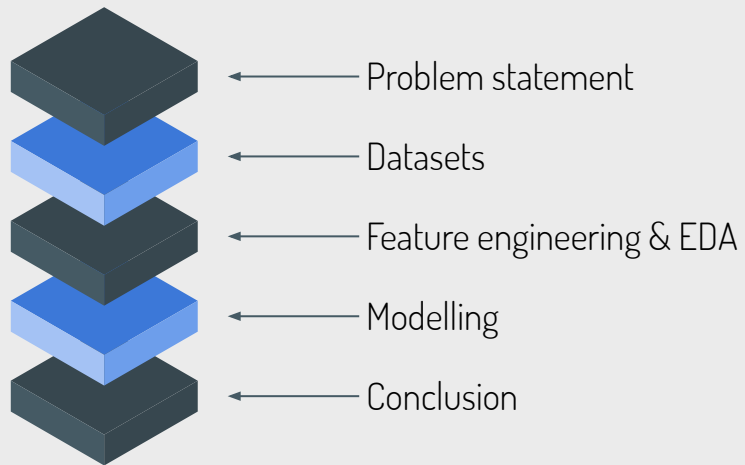


TRUMP, TWEETS AND THE STOCK MARKET

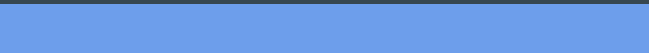


DSI Capstone Project
By: Russell Quah

TABLE OF CONTENTS



Problem statement



PROBLEM STATEMENT

The US market team at a local bank has seen literature on models that are able to predict market movement based on tweets by Donald Trump:

- JP Morgan creating a 'Volfeffe index to track tweets vs bond market
- Bank of America has stated that on the days when President Trump tweets a lot, the stock market falls

They have tasked the data science team to build a classification model using Natural Language Processing to predict if Donald Trump's tweets are market moving.



PROBLEM STATEMENT

- Logistic Regression
- XGBoost
- Long Short Term Memory Neural Network
- Evaluate the models based on:
 - accuracy (% predictions the model gets correct, both a significant movement and a non-significant movement)
 - precision (% predicted significant movement when it is actually significant movement)
 - sensitivity (% predicted significant movement out of all correct predictions)
- choose the best performing model to test it on the holdout csv



Datasets



@REALDONALDTRUMP TWEETS

Kaggle dataset

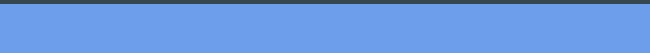
- 04 May '09 to 17 June '20
- 43352 tweets
- 8 columns
- Removed unnecessary features (id, link, mentions and hashtags)

S&P500 RETURNS

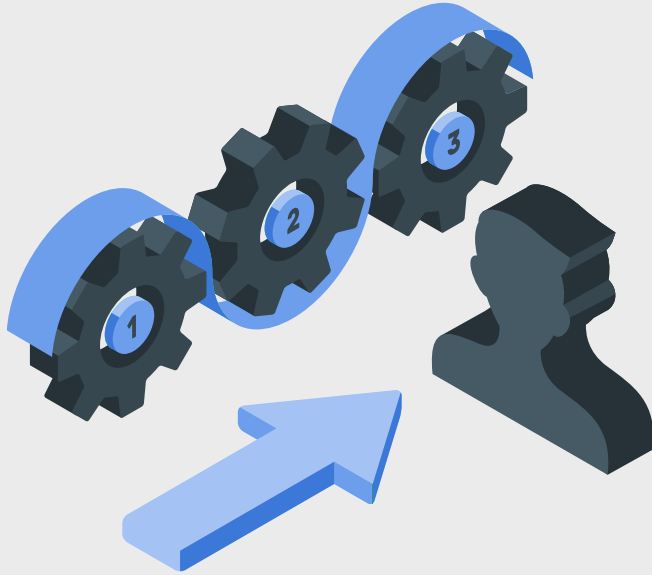
Yahoo finance API

- May '09 to June '20
- 2805 days
- 8 columns

Feature engineering & EDA



FEATURE ENGINEERING ON TWEETS



CYCLICAL DATA

Date was split into cyclical features:

- Month, Day, Hour, Minute
- Sin and Cosine

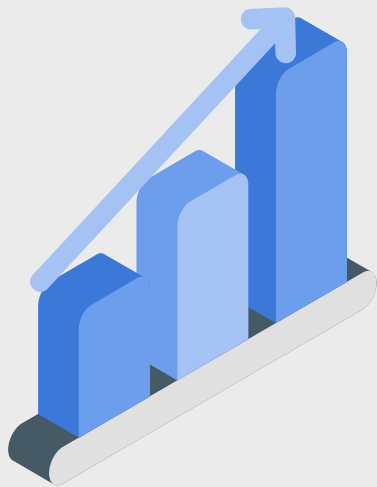
VADER SENTIMENT ANALYSIS

- Specifically designed to handle social media
- Compound = Positive + Neutral - Negative

LOUGHRAN MCDONALD FINANCIAL SENTIMENT ANALYSIS

- Dictionary
- Counts the number of times words appear in 9 different categories

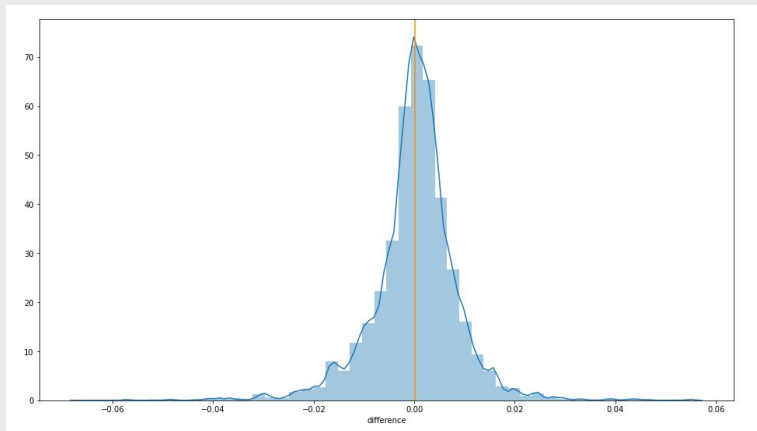
FEATURE ENGINEERING ON S&P500



INTRADAY DIFFERENCE

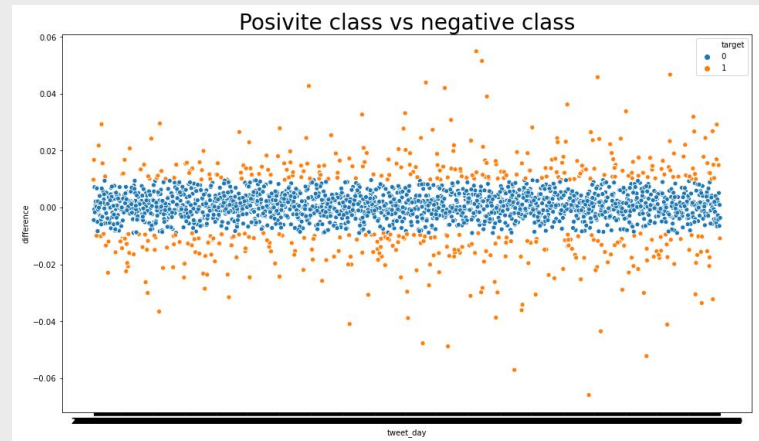
Target variable: Opening price - closing price

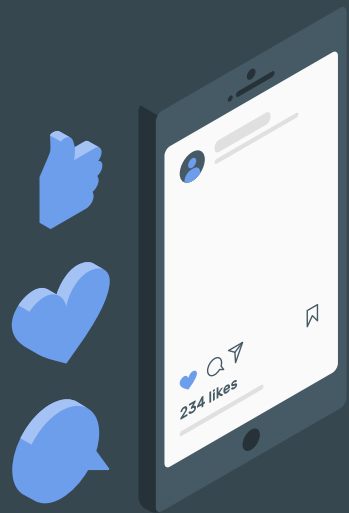
DEFINING THE POSITIVE CLASS



Positive Class

- 1 standard deviation away from the mean





TOTAL NUMBER OF
TWEETS SINCE MAY'09

34,595



TOTAL NUMBER OF TWEETS AS POTUS:

9,644



7.75 AVERAGE* TWEETS A DAY AS POTUS



+ CLASS 9.00 AVERAGE* TWEETS A DAY AS POTUS

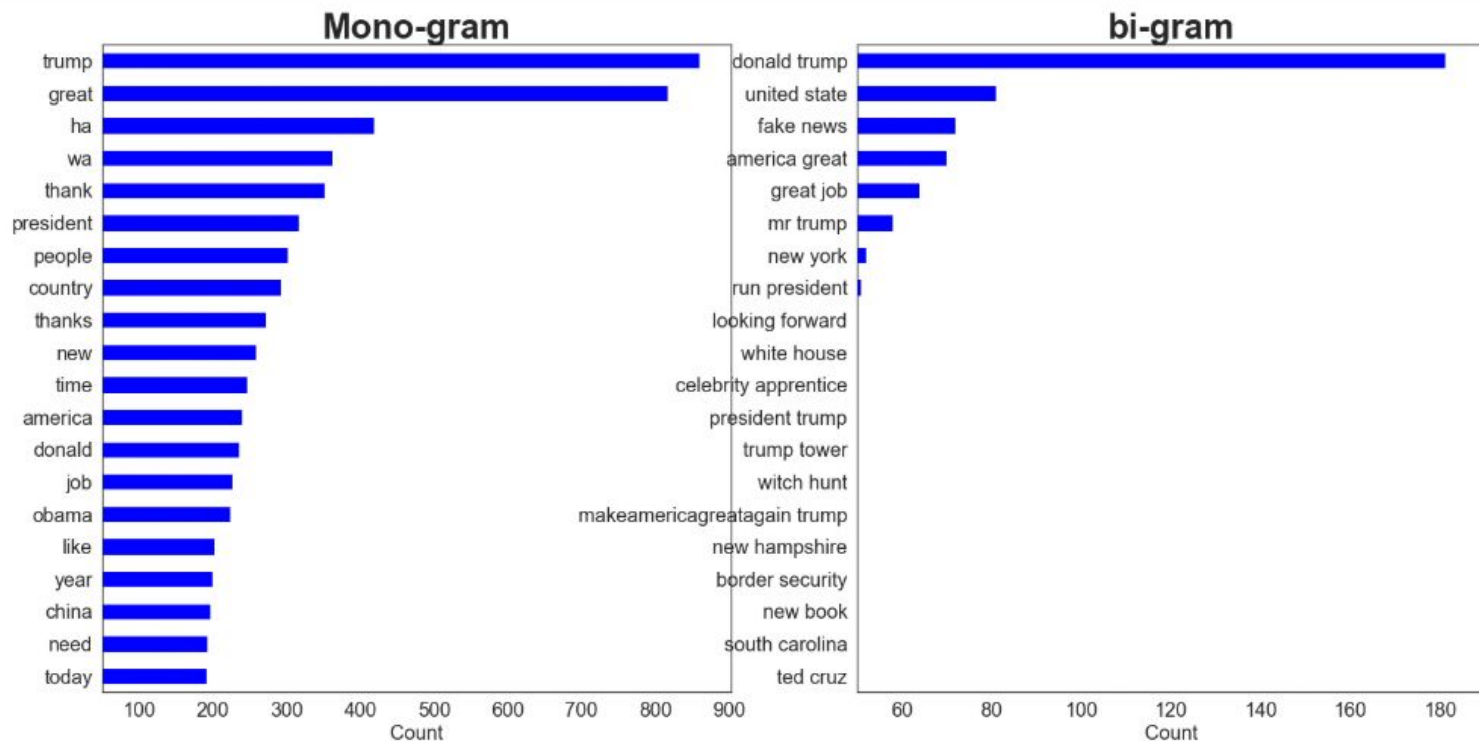
- CLASS 7.77 AVERAGE* TWEETS A DAY AS POTUS

[illegible]

© 2010 Blackwell Publishing Ltd *Journal of Internal Medicine* 267: 257–265

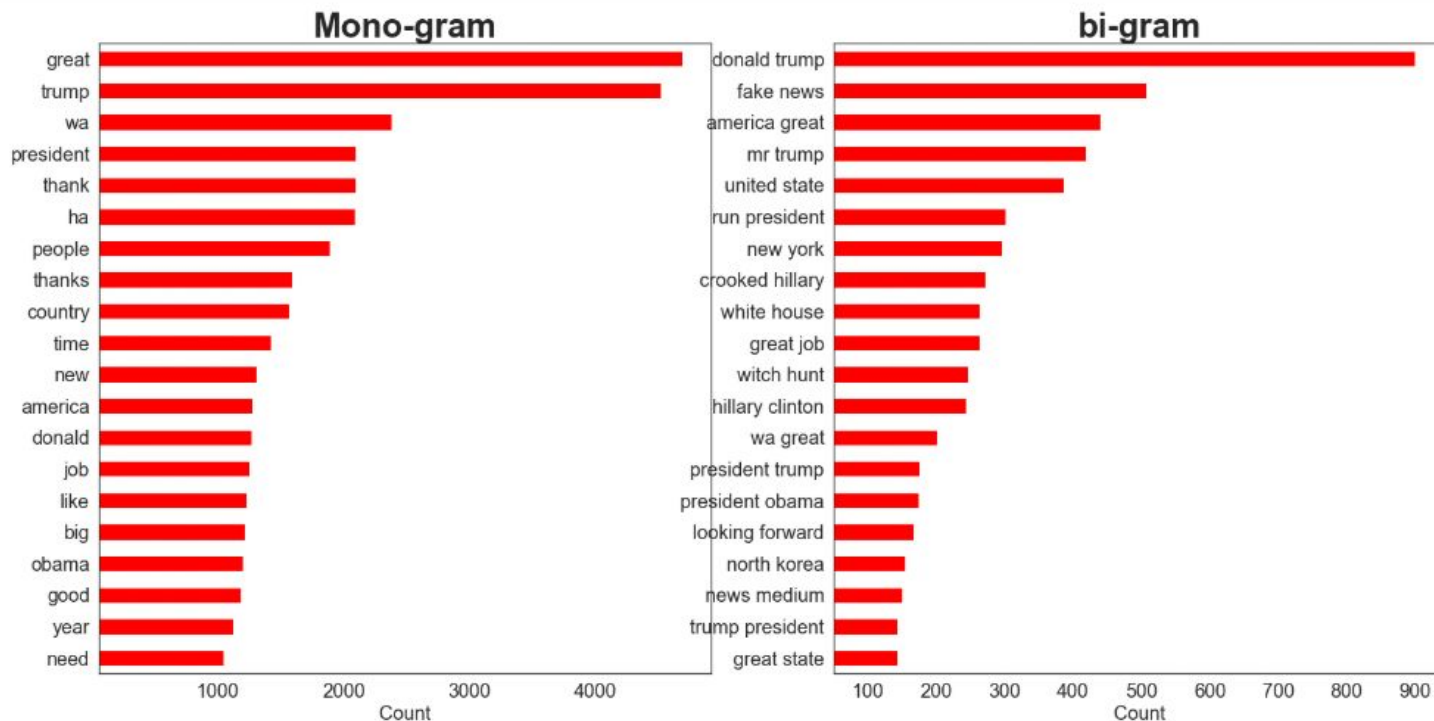
POSITIVE CLASS MOST USED WORDS

```
In [107]: #positive class  
barplot_cvec(1, 'Positive Class', ['Mono-gram', 'bi-gram'], 'blue', 50)
```



NEGATIVE CLASS MOST USED WORDS

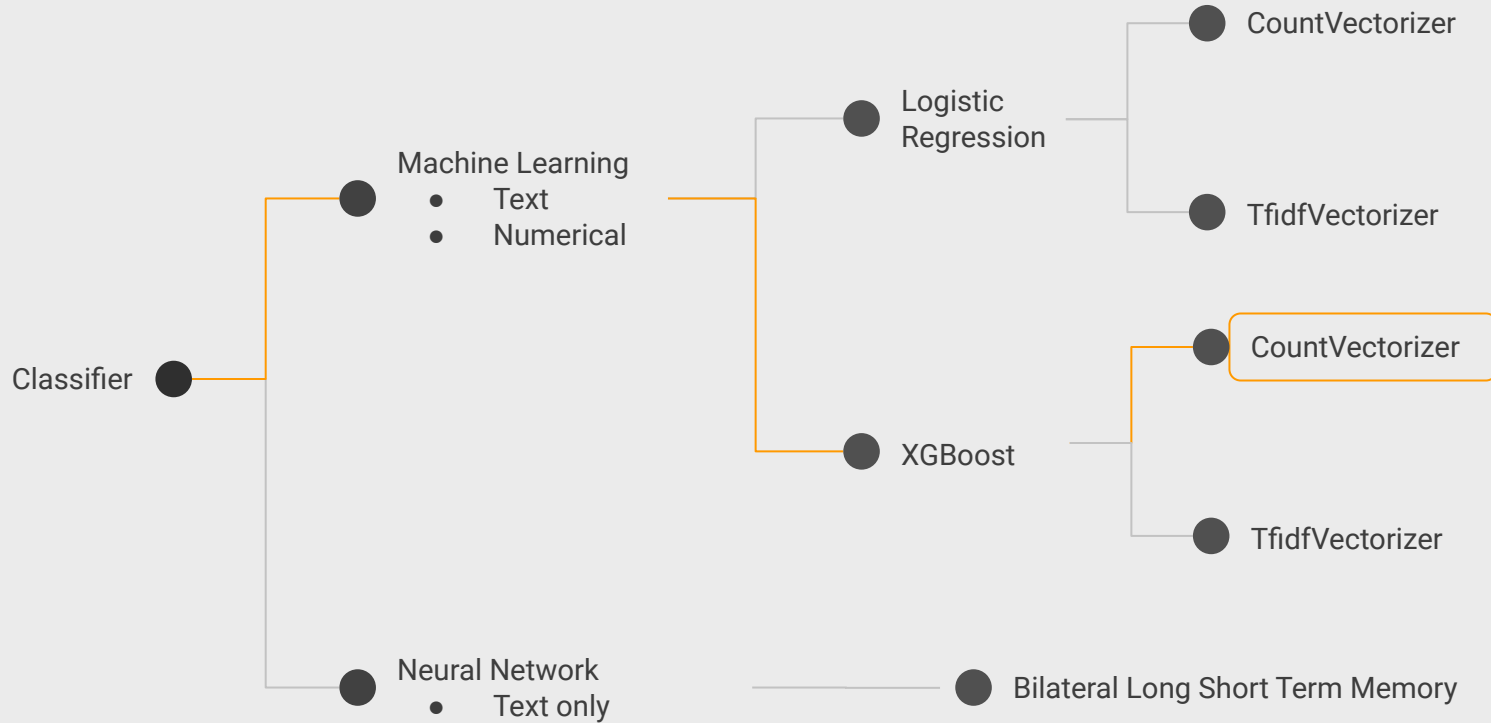
```
In [108]: #negative class  
barplot_cvec(0, 'Negative class', ['Mono-gram', 'bi-gram'], 'red', 50)
```



Modelling



MODELLING



MODELLING

Pipeline 1 - Gridsearch

Countvectorizer



XGBoost

+

Pipeline 2 - Gridsearch

Smote



Standard
Scaler



XGBoost

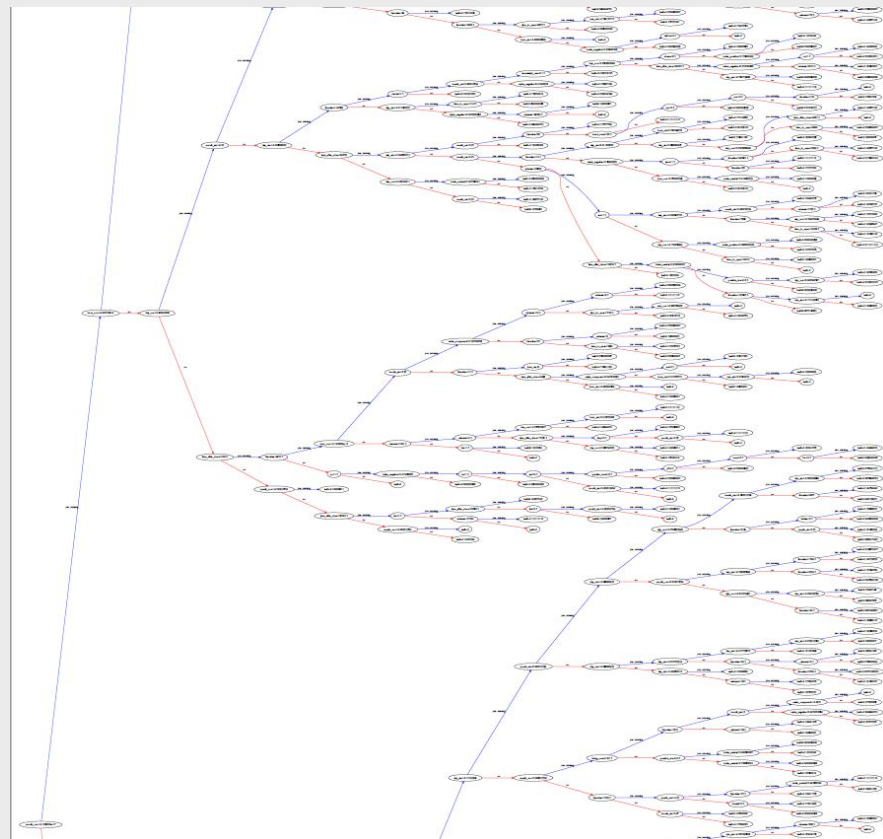
MODEL PERFORMANCE

	gs1 (cvec->Ls)	gs2 (tfidf->Ls)	gs3 (cvec->XGB)	gs4 (tfidf->XGB)	gs6 (Bilateral L STM)	best_model (cvecc->xgboost)
accuracy(train)	0.661	0.600	0.999	0.992	0.877	0.998
accuracy(val)	0.665	0.591	0.905	0.895	0.806	0.906
precision	0.331	0.455	0.462	0.401	0.087	0.469
sensitivity	0.180	0.177	0.854	0.830	0.200	0.857
F1	0.233	0.255	0.600	0.541	0.122	0.606
roc_auc	0.547	0.552	0.925	0.908	0.512	0.925

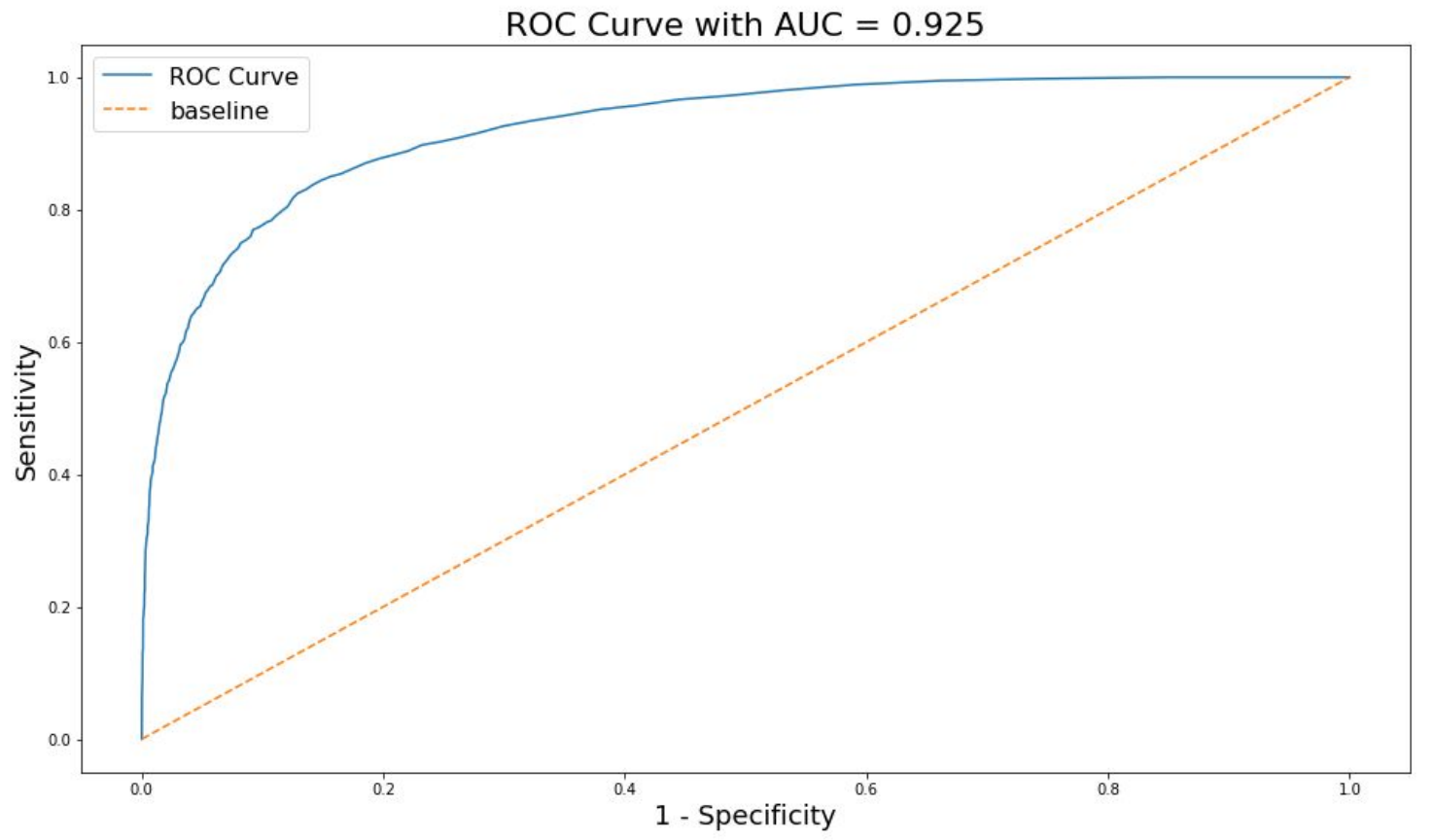
Baseline:

- 0.846 of negative class

MAP OF XGBOOST TREE



ROC AUC



Conclusion



CONCLUSION

Best model is CVEC into XGBoost,

- Accuracy of 0.906
- Precision of 0.469
- Sensitivity of 0.857

Limitations

- Model falls short on precision
- It is not a parametric model

Recommendation for the US market team:

- On days of positive class, Trump tends to tweet more on average
- The model is able to accurately predict when his tweets will not move the market





THANKS

Does anyone have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik** and illustrations by **Stories**

Please keep this slide for attribution.