GA DSI 26 Project 3: Wine and Beer

By: Lim Zhi Yong

Task:

- Understand consumer patterns
- Identify if the consumer wants winemaking or homebrewing info
- Train model with subreddit posts

Data Description

- 1,000 posts from each subreddit
    - r/winemaking
    - r/homebrewing
- Cleaned punctuation, stopwords, delimiters
- Considered both unigrams and bigrams

Notes

“wine”

“Wine” was top classified word for winemaking, but second in misclassified posts

Usual suspects

- hop', 'malt', and 'keg' for beer
- 'grape', 'skin', 'age' for wine

Overlapping words

- Sugar
- Yeast
- Ferment

Seeking advice

'first time' comes up relatively frequently

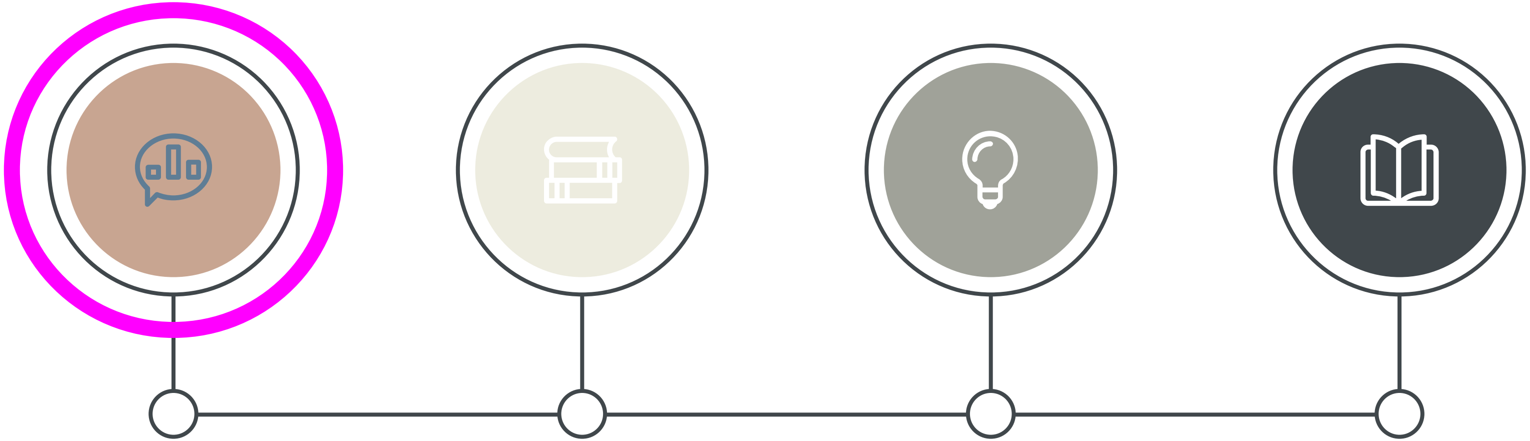
Types

Wine has more types (strawberry, elderberry, banana) than beer (pale ale, ginger) in top 20

Tokenize

Bigrams had more unique tokens than unigrams

Model Workflow



Data Cleaning

- Missing values
- Vectorization

Modelling

- Building models

Testing

- Scoring models

Recommendations

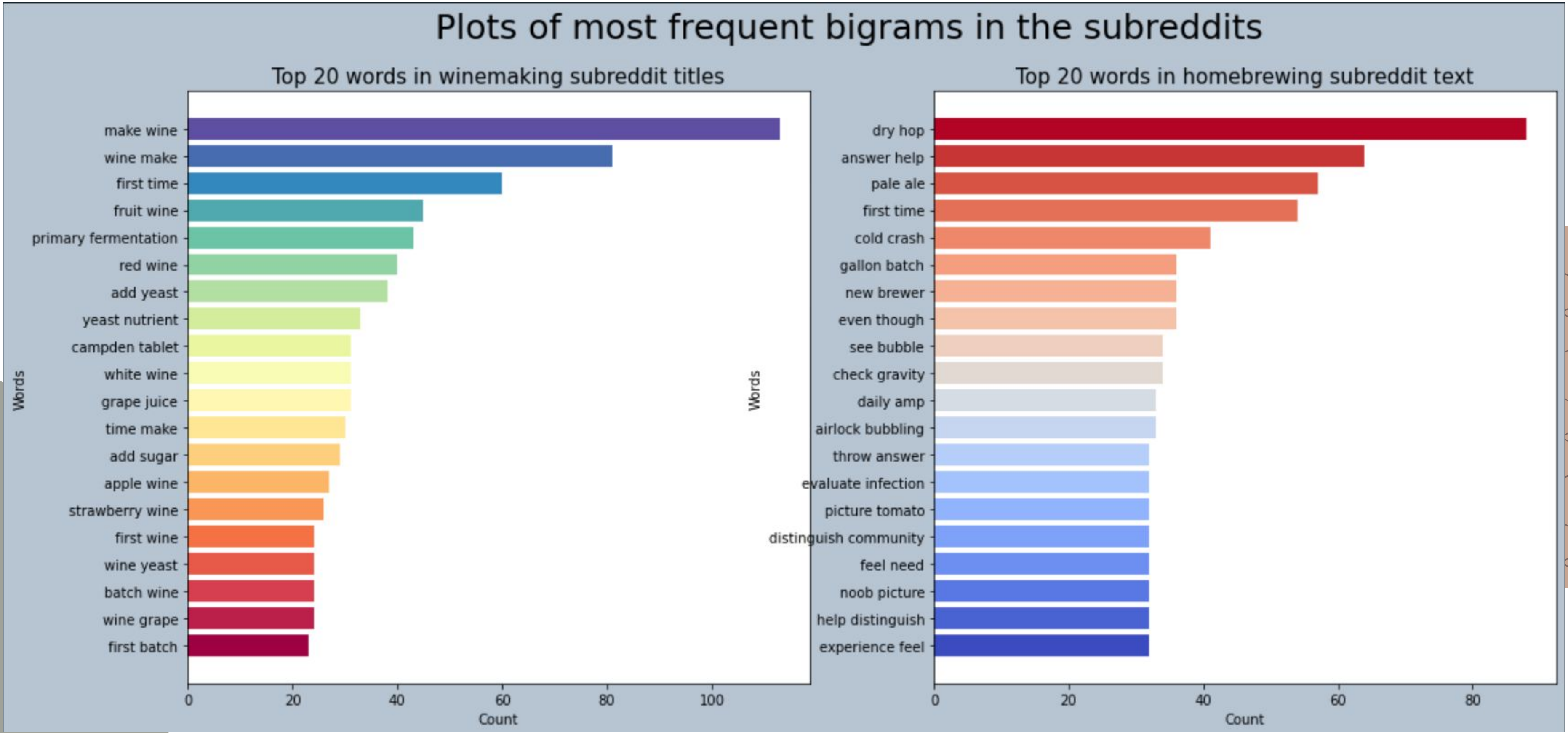
- Important features

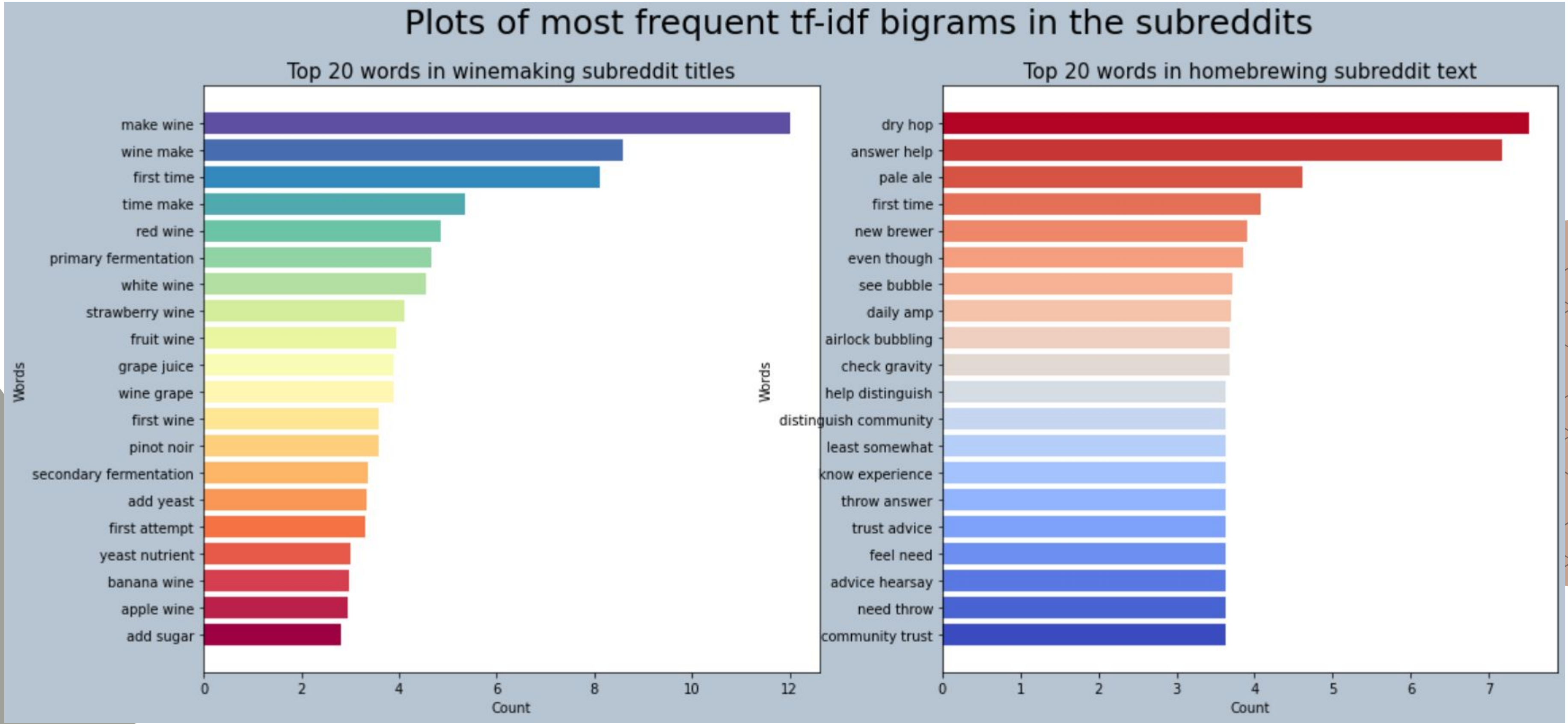
Missing Values

There are different types of missing values:

- Duplicate posts were removed
- Null and removed texts were replaced with the empty string
- One deleted post was miscategorized, we removed it

1969 rows left


Plots of most frequent bigrams in the subreddits
Top 20 words in winemaking subreddit titles
make wine
wine make
first time
fruit wine
primary fermentation
red wine
add yeast
yeast nutrient
campden tablet
white wine
grape juice
time make
add sugar
apple wine
strawberry wine
first wine
wine yeast
batch wine
wine grape
first batch
Words
0
20
40
60
80
100
Count
Top 20 words in homebrewing subreddit text
dry hop
answer help
pale ale
first time
cold crash
gallon batch
new brewer
even though
see bubble
check gravity
daily amp
airlock bubbling
throw answer
evaluate infection
picture tomato
distinguish community
feel need
noob picture
help distinguish
experience feel
Words
0
20
40
60
80
Count

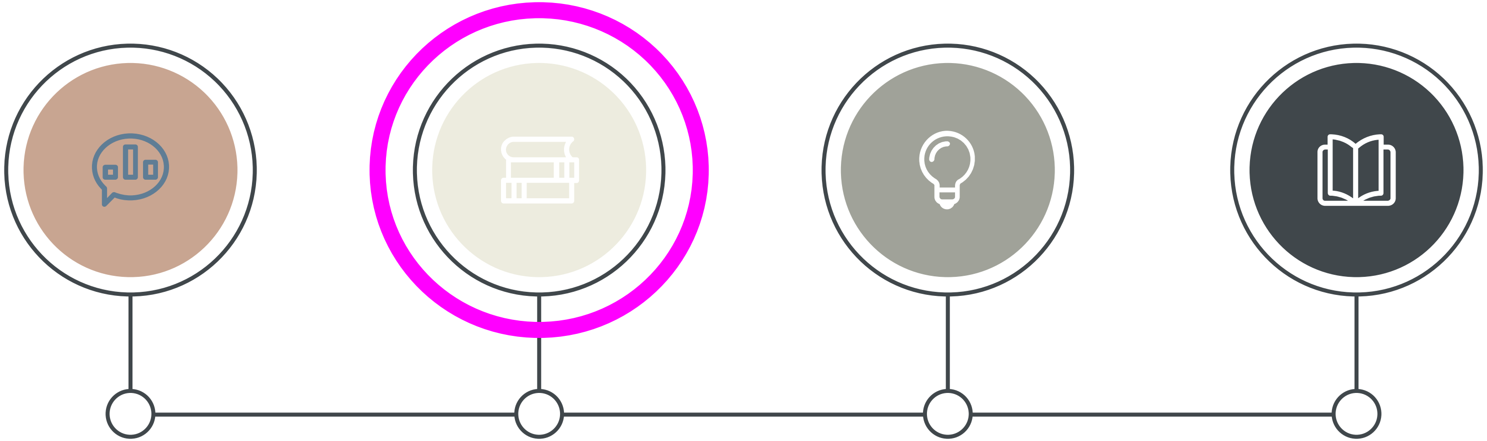


Plots of most frequent tf-idf bigrams in the subreddits

Top 20 words in winemaking subreddit titles

Top 20 words in homebrewing subreddit text

Model Workflow



Data Cleaning

- Missing values
- Vectorization

Modelling

- Building models

Testing

- Scoring models

Recommendations

- Important features

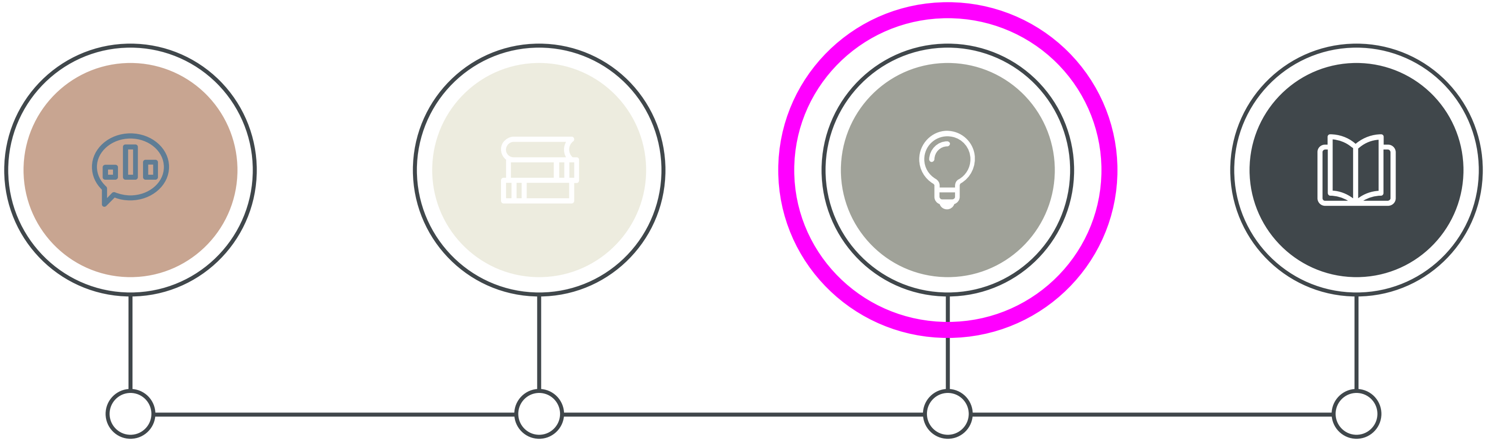
Modelling

- 8 models:
  - Logistic regression
    - Count
    - Tf-idf
  - KNN classifier
    - Count
    - Tf-idf
  - Naïve bayes
    - Count
    - Tf-idf
  - Random forest
    - Count
    - Tf-idf

Scoring

- ROC-AUC to determine best models
- F1 score to compare baseline score
- Accuracy to determine whether overfit

Model Workflow



Data Cleaning

- Missing values
- Vectorization

Modelling

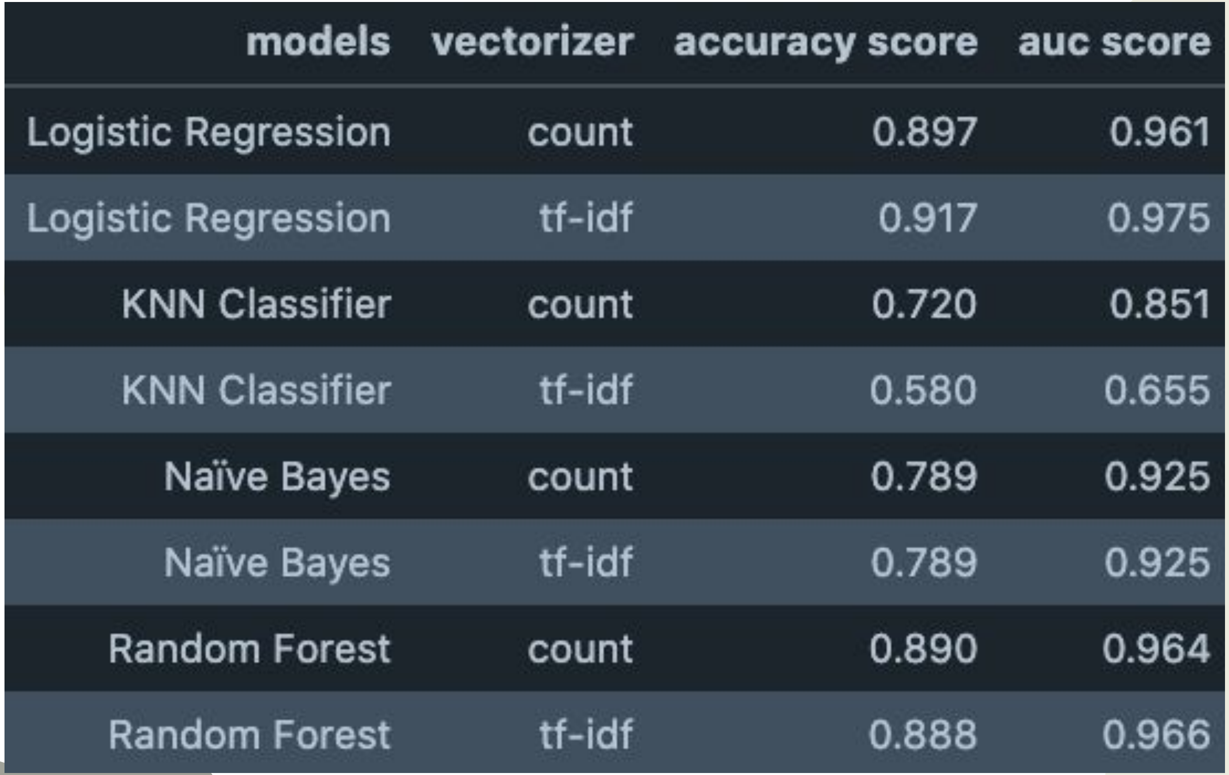
- Building models

Testing

- Scoring models

Recommendations

- Important features



models	vectorizer	accuracy score	auc score
Logistic Regression	count	0.897	0.961
Logistic Regression	tf-idf	0.917	0.975
KNN Classifier	count	0.720	0.851
KNN Classifier	tf-idf	0.580	0.655
Naïve Bayes	count	0.789	0.925
Naïve Bayes	tf-idf	0.789	0.925
Random Forest	count	0.890	0.964
Random Forest	tf-idf	0.888	0.966

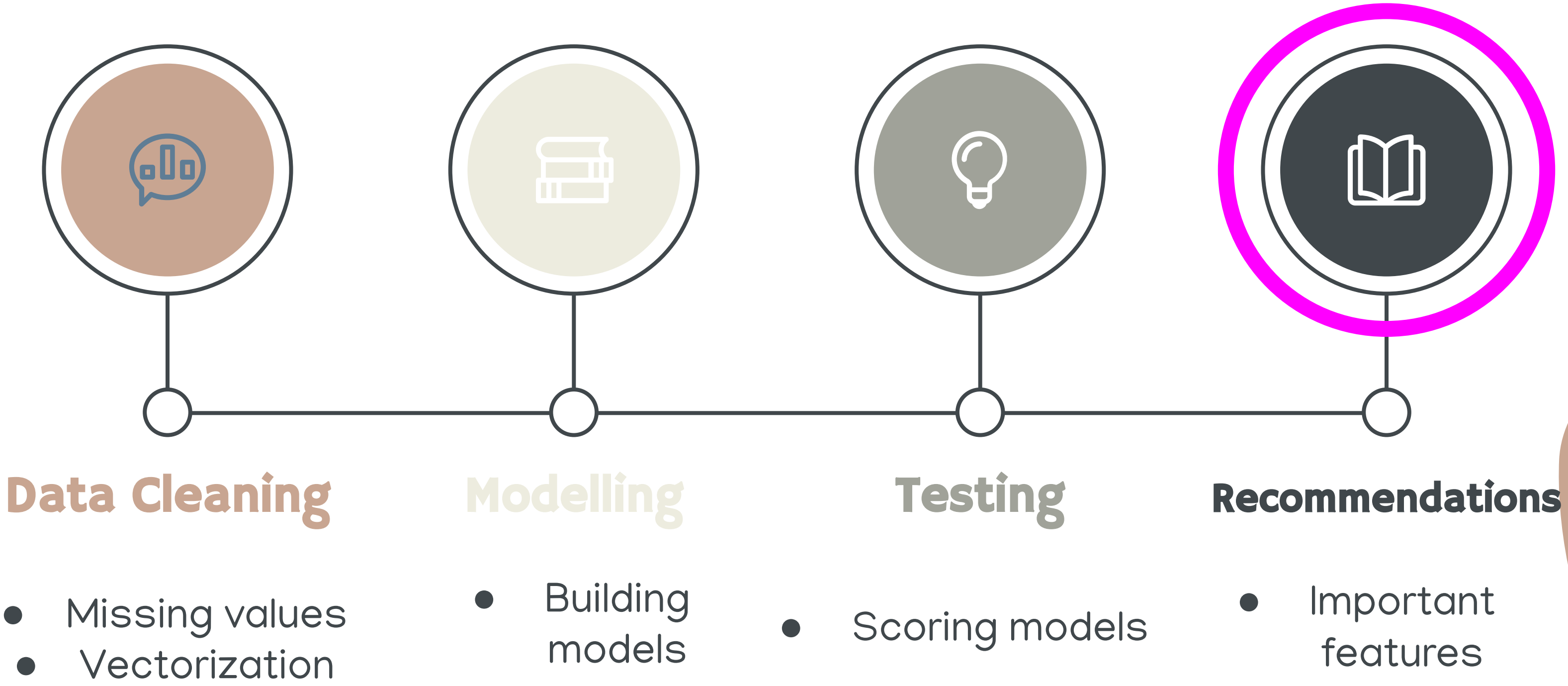
Logistic Regression performed the best

Pipeline params:

- TfidfVectorizer(max_features=4000 , min_df=2, ngram_range=(1, 2)))
- LogisticRegression(C=1, random_state=42, solver='liblinear')

0.97 ROC-AUC
0.92 accuracy

Model Workflow



Data Cleaning

- Missing values
- Vectorization

Modelling

- Building models

Testing

- Scoring models

Recommendations

- Important features

Best Model

Features

- Wordnet lemmatizer
- Tf-idf vectorizer
- Logistic regression with ridge penalty

Limitations

- Spell check before lemmatizing
- Slightly overfit, could remove more stopwords

Thanks

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

Please, keep this slide for the attribution