DATA-DRIVEN INVESTING ONLINE:

STOCKS & REAL ESTATE

FINDING INVESTMENT GOLD NUGGETS IN THE DIGITAL WILD WEST

DATA-DRIVEN INVESTING ONLINE: STOCKS AND REAL ESTATE

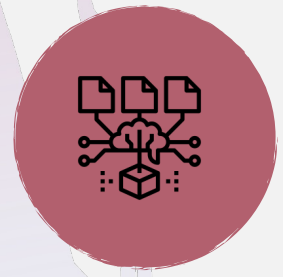
FINDING INVESTMENT GOLD NUGGETS IN THE DIGITAL WILD WEST

Content

Background &
Problem Statement

Work-flow

Data Cleaning &
EDA



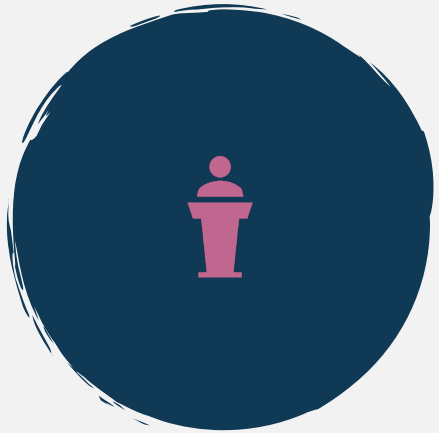
Modelling

Conclusion &
Recommendation

Benefits of investing with us

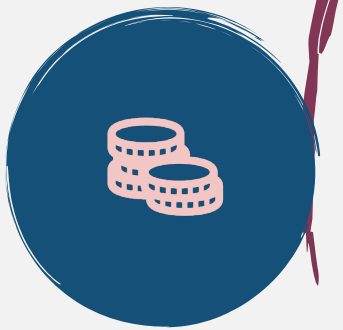
Data-Driven

Uses data-backed science to provide investor with valuable insights and informed decision-making strategies



Maximize returns

By partnering with us, investors can take advantage of cutting-edge tools and strategies to maximize their returns



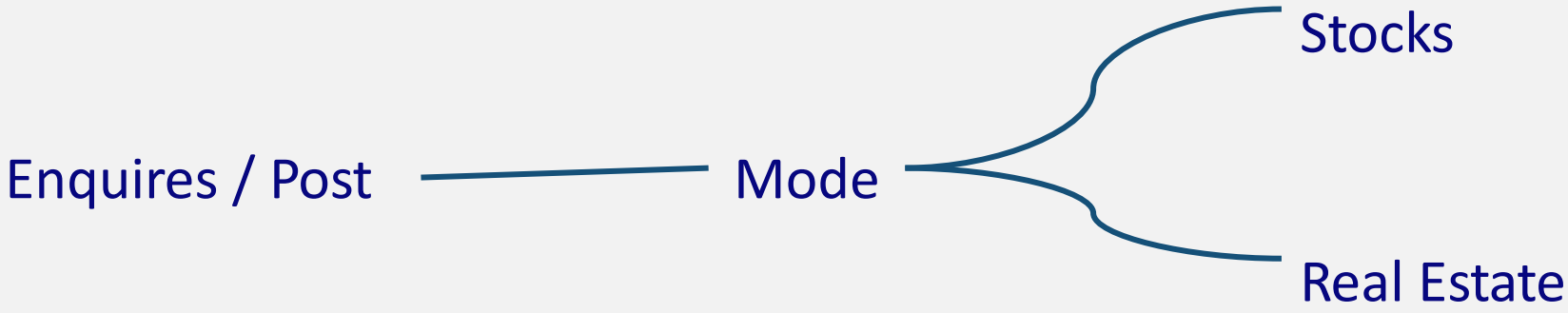
Innovative Investing

Forward-thinking, technology-driven partner in the world of investing, with a focus on helping investors grow their wealth and achieve their financial goals

BACKGROUND & PROBLEM STATEMENT

To develop a machine learning model that accurately predicts future investment performance based on online discussions and patterns

Stocks and real estate are popular long-term wealth growth options that can generate income and appreciation, with proper investment strategy and risk management.



HOW DO WE DO IT?

Contoso

WORK FLOW

Web-Scraping from Reddit

Post from Subreddit
- Real Estate Investing
- Stocks

Data Cleaning

- Cleaning of text
- Tokenization

EDA

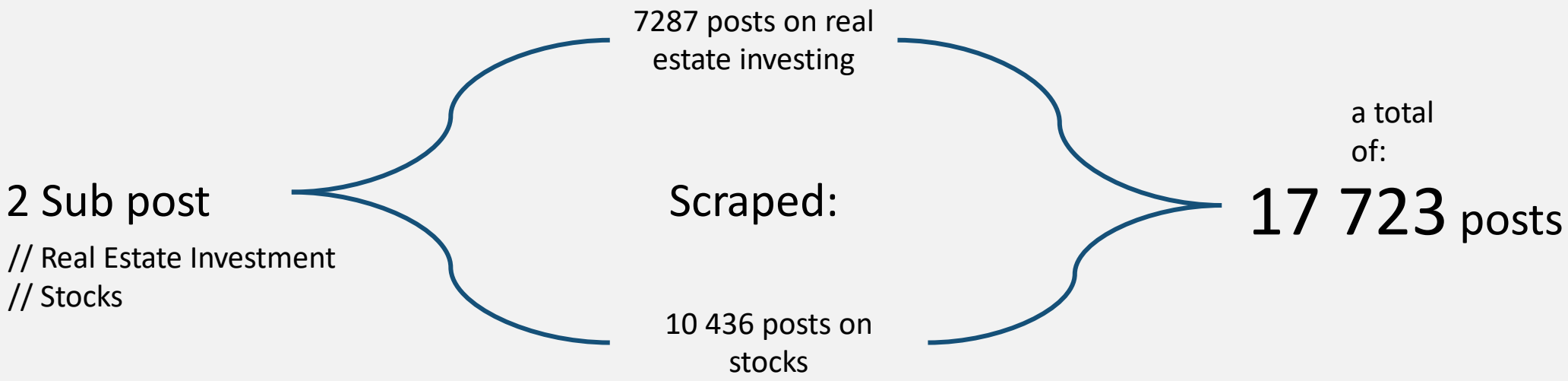
- Identifying common words

Modeling

- Build model to predict new text

WORK FLOW

Contoso



RAW DATA FROM REDDIT

Contoso

DATA CLEANING

Removal of empty / removed values

Removal of stop words

Lemmatization

Removal of symbols & punctuation

e.g: ?!@$,.

Tokenization

Convert long string into tokens

DATA CLEANING

Contoso

title_desc
CFP recommended "10% rule" for income?? In sho...
Average Turn costs on 1900-1950s SFH Hey all. ...
Buying a triplex with a partner and personally...
Buying a decent-sized house in Bay Area/LA for...
HOA rental restriction Hello, thinking about b...

preprocessed
cfp recommended rule income short wife debati...
average turn cost s sfh hey use management com...
buying triplex partner personally living one u...
buying decentsized house bay areala large amou...
hoa rental restriction hello thinking buying t...

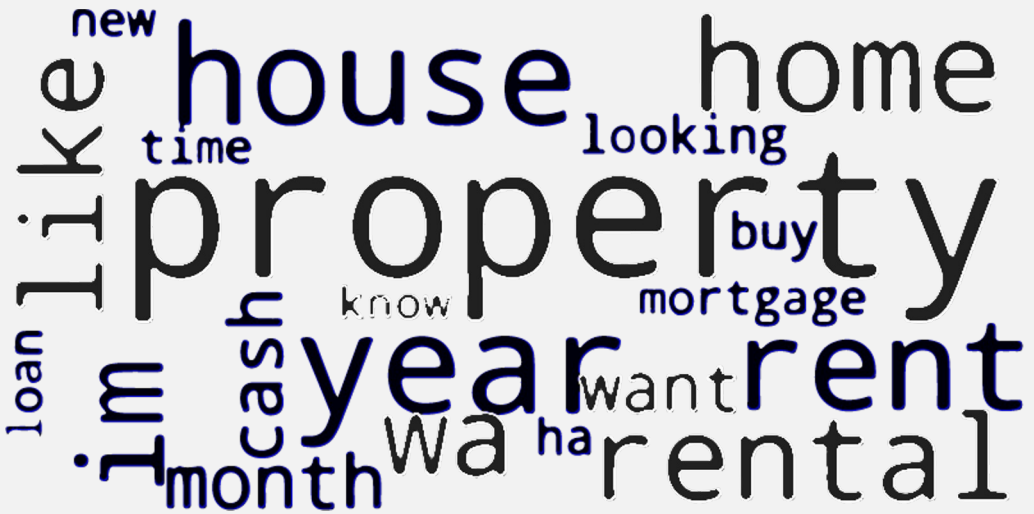
DATA CLEANING

Contoso

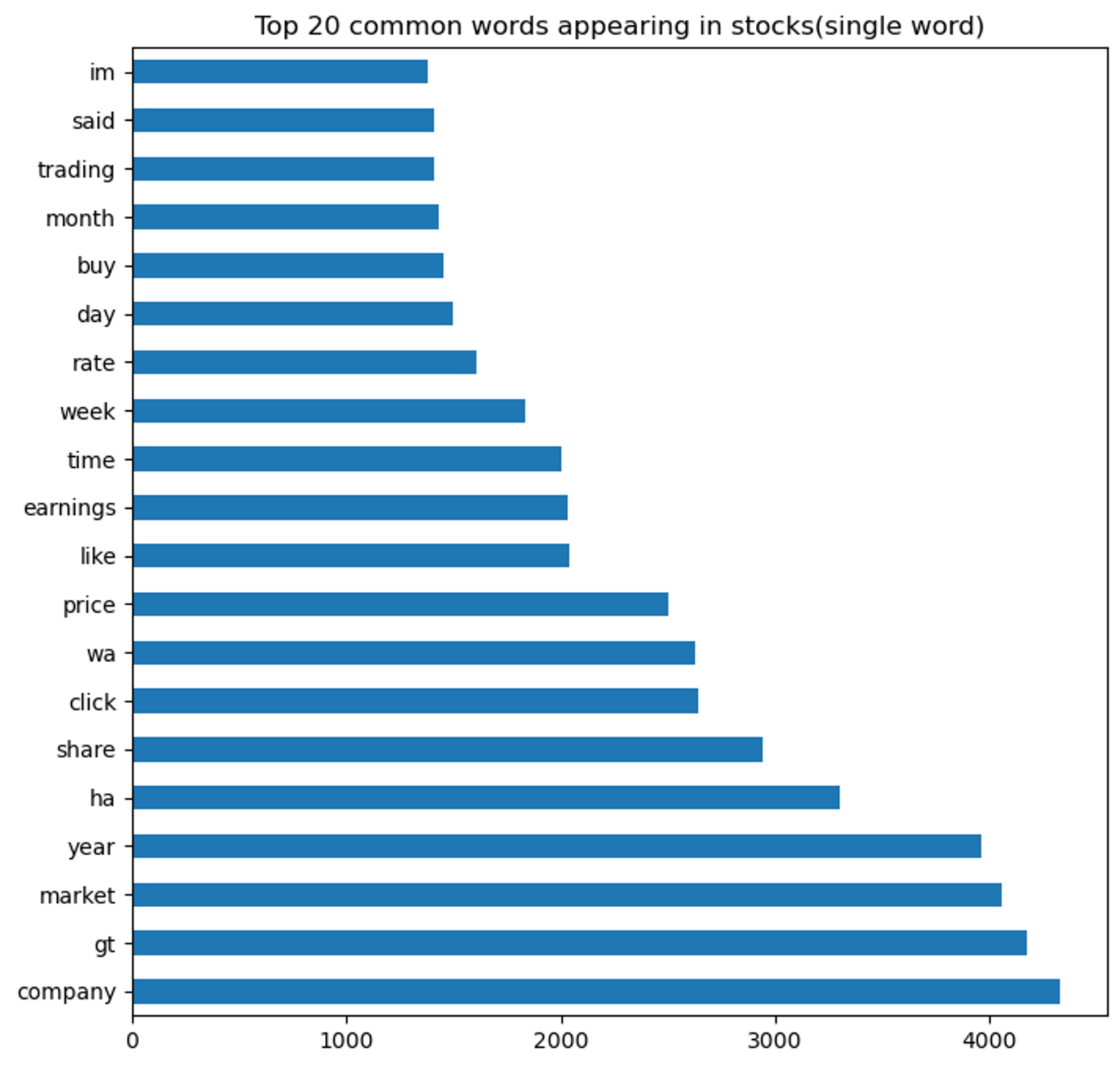
EXPLORATORY DATA ANALYSIS

R/Stocks

R/Real Estate Investing



Contoso

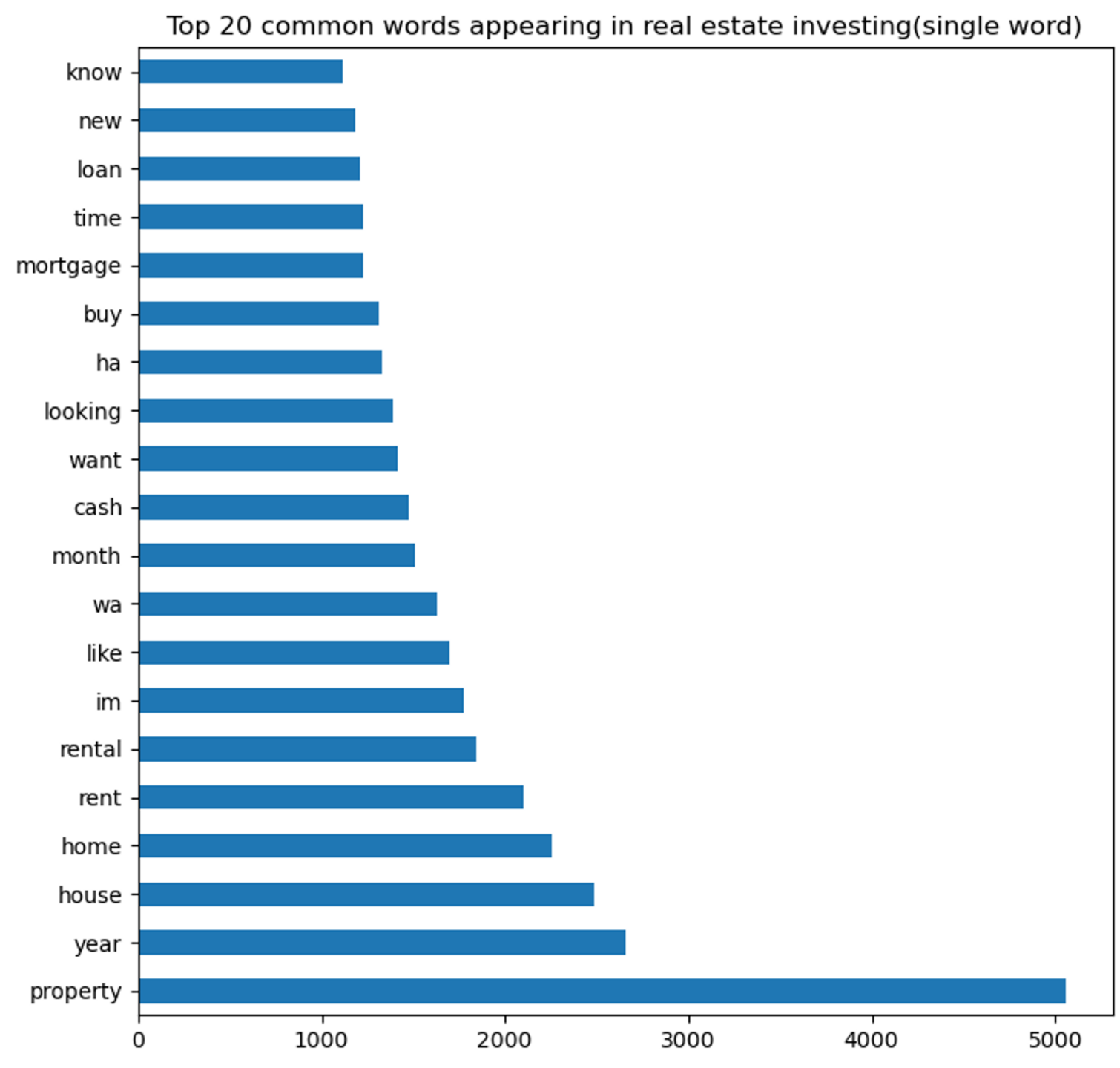


EDA (TOP 20 MOST COMMON WORD)

1.05 TIMES 'COMPANY' PER POST

1.02 TIMES 'GT' PER POST

Contoso

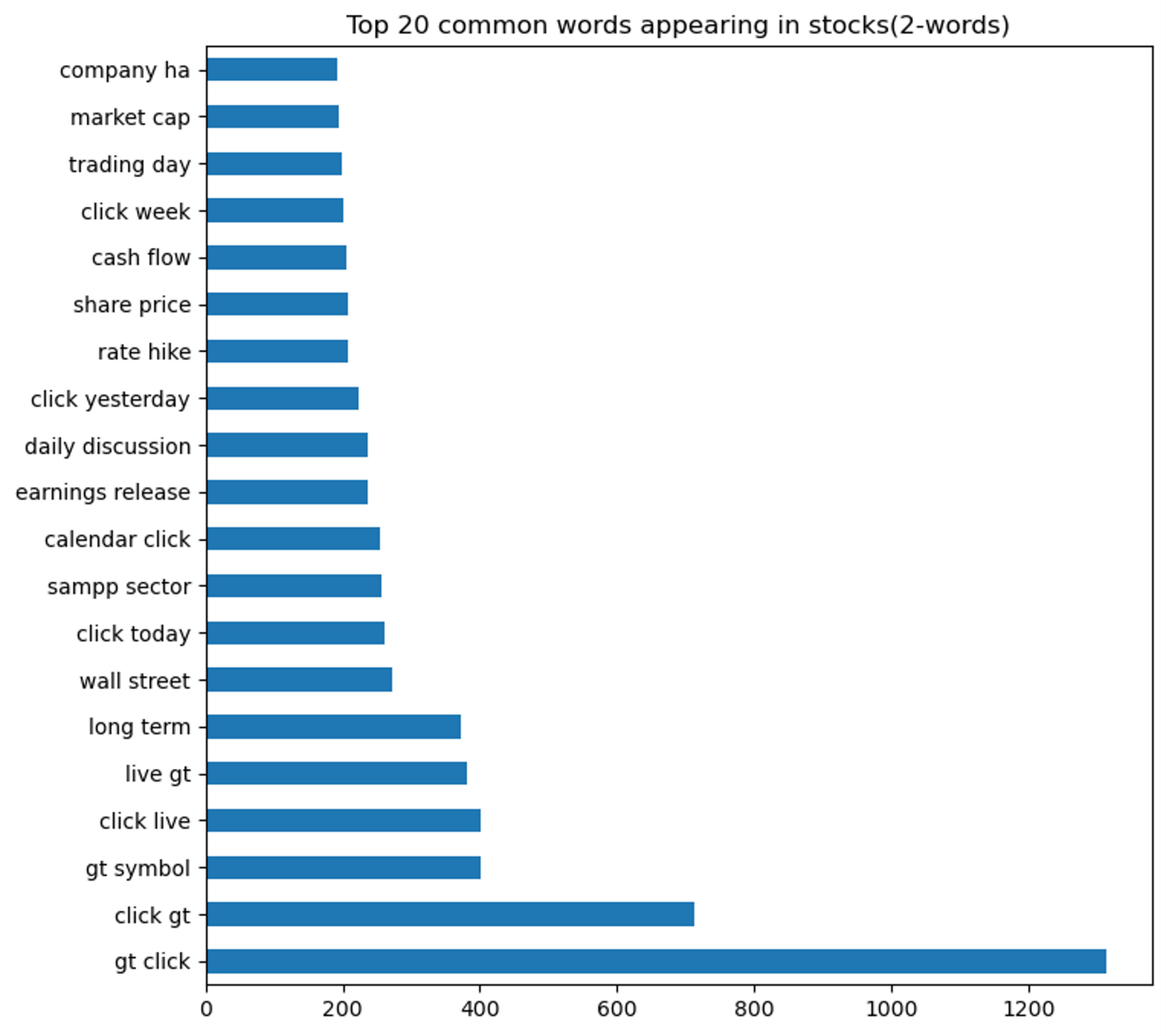


EDA (TOP 20 MOST COMMON WORD)

1.2 TIMES 'PROPERTY' PER POST

0.7 TIMES 'YEAR' PER POST

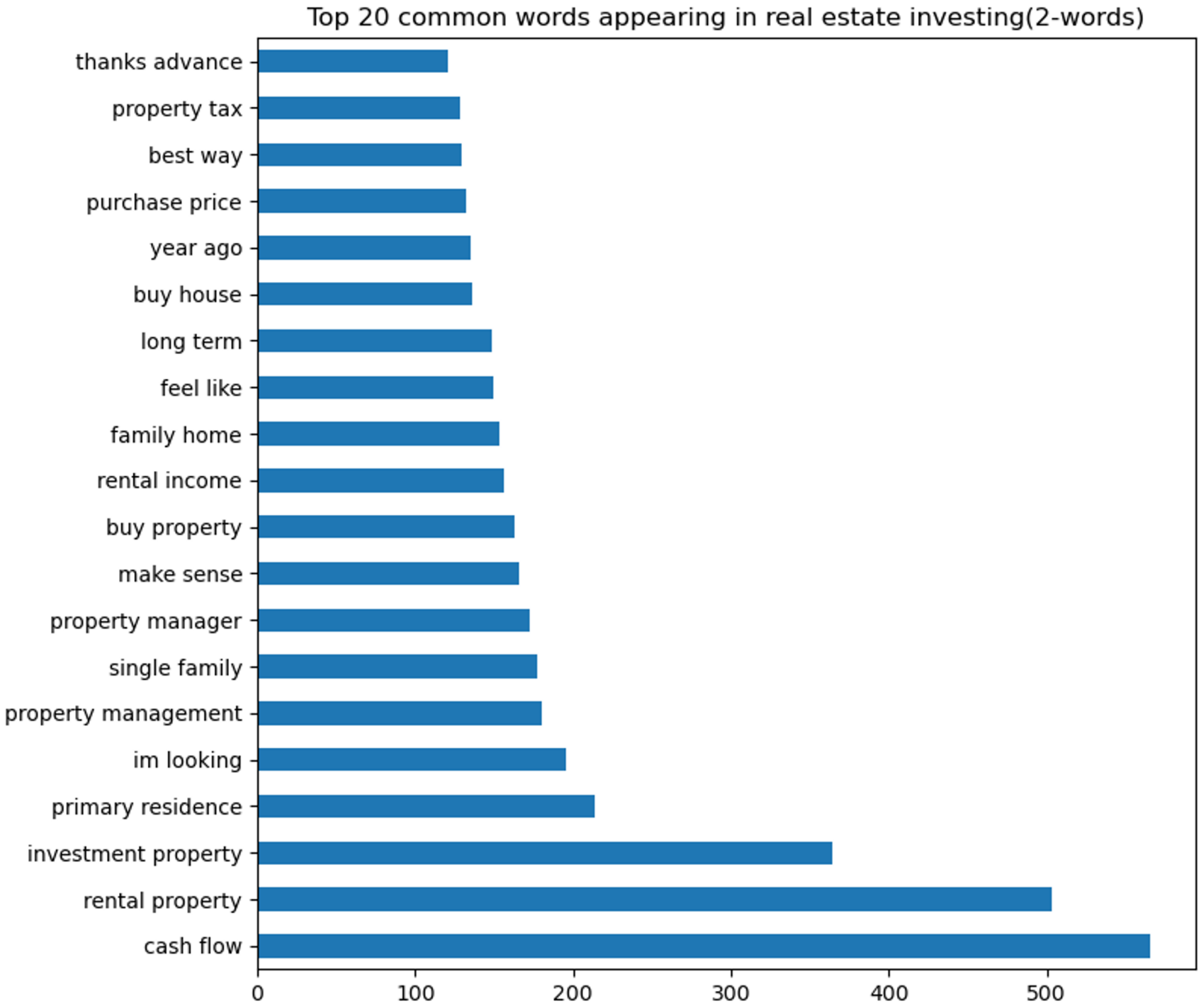
Contoso



EDA (TOP 20 MOST COMMON WORD)

0.8 TIMES ‘GOT CLICK’ PER POST

Contoso



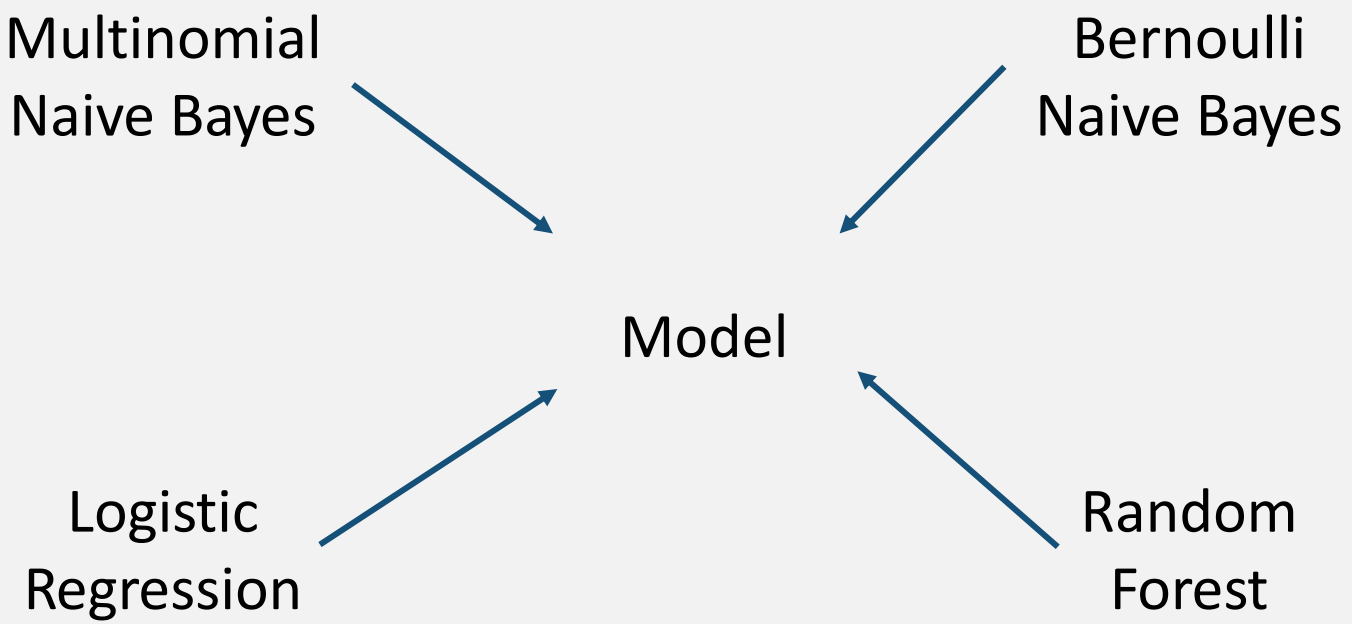
EDA (TOP 20 MOST COMMON WORD)

0.5 TIMES 'CASH FLOW' PER POST

0.12 TIMES 'RENTAL PROPERTY' PER POST

Contoso

MODELLING AND RESULT



TYPES OF MODEL USED

Contoso

- **Logistic Regression with TF-IDF vectorization selected**
- **Optimized for accuracy**

Model	Method	Train Score	Test Score	Precision(Stocks)	Precision(Real Estate)
Logistic Regression	BOW	0.999	0.9622	0.96	0.96
Multinominal Naive Bayes	BOW	0.9755	0.9464	0.97	0.93
Bernoulli Naive Bayes	BOW	0.9415	0.8946	0.97	0.83
Random Forest	BOW	0.9999	0.9651	0.95	0.98
Logistic Regression	**TF-IDF**	**0.9989**	**0.9662**	**0.96**	**0.98**
Multinominal Naive Bayes	TF-IDF	0.9836	0.9645	0.97	0.96
Bernoulli Naive Bayes	TF-IDF	0.9852	0.9566	0.96	0.96
Random Forest	TF-IDF	0.9999	0.9594	0.94	0.98
Logistic Regression	n-gram(2,2)	0.9997	0.9211	0.90	0.95
Multinominal Naive Bayes	n-gram(2,2)	0.9997	0.9278	0.94	0.91
Bernoulli Naive Bayes	n-gram(2,2)	0.9927	0.9228	0.93	0.91

Modelling Result

Contoso

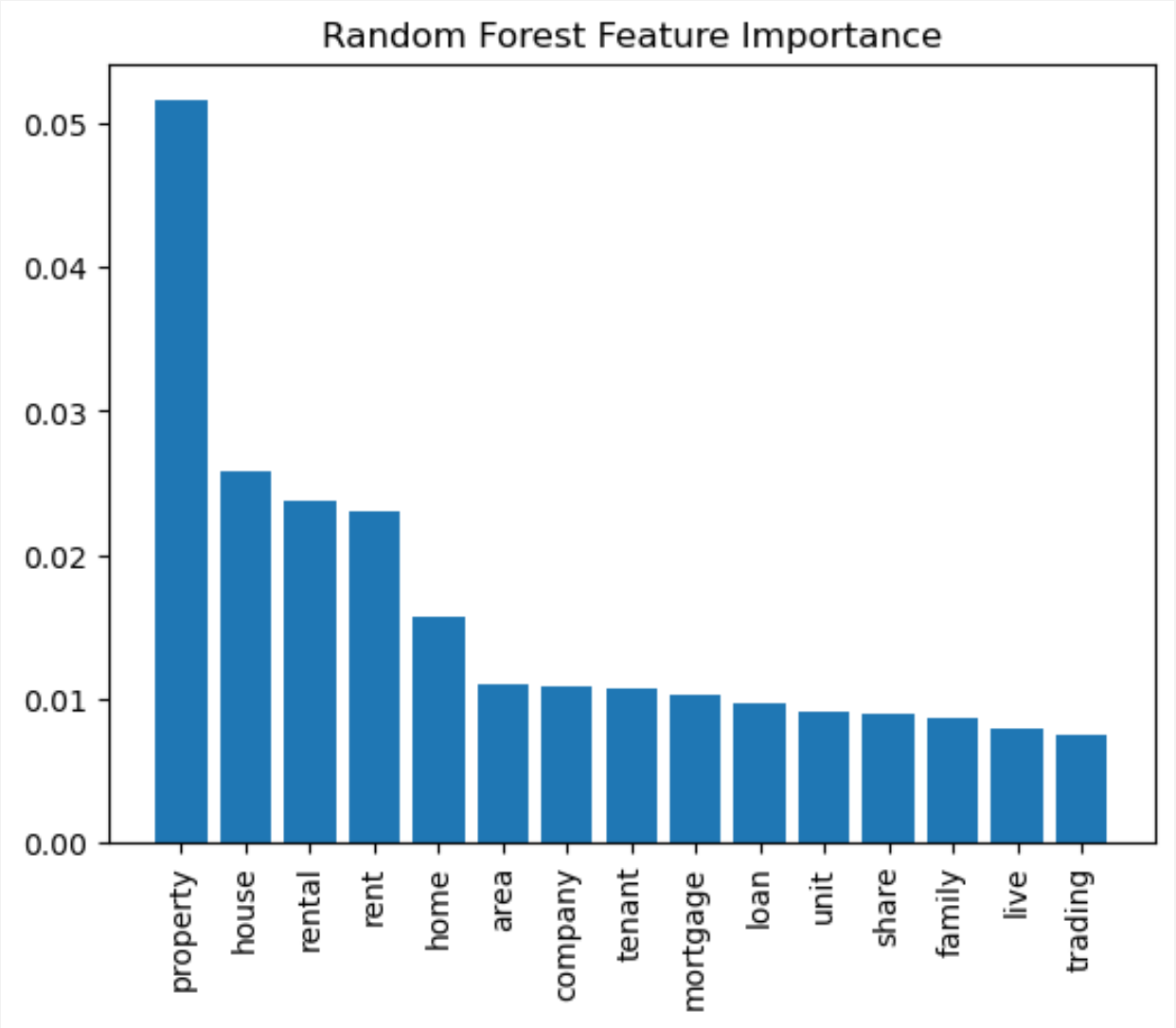
- Logistic Regression is able to predict 97% accuracy
- Training Set Score was at 0.9989
- Test Set Score was at 0.9662
- Both score are similar which could say there is no overfitting

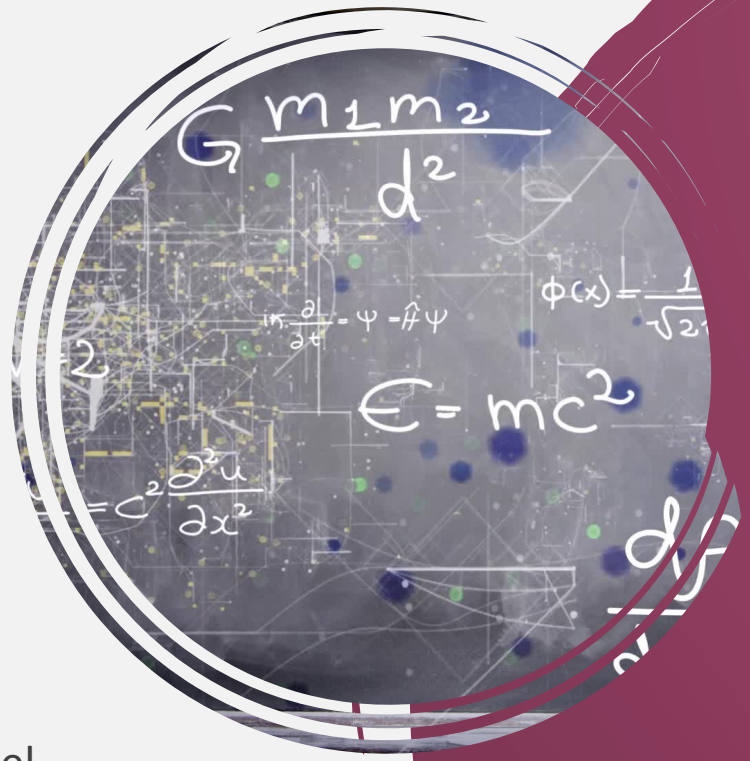
```
------------------------------------------------------------
Model: LogisticRegression()
Preprocessor: TfidfVectorizer()
------------------------------------------------------------
Train Score: 0.9989
Test Score: 0.9662
------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       947
           1       0.98      0.95      0.96       827

    accuracy                           0.97      1774
   macro avg       0.97      0.97      0.97      1774
weighted avg       0.97      0.97      0.97      1774

------------------------------------------------------------
```

LOGISTIC REGRESSION RESULTS

Random Forest Classifier

	Feature_Name	Importance_Score
16920	property	0.051536
10192	house	0.025780
18054	rental	0.023734
18047	rent	0.022947
10062	home	0.015751
1276	area	0.011031
4169	company	0.010904
21708	tenant	0.010751
13866	mortgage	0.010205
12457	loan	0.009692
23015	unit	0.009066
19572	share	0.008942
7669	family	0.008729
12411	live	0.007942
22351	trading	0.007428





Summary

- Logistic Regression showed the best model
- Usage of Random Forest classifier as it is less prone to overfitting and achieve higher accuracy.
- Lack of Data
  - Using subreddit as source

THANK YOU