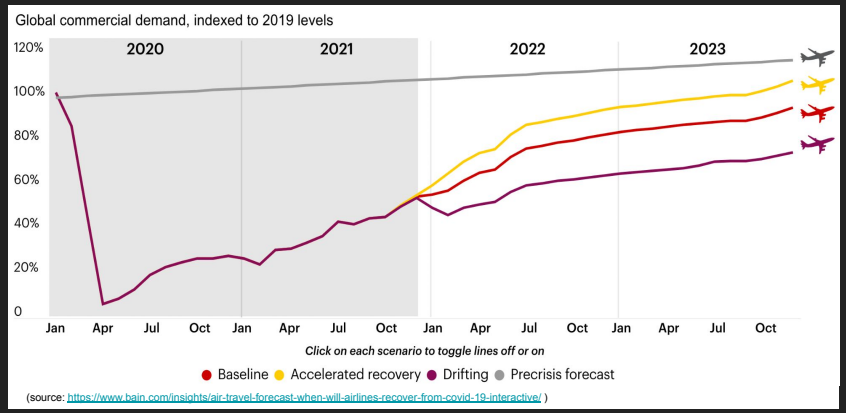
Building a Travel Information Portal using Reddit

Project 3 - Web Scraping and NLP Classification

CONTENTS

Background



Increased demand for air travel

Requirements and restrictions constantly change, difficult to find information

Good source of information - travellers who have experienced it before!

·01· PROBLEM STATEMENT

TAKEMYMONEY is a startup that aims to build a one-stop information portal to make trip planning hassle-free in the time of COVID-19. This would help different types of travellers easily find the information relevant for their style of travel.

In phase 1, TAKEMYMONEY will be focusing on budget and solo travel.

·02· PRE-PROCESSING

SCRAPING

2 subreddits were scraped using PushShift API

- r/solotravel
- r/shoestring (budget travel)

Important columns: title, selftext (body of each post)

At least 1,000 rows of data for each subreddit were obtained

- Excluding repeated posts
- Posts that were marked ‘removed’ or ‘deleted’
- Posts with empty selftext

Binarize subreddit column: 0 - solotravel, 1 - shoestring
Put both subreddits into one dataframe
Regex to clean urls, numbers, html special text (e.g. ampx200b) and whitespace
Dropped rows with moderator posts (author = automoderator or solotravelmods)

CLEANING

FEATURE CREATION

Concatenate both title and selftext into one column
Convert all text to lowercase
Lemmatize with spacy
Stem with snowball stemmer

EDA

Explore top words obtained from CountVectorizer and TFIDFVectorizer for both lemmatized and stemmed columns

Words (in red) that do not add value to model will be added to stopwords before modelling

- Words in subreddit (solo, travel, solotravel, shoestring)
- Common words (thank, really)
- Common english words (think, feel)
- Words that appeared in both subreddits (e.g. travel, city, country)

Unigrams

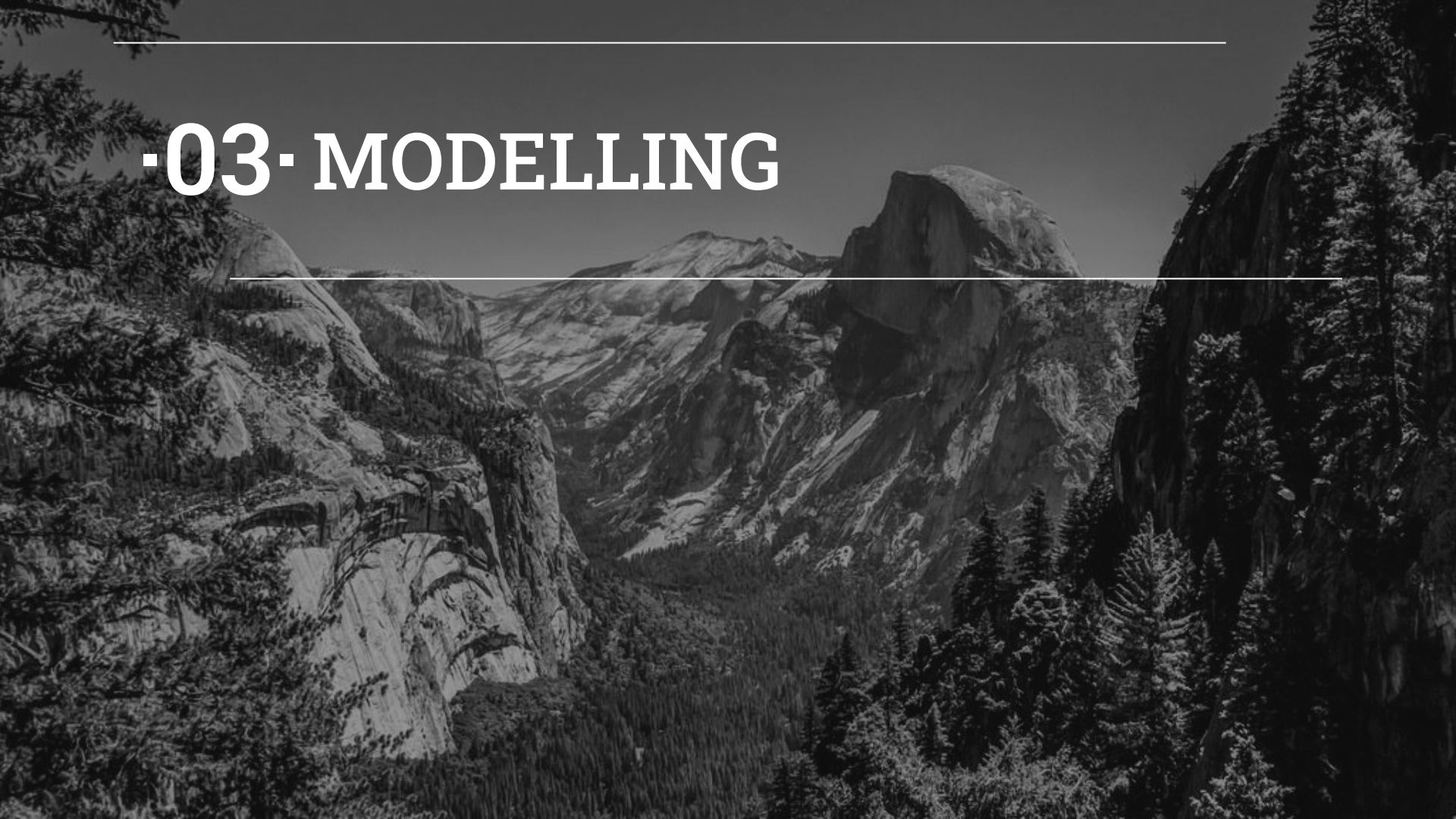
hostel
stay
budget
cheap
flight

travel
solo
week
city
country

Bigrams

national park
rent car
check bag
save money
street food

year old
just want
let know
thank advance



·03· MODELLING

MODEL PRE-PROCESSING

START

01

Stop Words

Add custom stop words to in-built english stop words in function

02

Baseline Accuracy

Solotravel - 50.8%
Shoestring - 49.2%

03

Train, Test, Split

X = lemmatized text column
Y = subreddit (1 or 0)

MODELLING

1. For each model, we ran a gridsearchCV on both vectorizers (Count and TFIDF)
2. Refitted models based on best parameters from gridsearch
3. Choose best model out of the 2 vectorizers

04

Logistic Regression

With TfidfVectorizer
Train / test scores: 0.83 / 0.72
Best parameters: C=0.2, max_features=3500

Multinomial Naive Bayes

With CountVectorizer
Train / test scores: 0.82 / 0.73
Best parameters: alpha=5.3, max_features=5000

05

06

Random Forest Classifier

With CountVectorizer
Train / test scores: 0.84 / 0.74
Best parameters: max_depth=25, n_estimators=300, max_features=3500

07

KNearestNeighbors

With TfidfVectorizer
Train / test scores: 0.71 / 0.72
Best parameters: leaf_size=5, n_neighbors=320, max_features=3500

Support Vectors

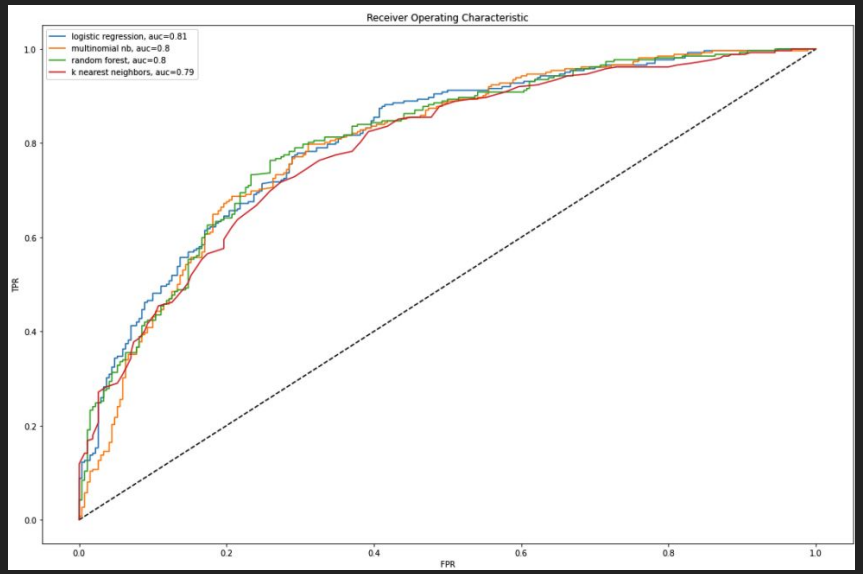
With TfidfVectorizer
Train / test scores: 0.90 / 0.73
Best parameters: C=1.1, max_features=5000

08

MODEL EVALUATION

Model	Count Vectorizer	TFIDF Vectorizer
Logistic Regression	Model 1	**Model 2**
Naive-Bayes	**Model 3**	Model 4
Random Forest	Model 5	**Model 6**
KNN	Model 7	**Model 8**
SVM	**Model 9**	Model 10

MODEL EVALUATION



Model	Vectorizer	Train Score	Test Score	Misclassification Rate
Logistic Regression	TFIDF	82.8%	72.2%	27.8%
Naive-Bayes	Count	81.9%	72.7%	27.3%
Random Forest	TFIDF	84.2%	73.9%	26.1%
KNN	TFIDF	71.2%	72.0%	28.0%
SVM	Count	89.7%	72.7%	27.3%

- ROC-AUC score used to evaluate our model, as we had balanced classes between for the 2 subreddits; misclassification would not have a large impact
- Majority of the models were severely overfit
- To go with KNearestNeighbors as it was a good compromise between model overfitting and ROC-AUC scores

KNearestNeighbors
(with TfidfVectorizer)

	Predicted r/solotravel	Predicted r/shoestring
Actual r/solotravel	200	70
Actual r/shoestring	79	183

Misclassified Text

1 **Truly misclassified** (put in r/shoestring)

I would love to go outside the country but I know there be just a boat load of restriction everywhere right now so I 'm look for idea of place or inspiration to go in the us . length of trip would be a long weekend to squeeze in between class

2 **Belongs in both**

I want to solo travel to an asian country that march for a week , at the beginning of my college vacation (we get four week of vacation) . I will finance the trip myself . however , my parent be concern with the possibility that covid restriction might be force at the last minute , and that I end up be stick there for week . they mention a similar case that happen in turkey and morocco . how can I convince they that this risk do not really represent a strong case against my trip ? I do agree that this be a risk , but if we start count all similar risk as a threat , then we would not travel at all . what do you guy think

3 **Posted in wrong subreddit** (in r/solotravel)

I be look for body , mind and spirit rejuvenation . clean , healthy eating , exercise , massage and maybe even some one on one therapy . I know place like this exist in india and other place far away from the state but I do not want to deal with + hour of fly each way for just a few week; budget would ideally be around $, for + day not include airfare

4 **Does not belong in either**

I want to travel to all these different country right and experience local place and all . but I be so bad with food . I have such a hard time try new food and I dislike so many thing that be really big in other country like rice . there ' a lot of country that rice be in almost every meal so I think you can see the issue . anyone else have this issue and what have you do about it ?

·04· CONCLUSION

CONCLUSION

Recommendations

- 2 subreddits might not be very different and could be combined into 1
- Grouping information portal by traveller type (solo or group) might be more useful
    - Possible relationship between solo and budget travel

Future Improvements

- Experiment with different train/test splits to reduce overfitting
- Modelling with XGBoost to see if it yields better results
- Explore relationship between types of travellers and what their considerations are
- Sub-categorizing posts into regions (e.g. Europe, SEA) to return more targeted information search


Questions?