NLP project to classify Subreddits for ESPN

… Cricket Vs Soccer

Kathirvel Kumararaja
January 31, 2020

Kathirvel Kumararaja

Vice President - Operations,
DevJee Inc.

Credentials:

- Data Scientist with 20 + years of IT consulting experience.
- Holds an Engineering degree and and MBA.
- Experience in handling data analysis for multi-billion dollar capital development projects
  - Burj Khalifa tower in Dubai
  - Pentagon Renovation project, Arlington Virginia.

Agenda

- Problem Statement
- Bird's eye view
- Findings
    - Data overview - EDA
    - Model evaluation
- Conclusions and recommendations

Problem Statement

Business Objective

ESPN Market research team is looking for data on
 - Two most popular sports

- Soccer 3.5 billion Fans
- Cricket 2.5 billion Fans

Challenges

- Untrained interns
- Similar looking blogs

Desired Outcome

- Using API to extract data
- Train the classifier using NLP
- Evaluate the model using Accuracy as the criteria

Birds eye view

Qualitative data 1

- Two Subreddits
    - Cricket - Positive and Soccer - Negative

Qualitative data 2

- Time scale
    - 250 days of blog posts

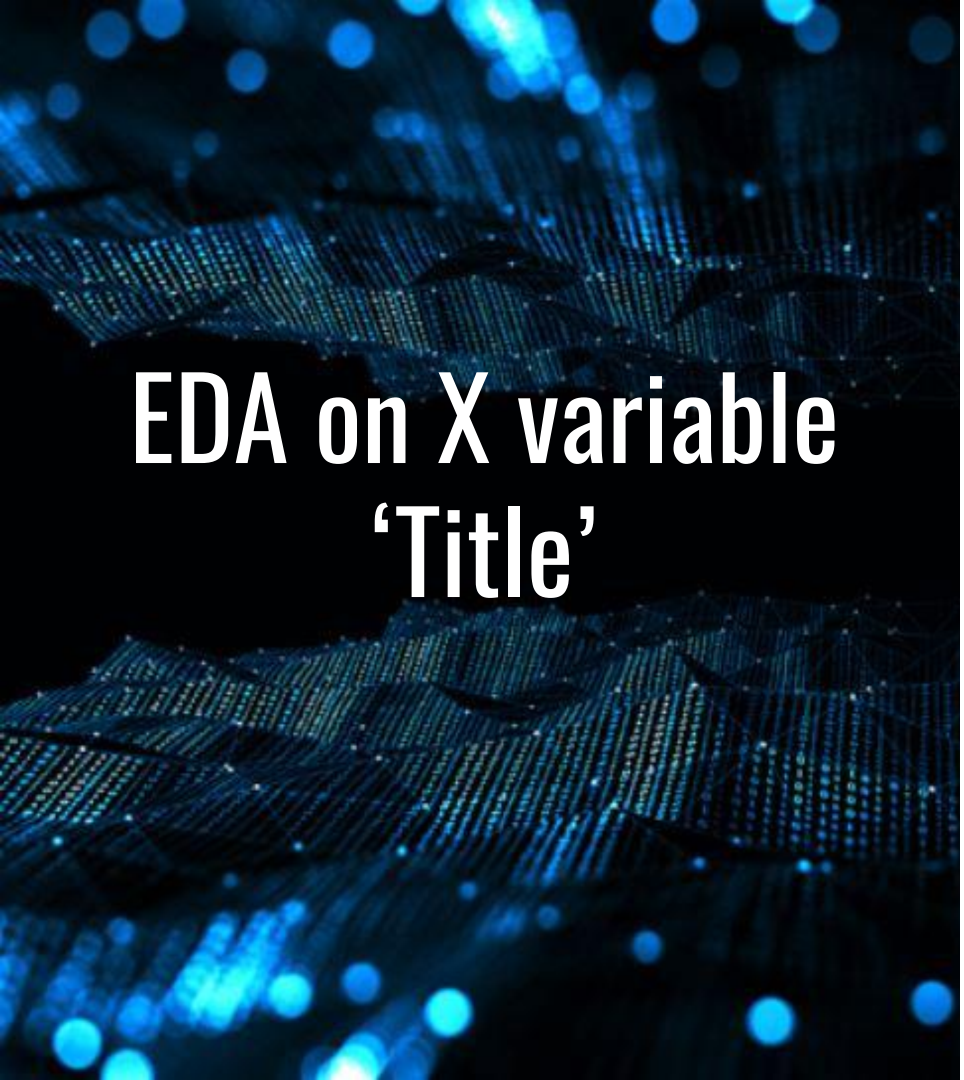
Qualitative data 3

- Baseline Score
    - 0.54

Qualitative data 4

- Models / Classifiers Explored
    - Logistic Reg, GaussianNB, Multinomial, Random Forest.

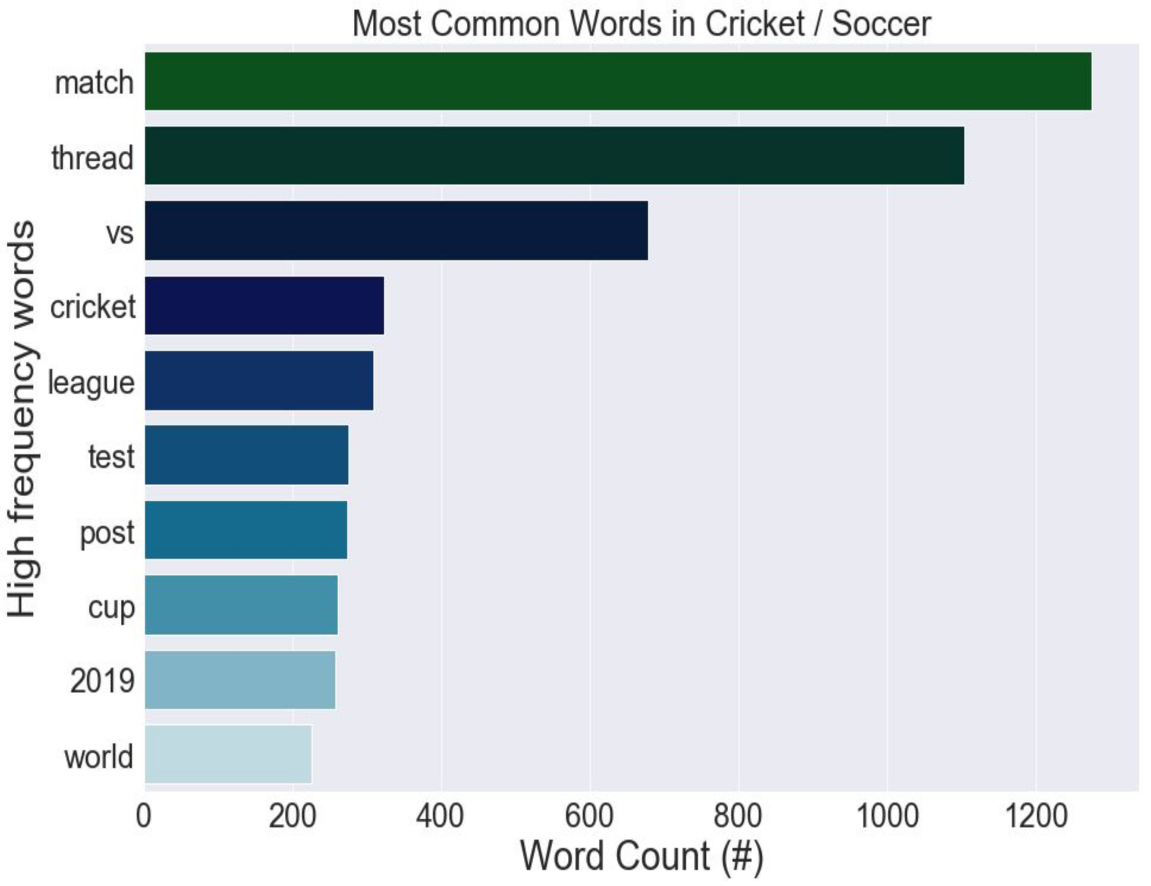
Data Overview - EDA



EDA on X variable ‘Title’

- Tokenize
- Stemming
- Lower Case

Data overview - EDA



- Vectorized the 'Title' feature
- The common ten words found.
- Used them in my model training by augmenting standard Stopwords

Model Evaluation

Criteria :

- Testing Accuracy
- Confusion Matrix
- ROC with AUC curve
- Model coefficants

Model Performance - overview

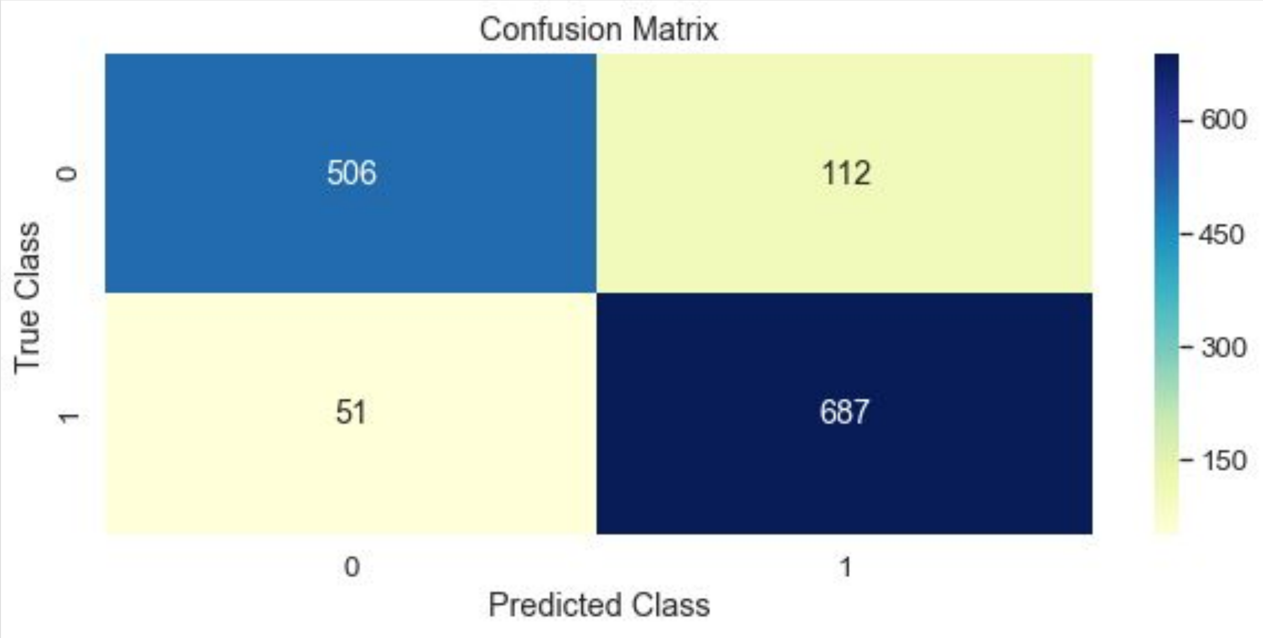
Classifier	Train Score	Test Score	Vectorizer
Base Line	0.54		
Logistic Regression	0.91	0.88	CountVect
Logistic Regression	0.92	0.88	TFIDF
Gaussian	0.87	0.86	CountVect
Multinomial	0.87	0.86	TFIDF
Random Forest	0.87	0.87	TFIDF

Model Evaluation

Logistic regression with TFIDF vectorizer produced 0.88 accuracy score - better than other machine models.

Data overview - Confusion Matrix

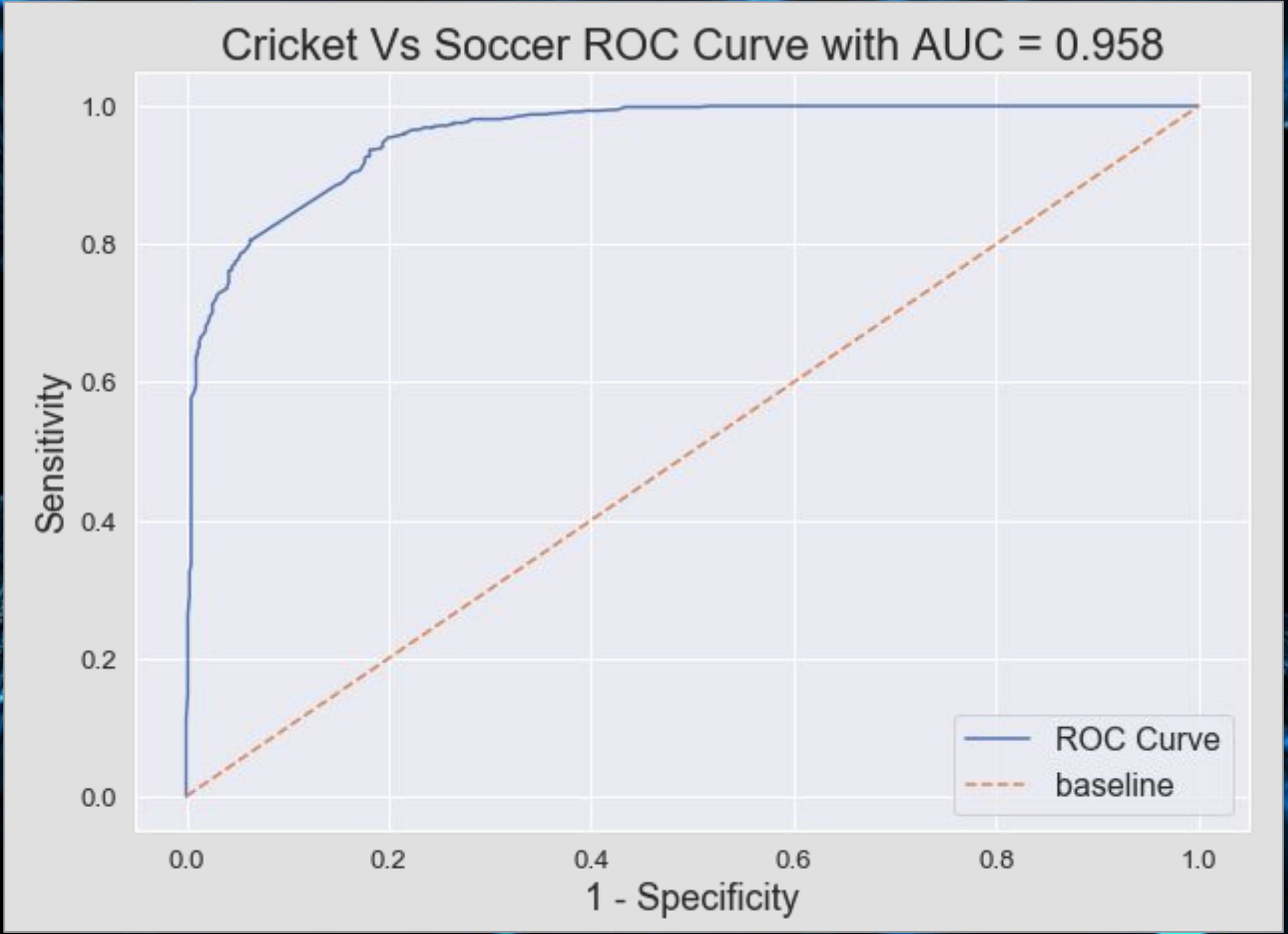
- Accuracy = 0.88
- Misclassification Rate = 0.12
- Specificity = 0.82
- Precision = 0.86
- Sensitivity = 0.93



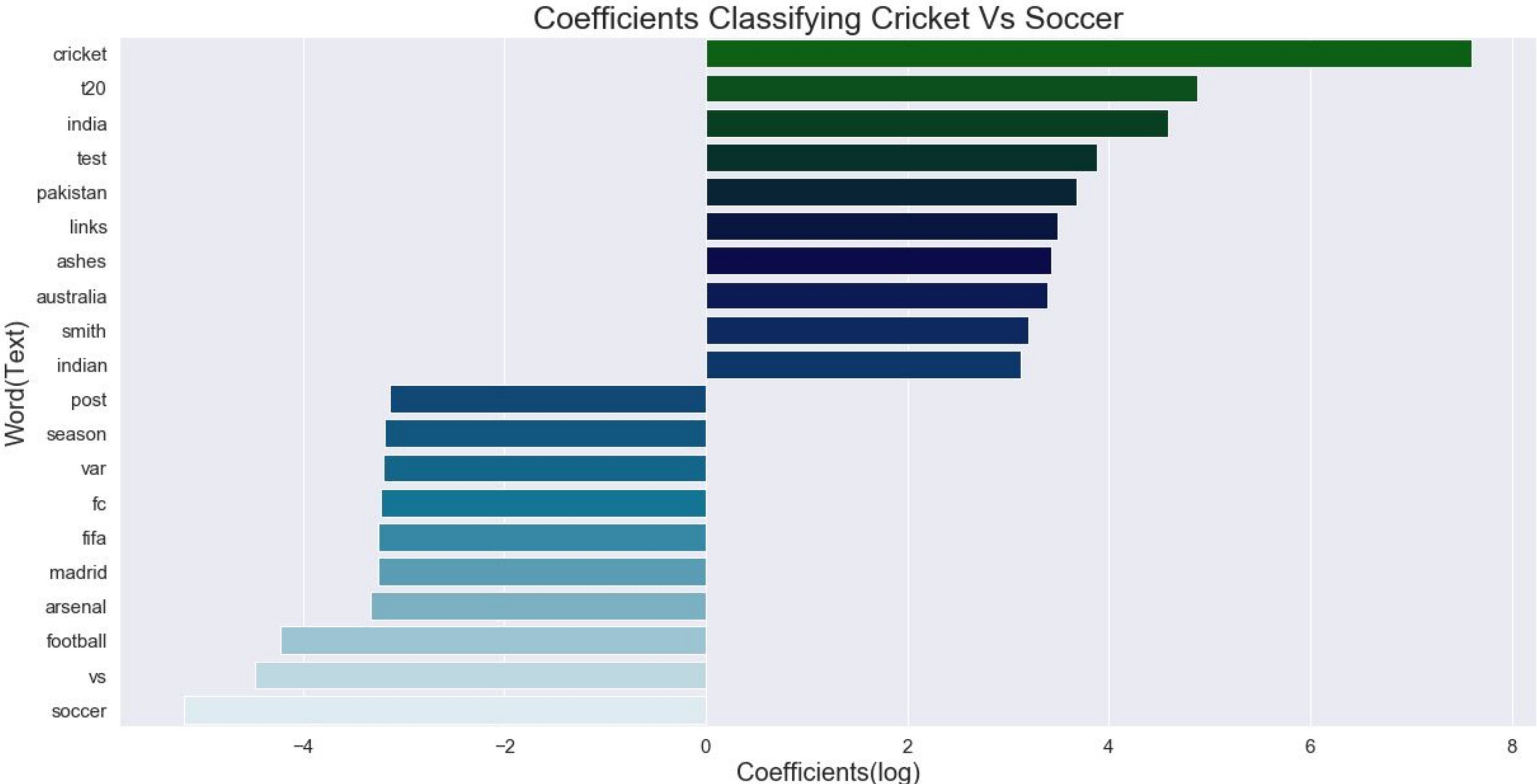
Data overview - ROC AUC Curve

- ROC AUC of close to 1

- Positive and Negative classes are perfectly separated



Data overview - Model Coefficents



Conclusion

- Logistic Regression with TFIDF performed the best (at 89% accuracy).
- Our model will help to differentiate the Cricket and Soccer Blog posts for the Market Research Team
- Would like to get more data from other sports blogs to train my model and bring down the variance.
- Would like to do better feature engineering using the Comments, Self text.

Thanks - Questions?

The Team

Kathirvel Kumararaja

Mahdi

Ambar

GA DSI 10 Cohort