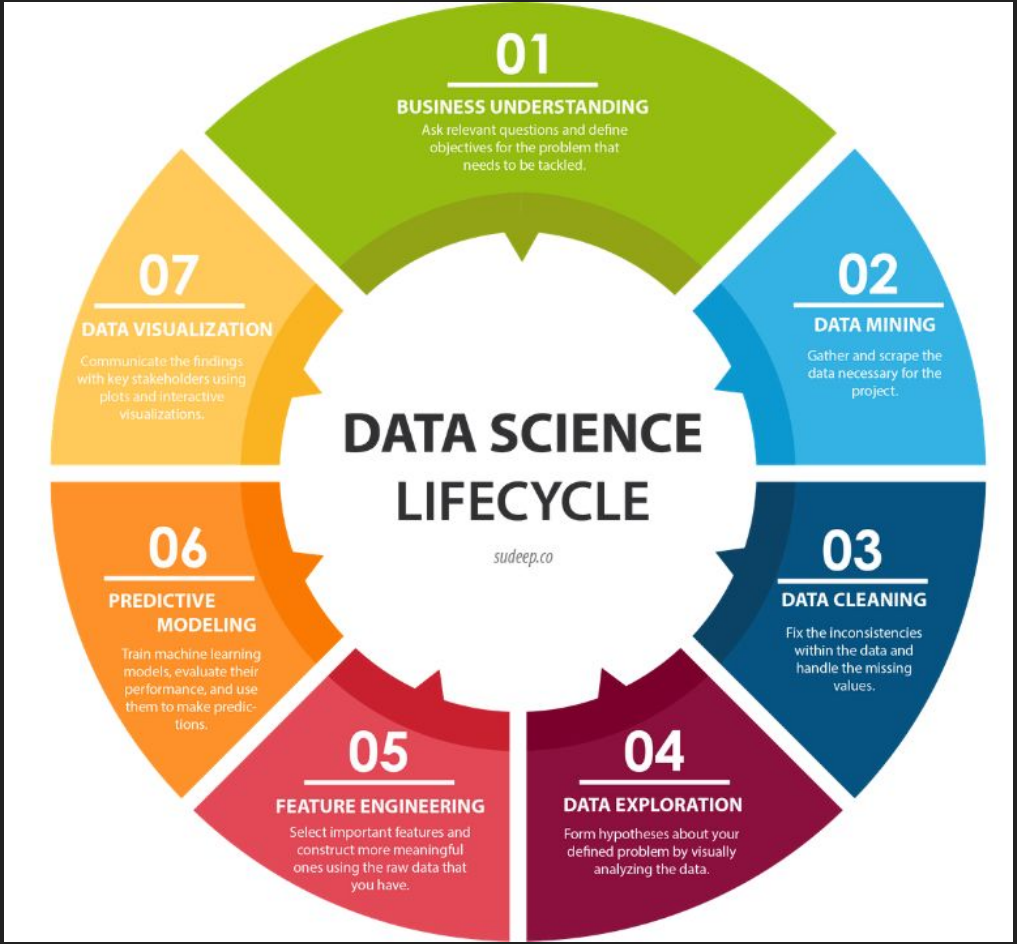
SubReddit Classification

Using webscraping, NLP & classification ML techniques

Done by Ethan Lee
13 Aug 2022

Agenda

1. Define the problem
2. Gather & clean the data
3. Explore the data
4. Model the data
5. Evaluate the model
6. Answer the problem



DATA SCIENCE LIFECYCLE

sudeep.co

01 BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING
Gather and scrape the data necessary for the project.

03 DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.

05 FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.

06 PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.

07 DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

Problem statement

How well can we classify a post belonging to one of two popular subreddits:

/r/bicycling
1.1m subs

/r/motorcycles
1.5m subs

Reason for choosing these 2 subreddits:

- Large number of subs = likely higher quality posts
- Similar (as a mode of transport, mostly 2-wheeled, used for sports etc) but different

Problem Statement

Gather & Clean Data

Explore Data

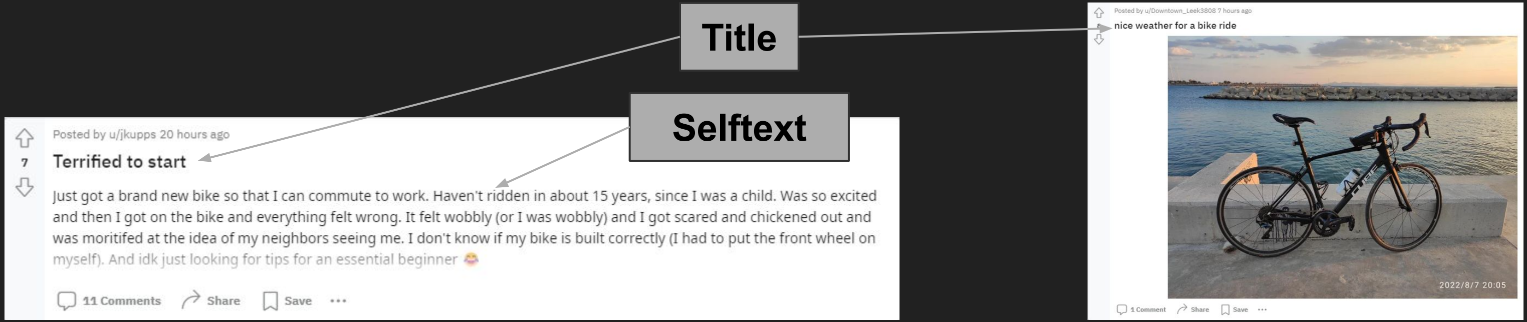
Model Data

Evaluate Model

Answer the Problem

Data gathering & cleaning

- Gathered 3,000 posts per subreddit
- Data cleaning:
  - Joined title and selftext (many empty self texts)
  - Removed duplicates
  - Removed urls, special characters, emojis
- After cleaning, 5,329 posts left
  - 53.5% bicycling and 46.5% motorcycles



Posted by u/jkupps 20 hours ago

7

Terrified to start

Just got a brand new bike so that I can commute to work. Haven't ridden in about 15 years, since I was a child. Was so excited and then I got on the bike and everything felt wrong. It felt wobbly (or I was wobbly) and I got scared and chickened out and was moritifed at the idea of my neighbors seeing me. I don't know if my bike is built correctly (I had to put the front wheel on myself). And idk just looking for tips for an essential beginner 😂

11 Comments Share Save ···

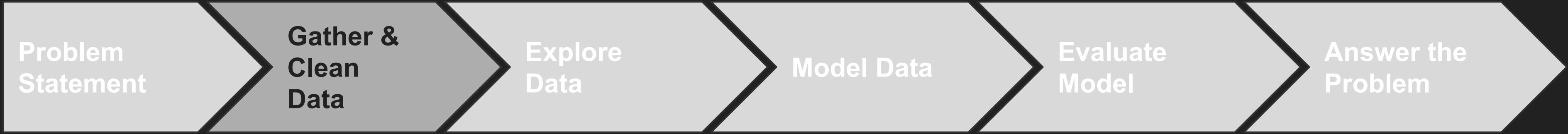
Posted by u/Downtown_Leek3808 7 hours ago

8

nice weather for a bike ride

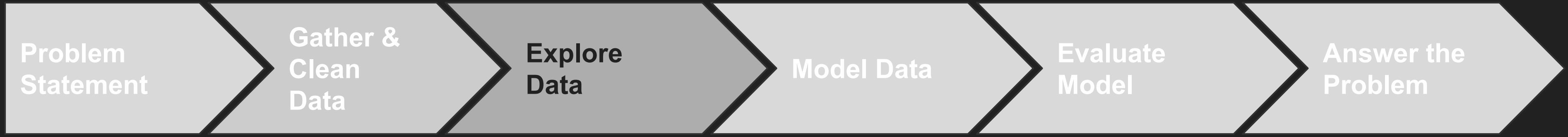
2022/8/7 20:05

1 Comment Share Save ···



Preprocessing and EDA

- Stop words removal
  - Added stopwords such as “bicycle, motorcycle, motorbikes, cyclist” to the original list of “English” stopwords
- Porter Stemming
  - Used to extract the base form of the words by removing affixes
  - E.g. apartment -> apart, garage -> garag, preventing -> prevent
- Lemmatizer
  - Grouping together different forms of the same word
  - E.g. riding-> ride, rides -> ride, rider-> ride



Problem Statement

Gather & Clean Data

Explore Data

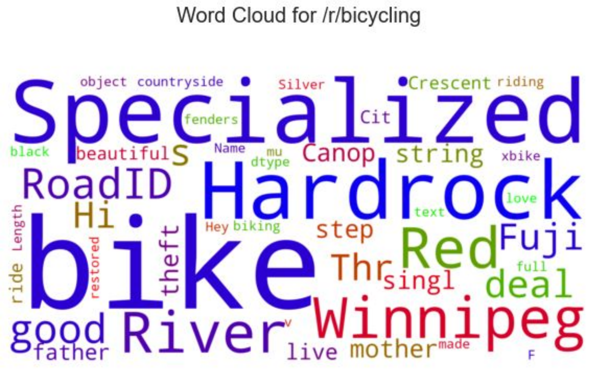
Model Data

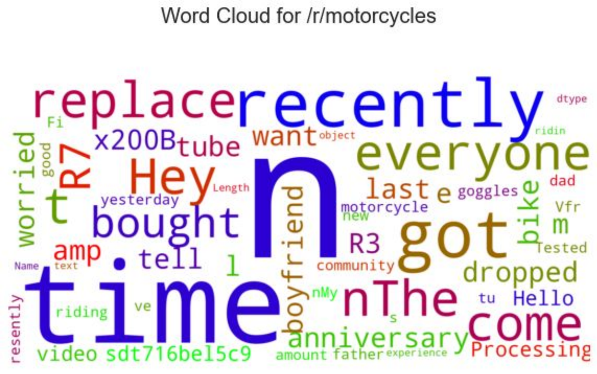
Evaluate Model

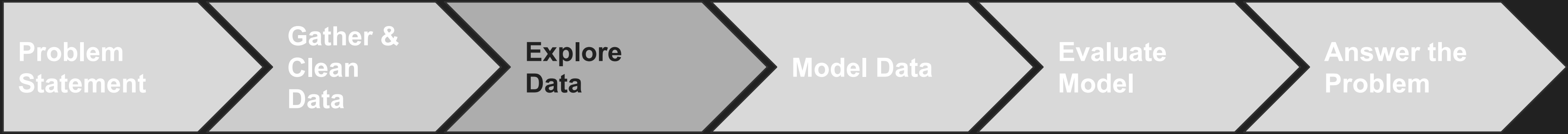
Answer the Problem

Preprocessing and EDA

- WordCloud to visualise common words
- N-grams
  - Continuous sequence of words in a document
  - E.g. First bike, brand new, need help
  - Bigrams more common that tri-grams, and mostly associated with people asking for advice or help


Word Cloud for /r/bicycling


Word Cloud for /r/motorcycles


Problem Statement

Gather & Clean Data

Explore Data

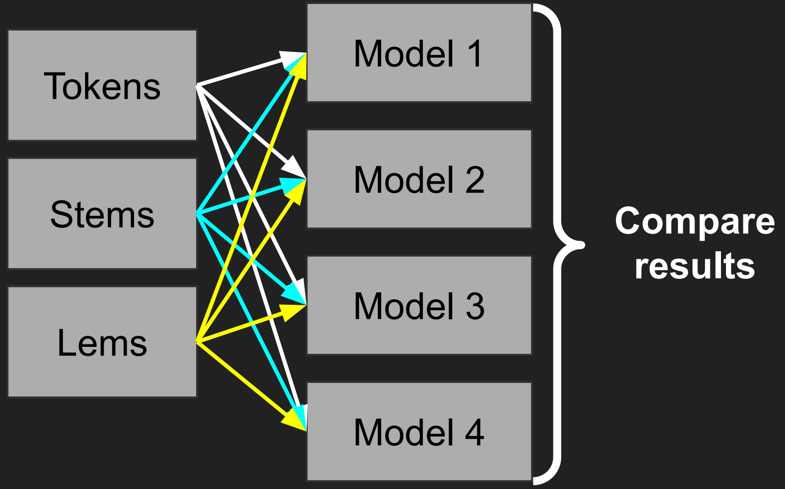
Model Data

Evaluate Model

Answer the Problem

Modelling

- Train-test split done on 3 datasets (cleaned tokens, stemmed tokens and lemmatized tokens)
- Train size 70%, test size 30%
- 4 models:
  - CountVectorizer with MultinomialNB
  - TfidfVectorizer with MultinomialNB
  - CountVectorizer with RandomForest
  - TfidfVectorizer with RandomForest



Tokens

Stems

Lems

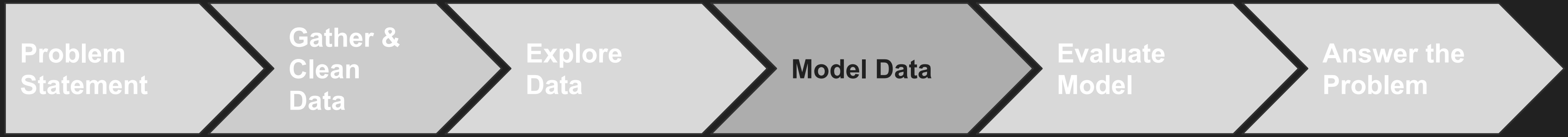
Model 1

Model 2

Model 3

Model 4

Compare results



Problem Statement

Gather & Clean Data

Explore Data

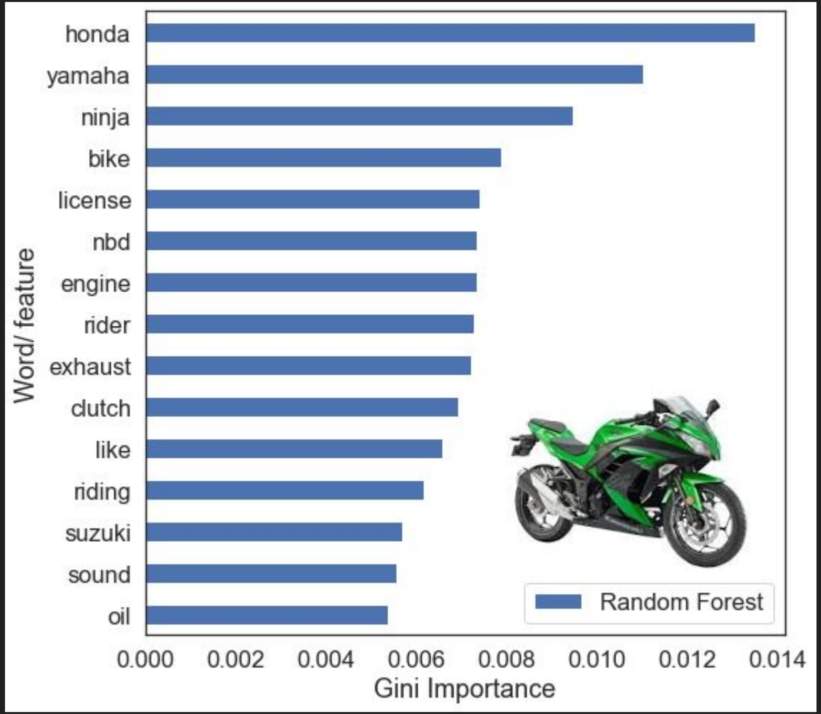
Model Data

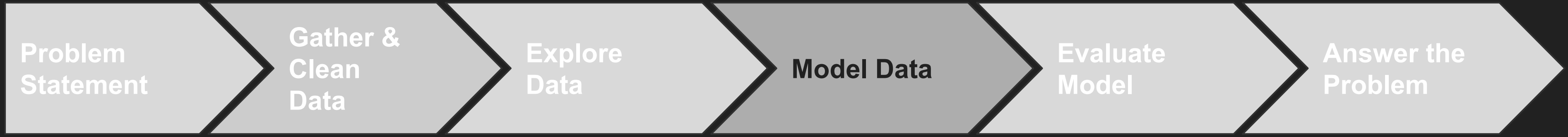
Evaluate Model

Answer the Problem

Feature Importance

- Higher score means that the feature will have a larger effect on the model prediction
- Words used to distinguish motorcycles from bicycles
- Brands: Honda, Yamaha, Ninja, Suzuki
- Motorcycle-specific words: Clutch, engine, exhaust, oil




Problem Statement

Gather & Clean Data

Explore Data

Model Data

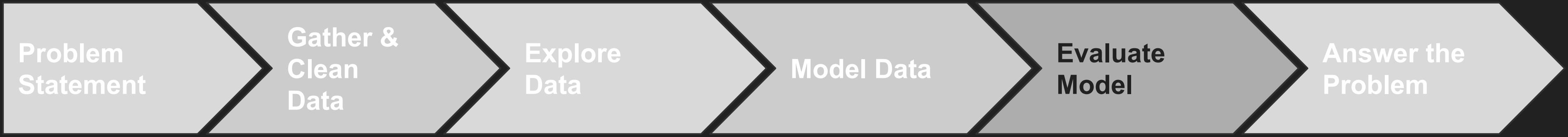
Evaluate Model

Answer the Problem

Model Evaluation

- Compare on accuracy, generalisation, precision, f1 score and AUC ROC score (since dataset in fairly balanced)

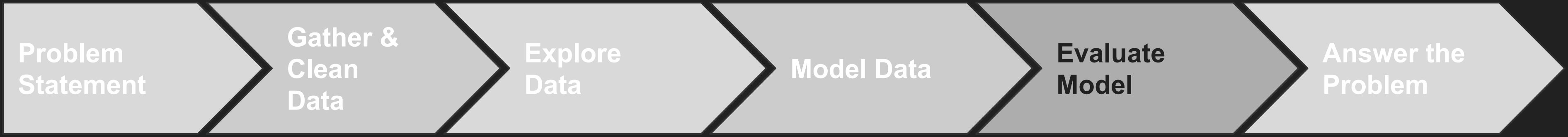
	Accuracy	Generalisation	Precision	F1 score	AUC ROC score
CountVect with NB	0.8737	6.49%	0.8455	0.8677	0.9465
TfidfVect with NB	0.8699	9.11%	0.8534	0.8613	0.9522
CountVect with RF	0.8418	15.6%	0.8375	0.8278	0.9111
TfidfVect with RF	0.8443	15.3%	0.8479	0.8286	0.9138



Model Parameters

- **Max features:** 7,000 (Only top 7000 words from corpus saved)
- **Max df:** 0.95 (Ignore words that occur in >95% of documents from corpus)
- **Min df:** 2 (Word must occur in at least 2 documents from corpus)
- **Ngrams:** (1,2) (Capture every 1 and 2 word phrases)
- **Stopwords:** ‘english’ + common words like ‘bicycles, motorcycles, cyclists’ etc

	Accuracy	Generalisation	Precision	F1 score	AUC ROC score
CountVect with NB	0.8737	6.49%	0.8455	0.8677	0.9465



Problem Statement

Gather & Clean Data

Explore Data

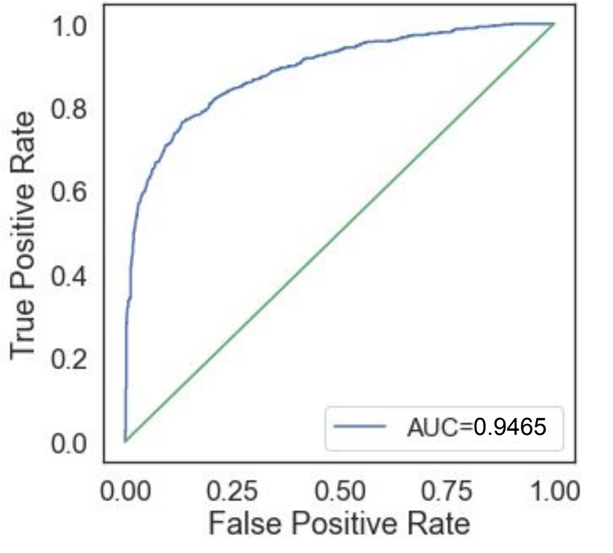
Model Data

Evaluate Model

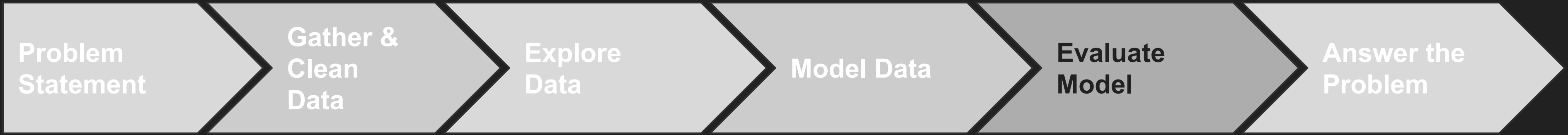
Answer the Problem

Model Evaluation

- CountVectorizer with MultinomialNB was chosen because of its good generalization, accuracy, precision, F1 score and AUC ROC score
- Not as overfitted as the Random Forest models
- Takes lesser time to run (9 seconds) compared to the RandomForest model (28 seconds)



	Accuracy	Generalisation	Precision	F1 score	AUC ROC score
CountVect with NB	0.8737	6.49%	0.8455	0.8677	0.9465



Problem Statement

Gather & Clean Data

Explore Data

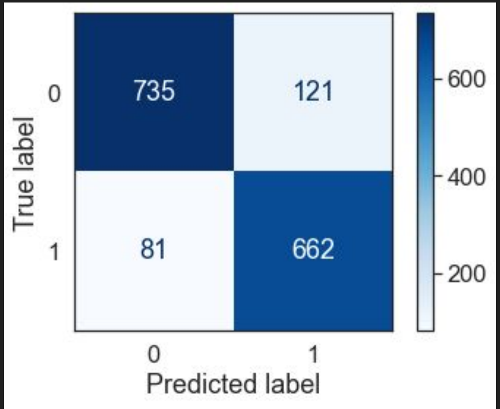
Model Data

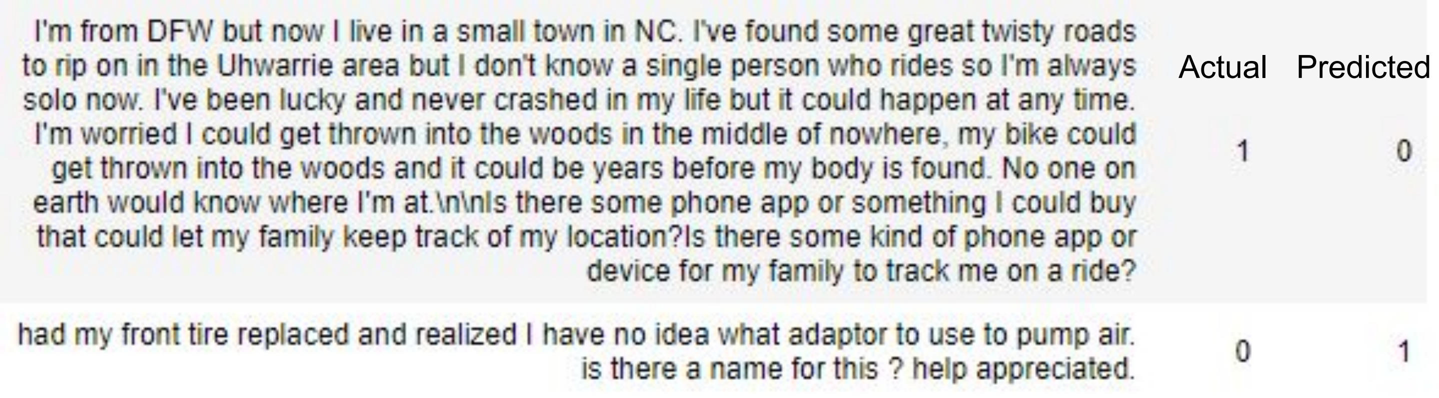
Evaluate Model

Answer the Problem

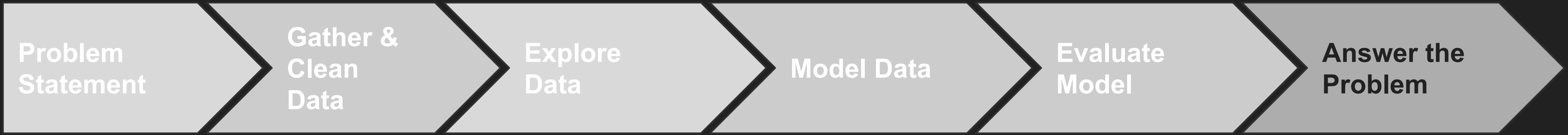
Answering the Problem

- How well can we classify a subreddit post?
- **87% accuracy,** but possible to increase the accuracy and precision of the models using GridSearch instead of Randomized Search (only 25 iterations done in model)





	Actual	Predicted
I'm from DFW but now I live in a small town in NC. I've found some great twisty roads to rip on in the Uhwarrie area but I don't know a single person who rides so I'm always solo now. I've been lucky and never crashed in my life but it could happen at any time. I'm worried I could get thrown into the woods in the middle of nowhere, my bike could get thrown into the woods and it could be years before my body is found. No one on earth would know where I'm at.\n\nIs there some phone app or something I could buy that could let my family keep track of my location?Is there some kind of phone app or device for my family to track me on a ride?	1	0
had my front tire replaced and realized I have no idea what adaptor to use to pump air. is there a name for this ? help appreciated.	0	1



Problem Statement

Gather & Clean Data

Explore Data

Model Data

Evaluate Model

Answer the Problem

Conclusion

- Model performed ok despite some spam posts left over. M**ore thorough cleaning** can be done to remove more spam
- Test out on **other models** other than Naive Bayes and Random Forest
- Explore **new features** within Reddit such as the upvotes and downvotes, as well as post comments
- Metrics are easy to compare for a balanced dataset. In the future, need to make use of sensitivity and specificity for imbalance datasets. The ROC-AUC curve would also have to be substituted with the Precision Recall curve.

Q&A

- Any questions? Thank you!

Appendix

- Random Forest seemed to overfit, generalization is not good.
- Current hyperparameters tested
  - N_estimators: [100, 150, 200]
  - Max_depth: [None, 1, 3, 5]
- How to improve?
  - Reduce number of variables sampled at each split (reduce max_features)
  - Use more data
  - Increase the number of n_estimators
  - Use other hyperparameters such as min_samples_leaf (set >1)