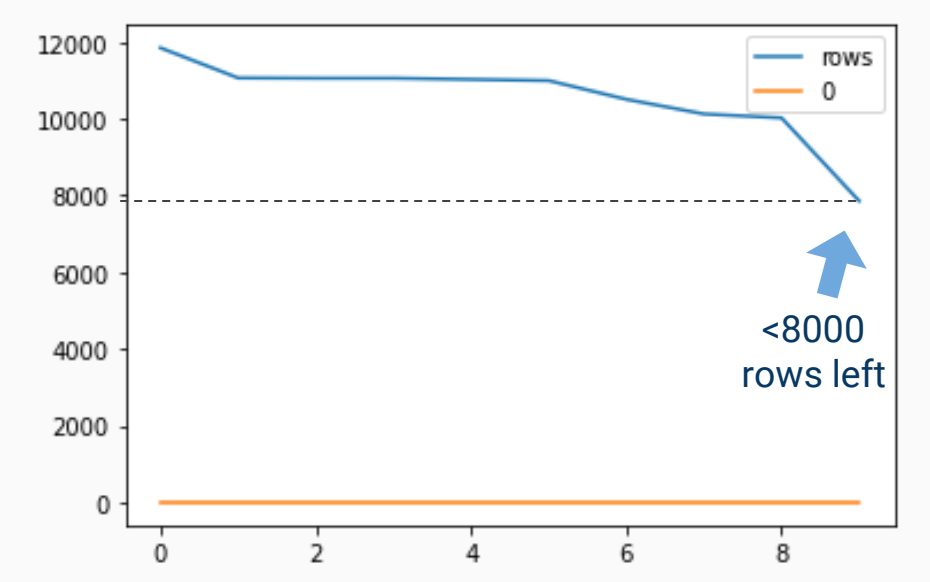
1

Data Cleaning

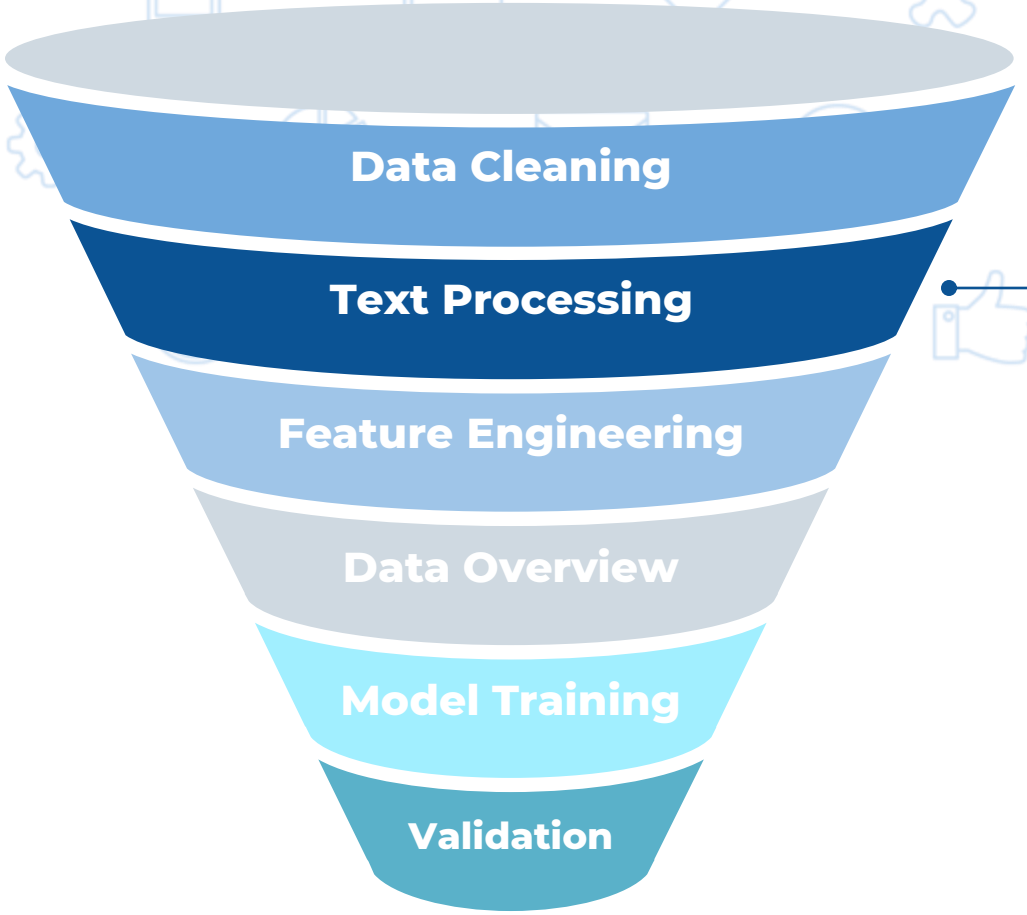
Step 4 / 4

Remove Test Case

Deleting rows that total delivered number <=100



2



2

TEXT PROCESSING

2

TEXT PROCESSING

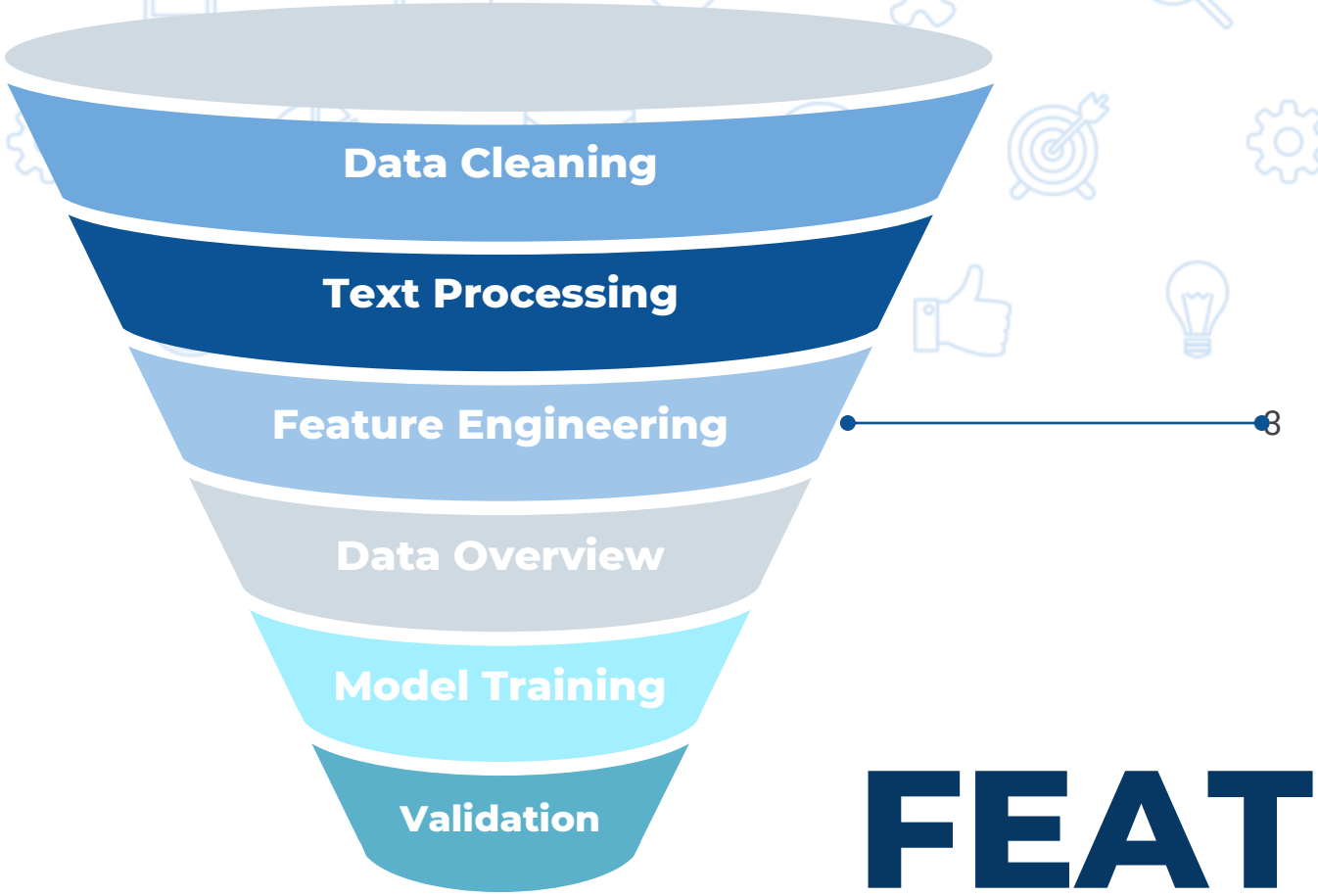
Focusing on the Email Subject to understand the meaning

- Removed Punctuations
- Extracted Date

- Extracted Username
- Extracted emojis
- Turned them into separate tokens

- Stemming
- Lemmatizing
  - **Clubbing words with similar meaning**

3



FEATURE ENGINEERING

3

Feature from Text

Generating Features from Email Subject and Parameters

Count

Word / Char/ Emoji
Emoji Type
Punctuation
User Token

NER

Total Count
Organization
Location
Person
Facility

Date

Weekday
Month
Day

WORDS

Stemming
Lemmatizing

3

Feature from Numbers

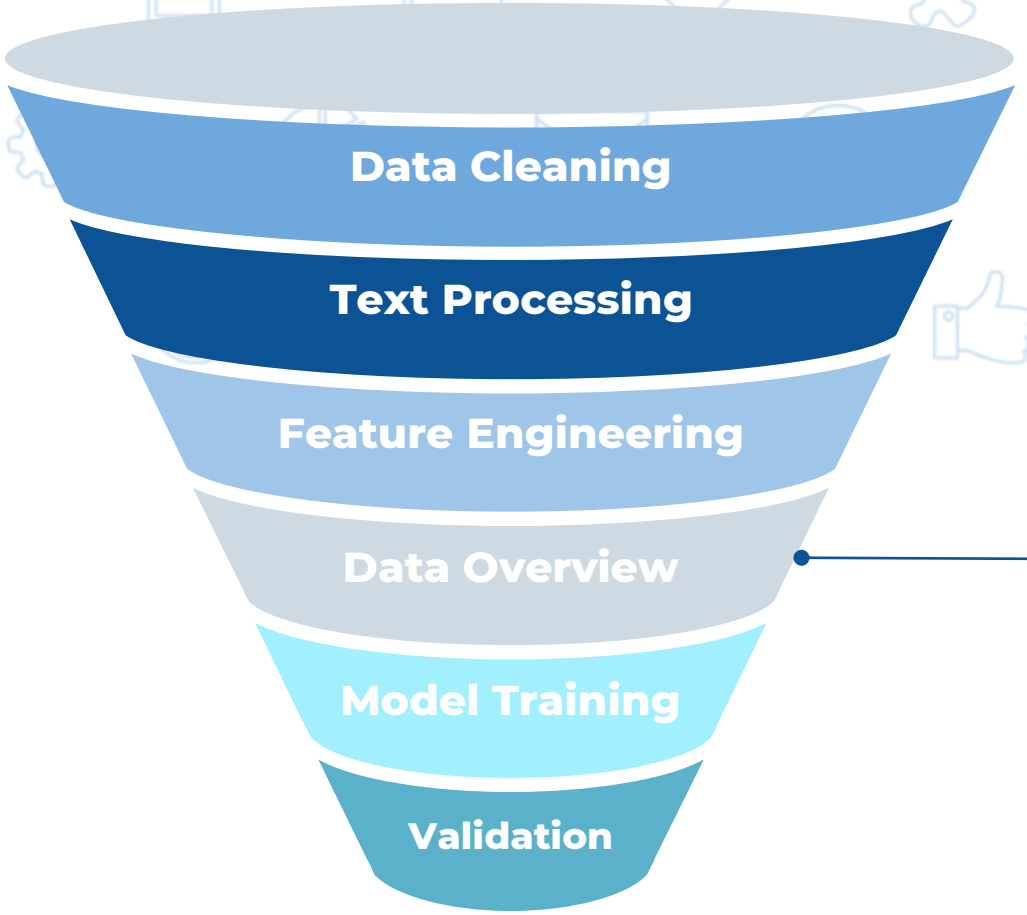
Feature Transformed

unsubscribe
click
delivered
open

Fixed feature

Successful boolean
unique open rate without
subscribe

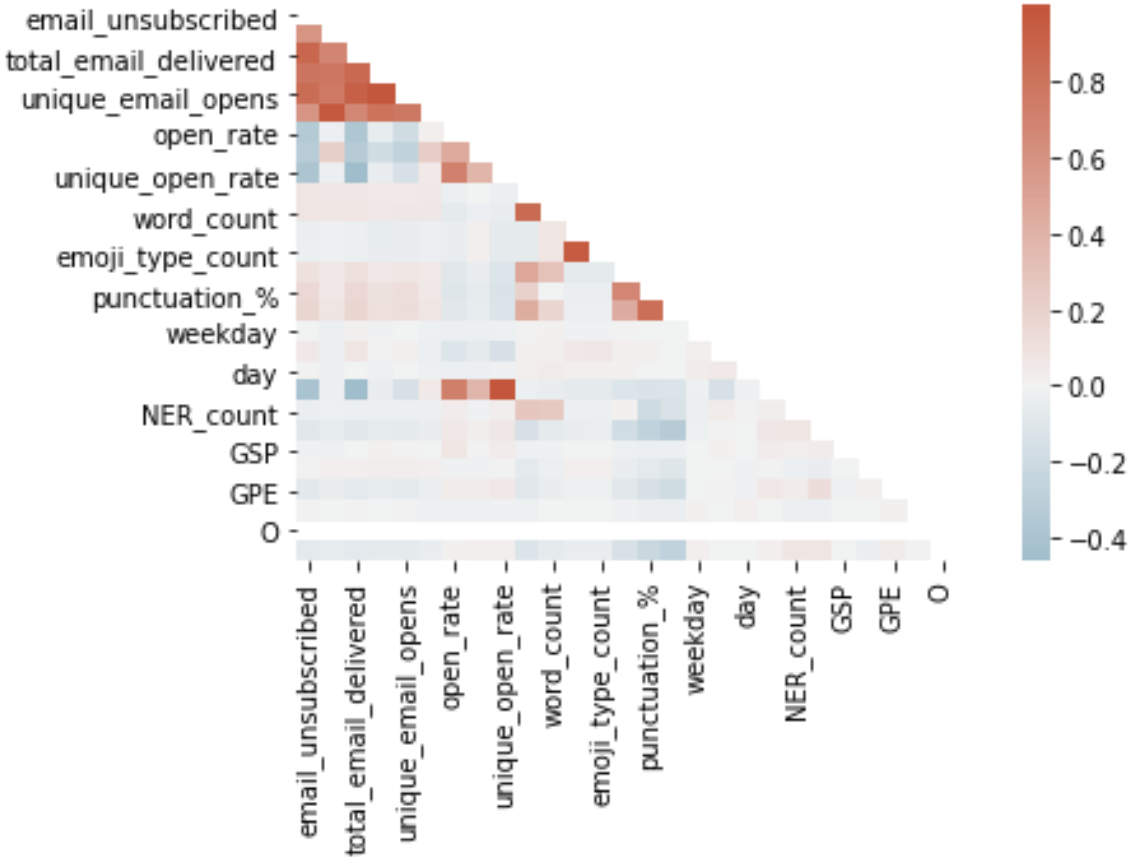
4



DATA OVERVIEW

4

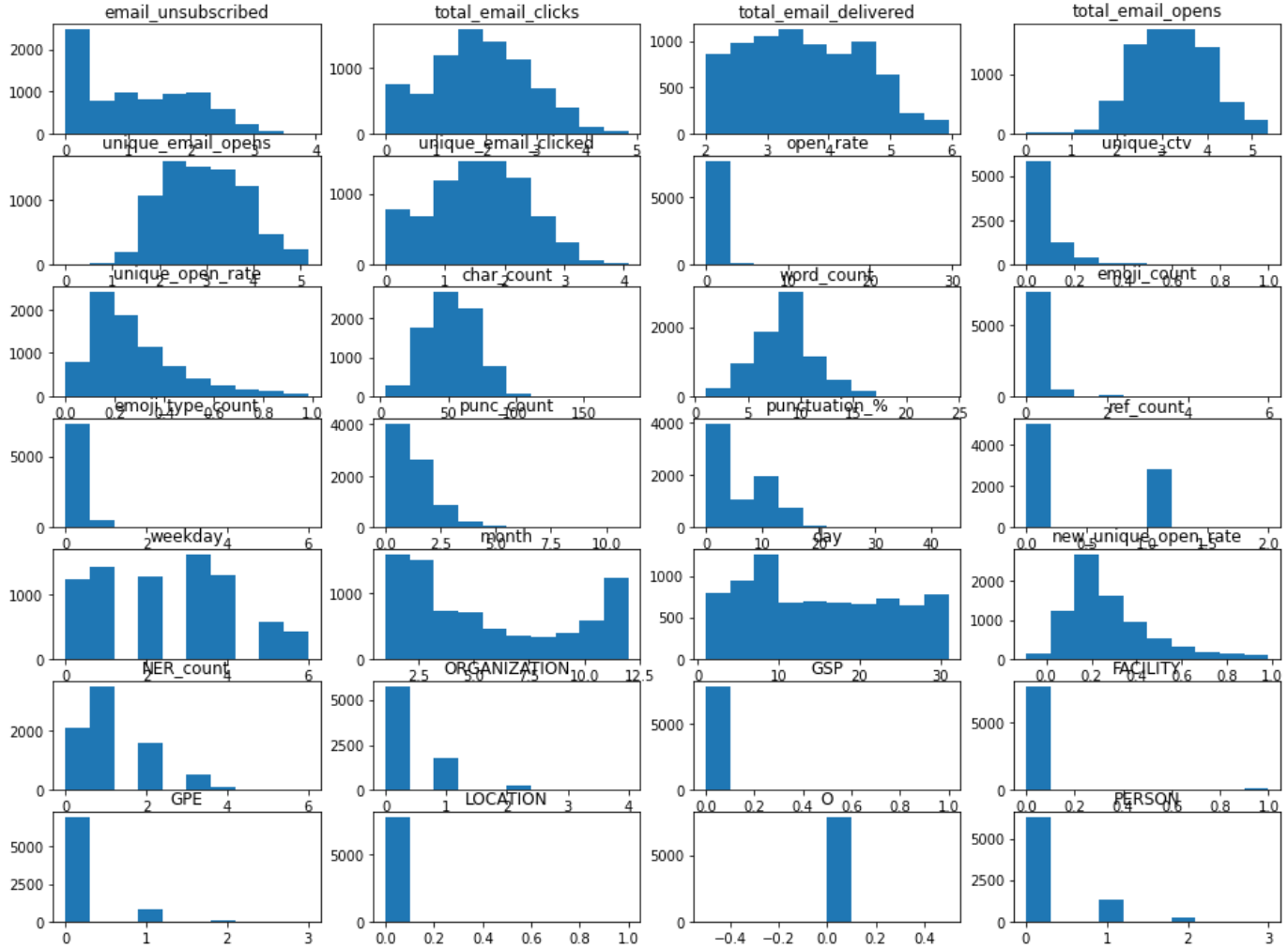
Correlation plot



Day is highly correlated with Unique open rate

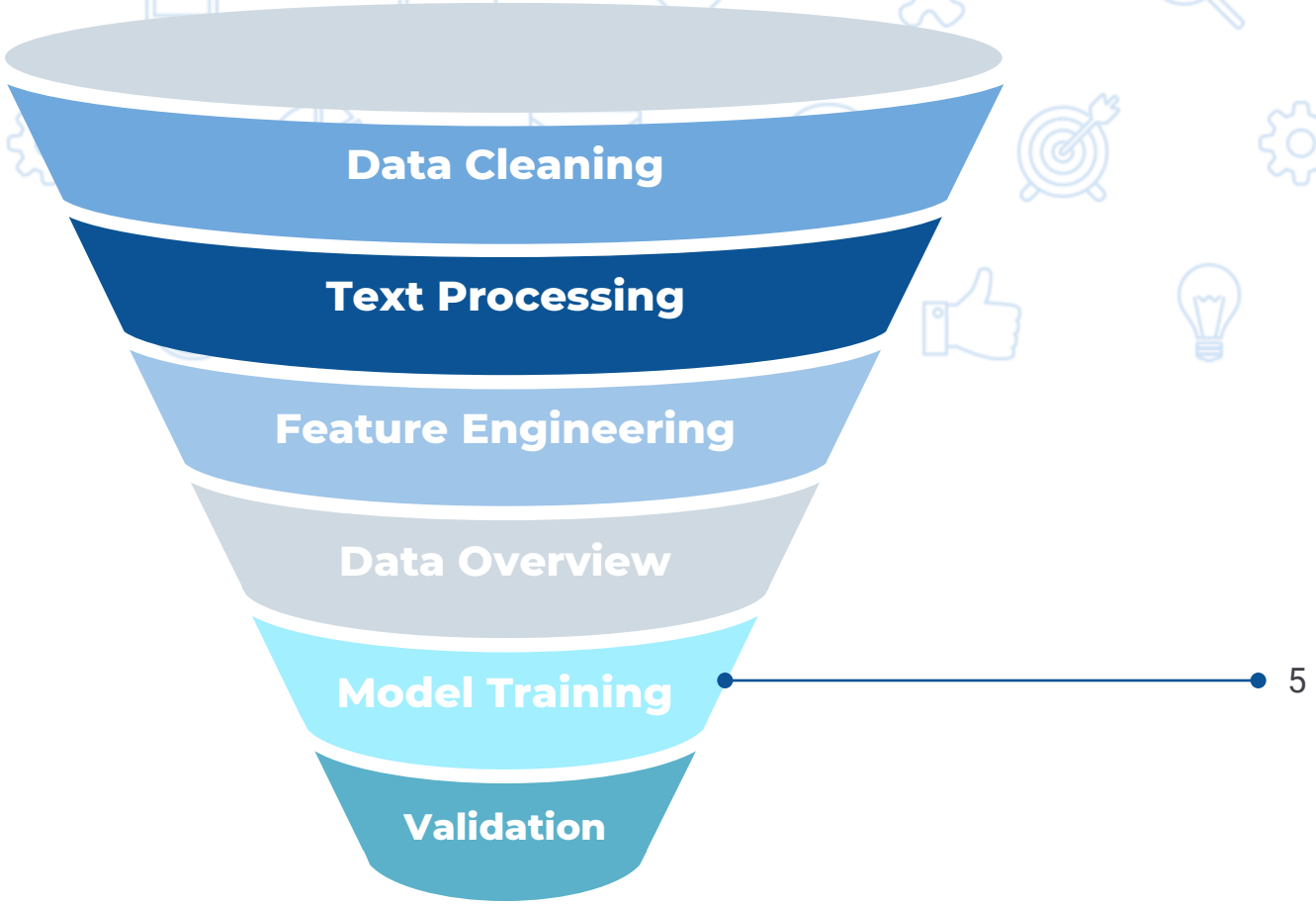
4

Histogram



After transformation, no feature is highly skewed but most are not normally distributed

5

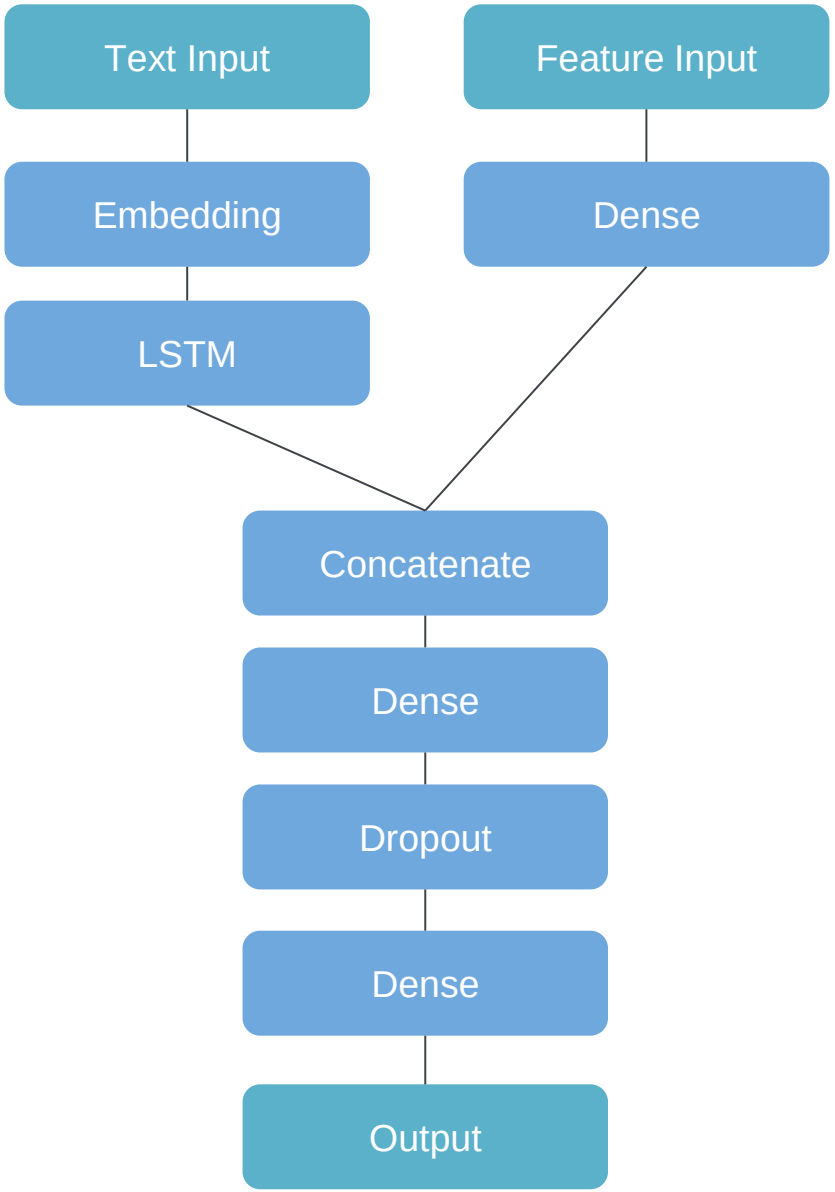


MODEL TRAINING

5

Model setup

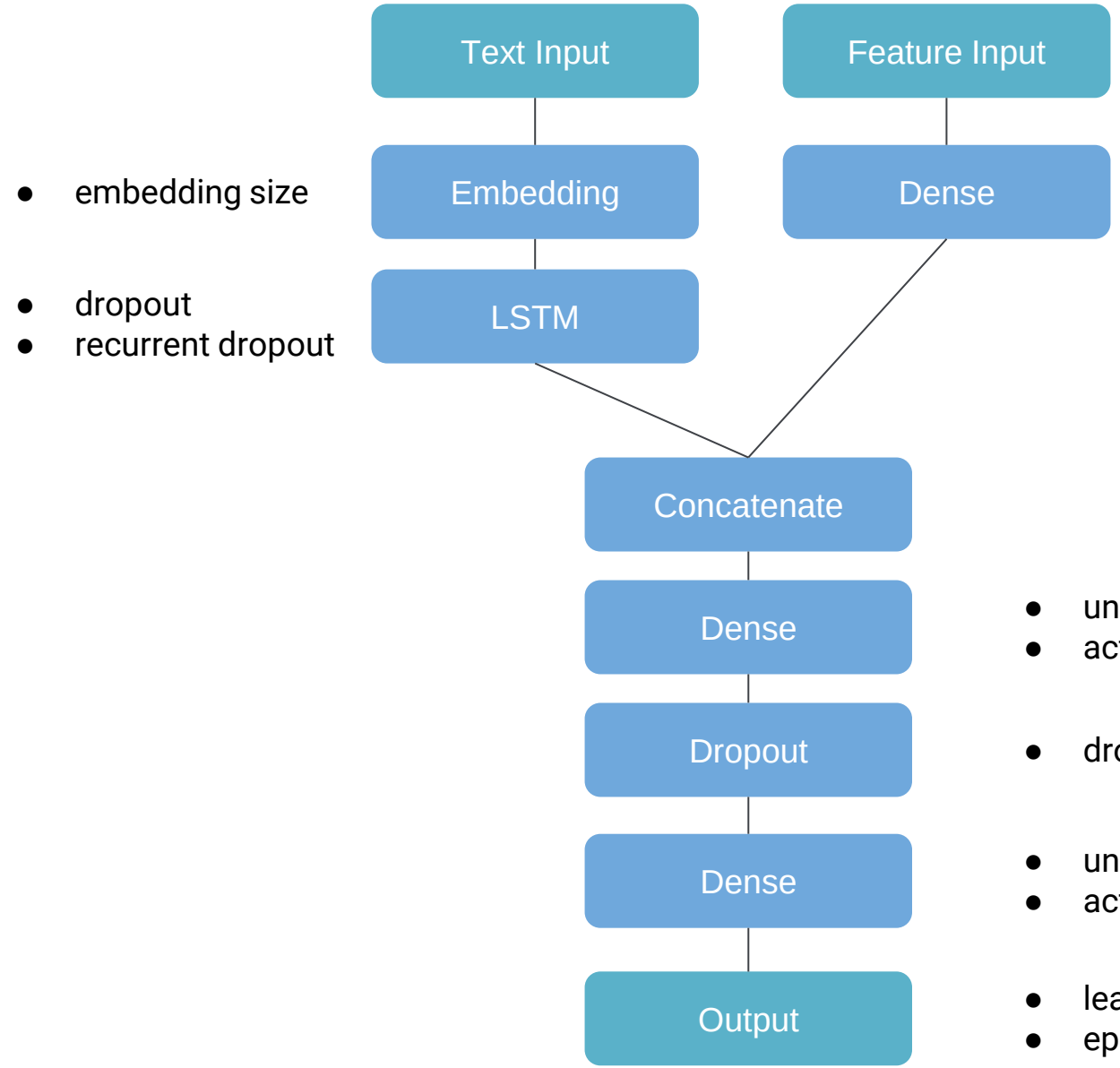
Layer Overview of the ML Model



5

Auto Tuning

HyperParameters for Auto Tuning Model



- embedding size

- dropout
- recurrent dropout

- units- nodes in output layer
- activation function

- units
- activation function

- dropout

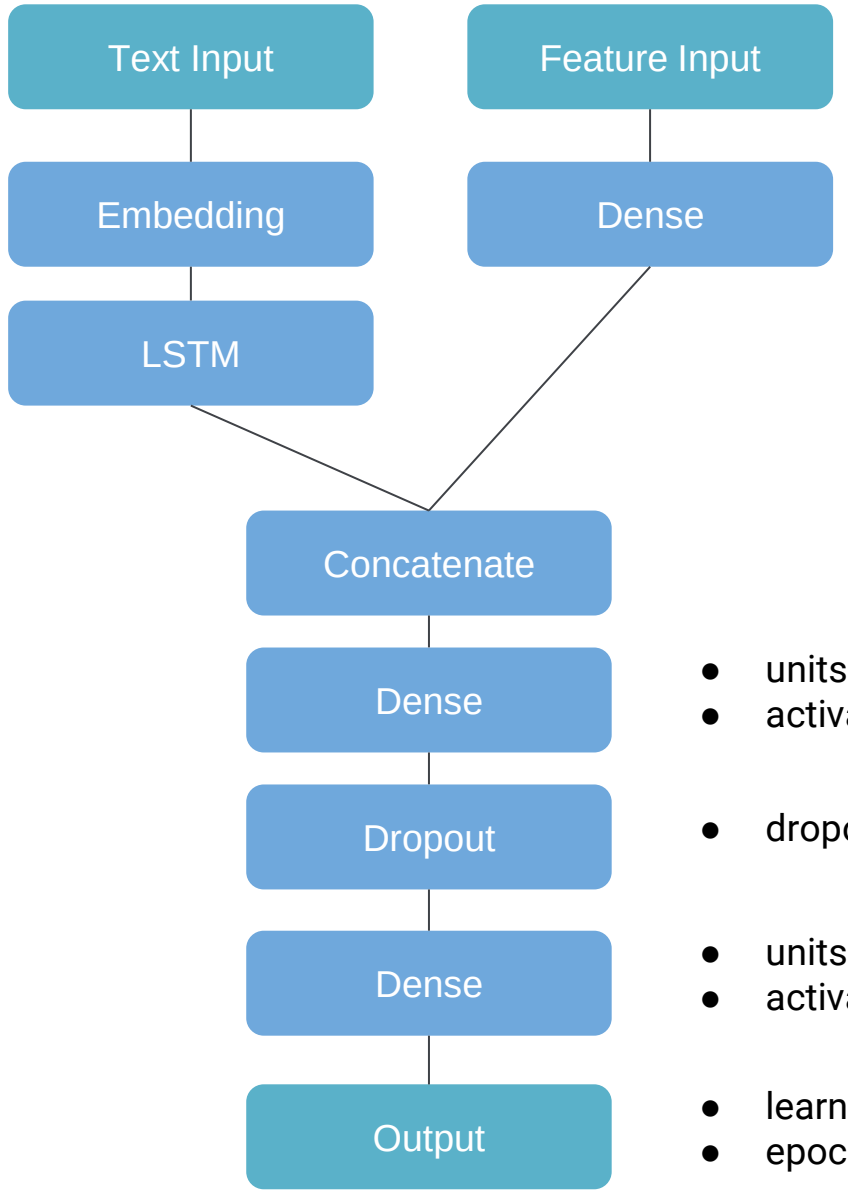
- units
- activation function

- learning rate
- epochs

5

Optimized Parameter Auto Tuning

Optimized Parameters for open rate prediction



- embedding size = 50

- dropout=0.6
- recurrent dropout=0.4

- units=48
- activation function=relu

- units=64
- activation function=softmax

- dropout=0.4

- units=48
- activation function=selu

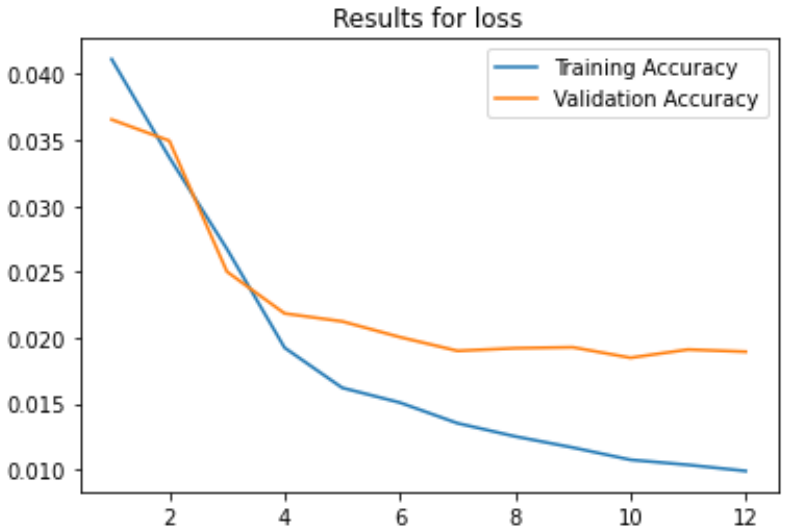
- learning rate=0.005
- epochs=12

5

Text Processing

Before Autotuning

After Autotuning

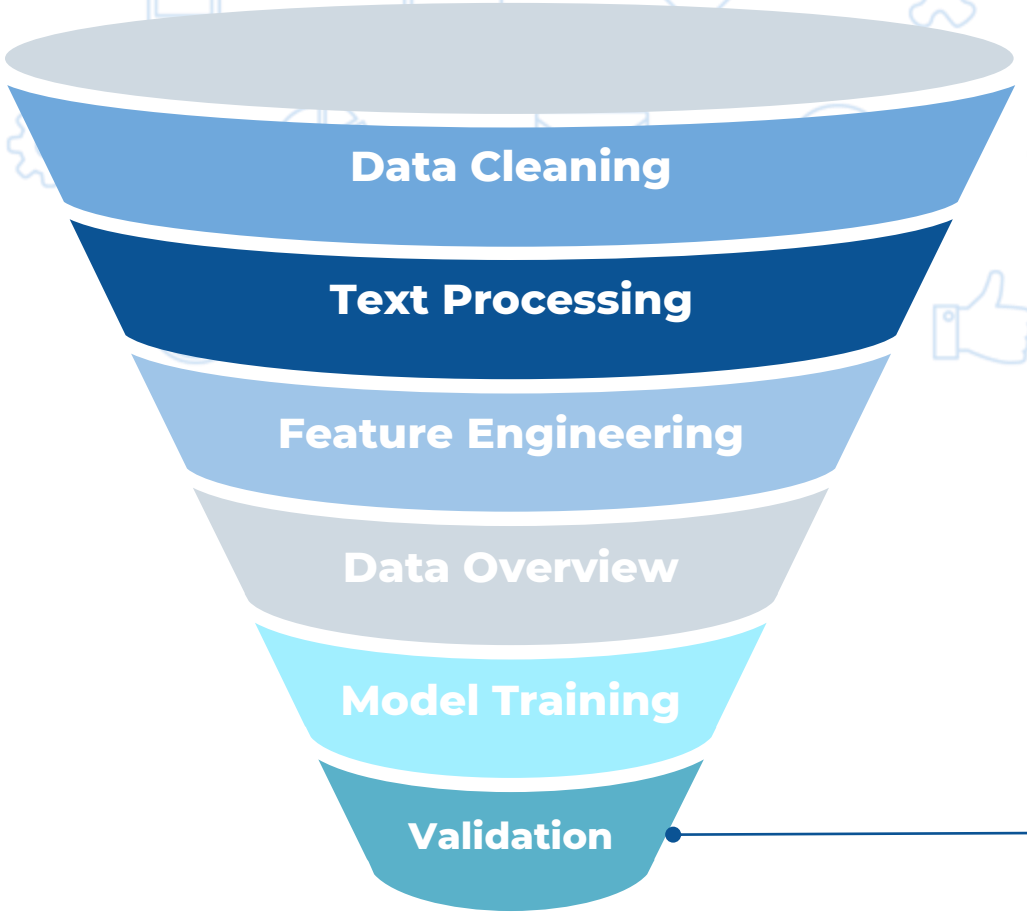


Loss :

0.1 -> 0.02

(-80%)

6

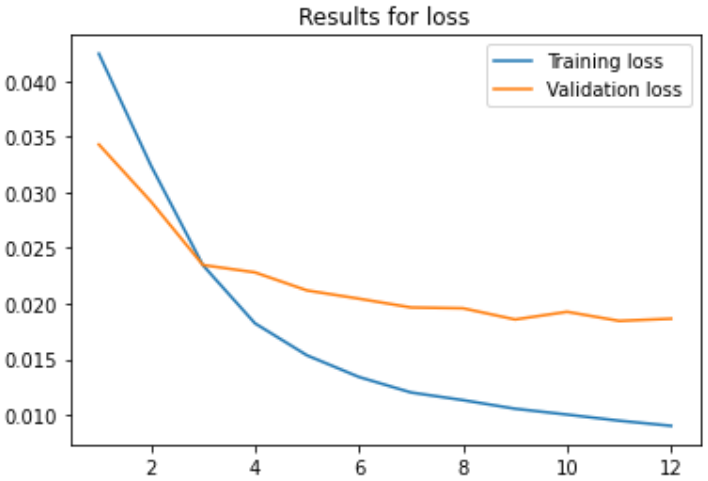


6

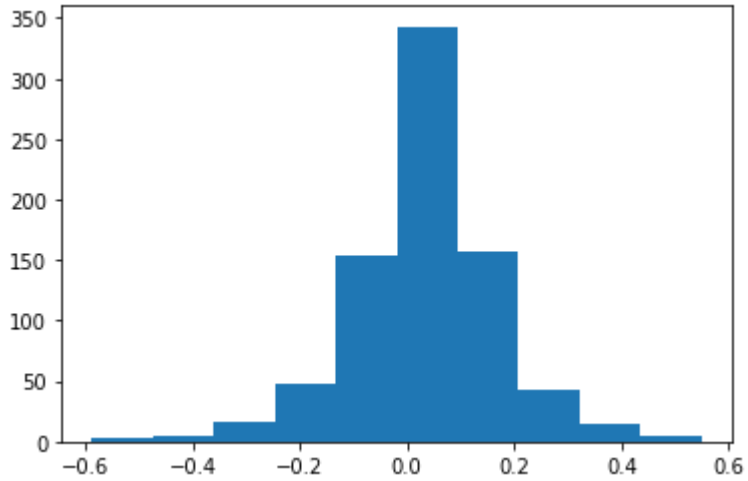
VALIDATION

6

Text Processing



Loss History

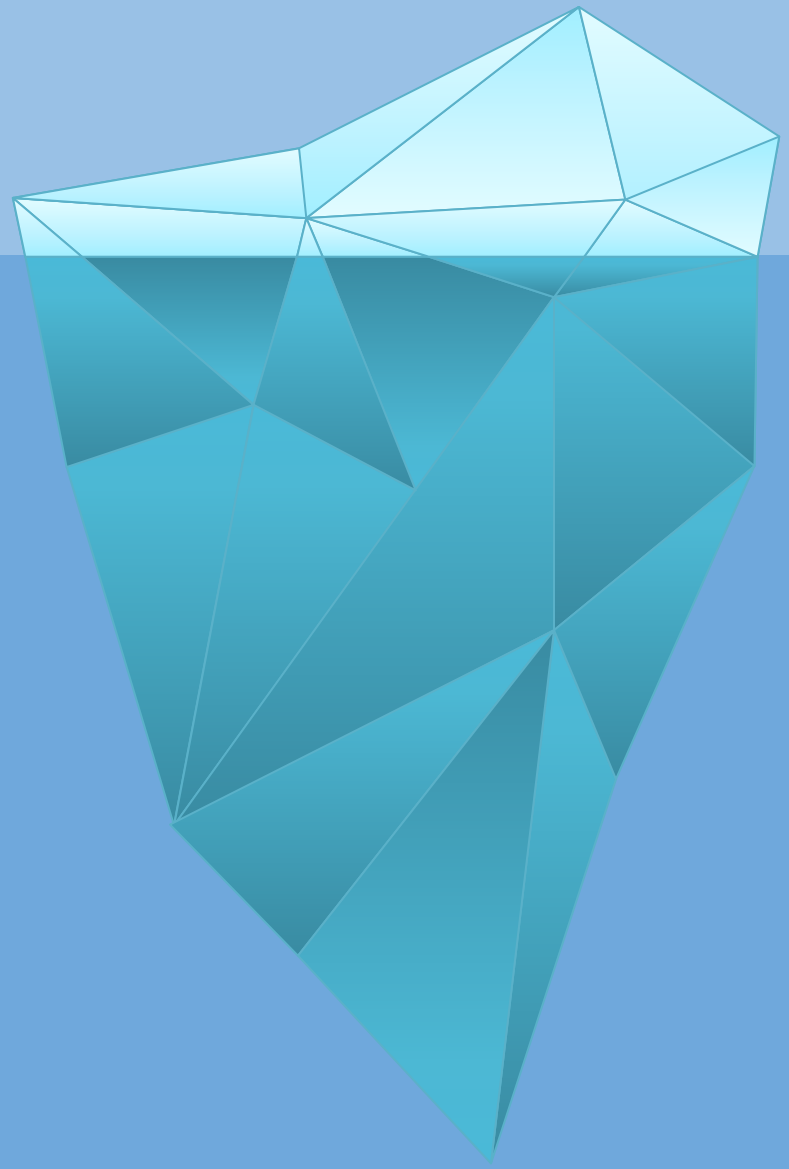


Deviation histogram

Validation loss: 0.189
Deviation : 0.0185

7

From the deviation plot we can conclude 75% prediction is fall within +-0.2 of the real situation



Conclusion

Our model measured by error. Since we are unaware of the standard of the loss, we do not know how good our model is.

To Explore

Bigger dataset can get us a better accuracy as there will be bigger data to test and train

Defining a successful email will allow us to create a binary column which can help us run classifiers