Predicting severity of side effects of medicinal drugs marketed in the UK

General Assembly
Data Science Immersive
Capstone Project

Jocelyn Ho
26 February 2021

Background

Purpose

- To utilise data on side effects of drugs
  - Little is known about how data from the Yellow Card Scheme is used

Goal

- To predict whether certain types of side effects/ drugs are more likely to result in a severe or fatal side effect

Metrics

- Classification model predictions
- Mean cross-validated accuracy score

MHRA

- Executive Agency of the Department of Health and Social Care
- Acts on behalf of the Ministers
- Protect and promote public health and patient safety
- Ensure medicines and medical devices meet appropriate standards of safety, quality and efficacy

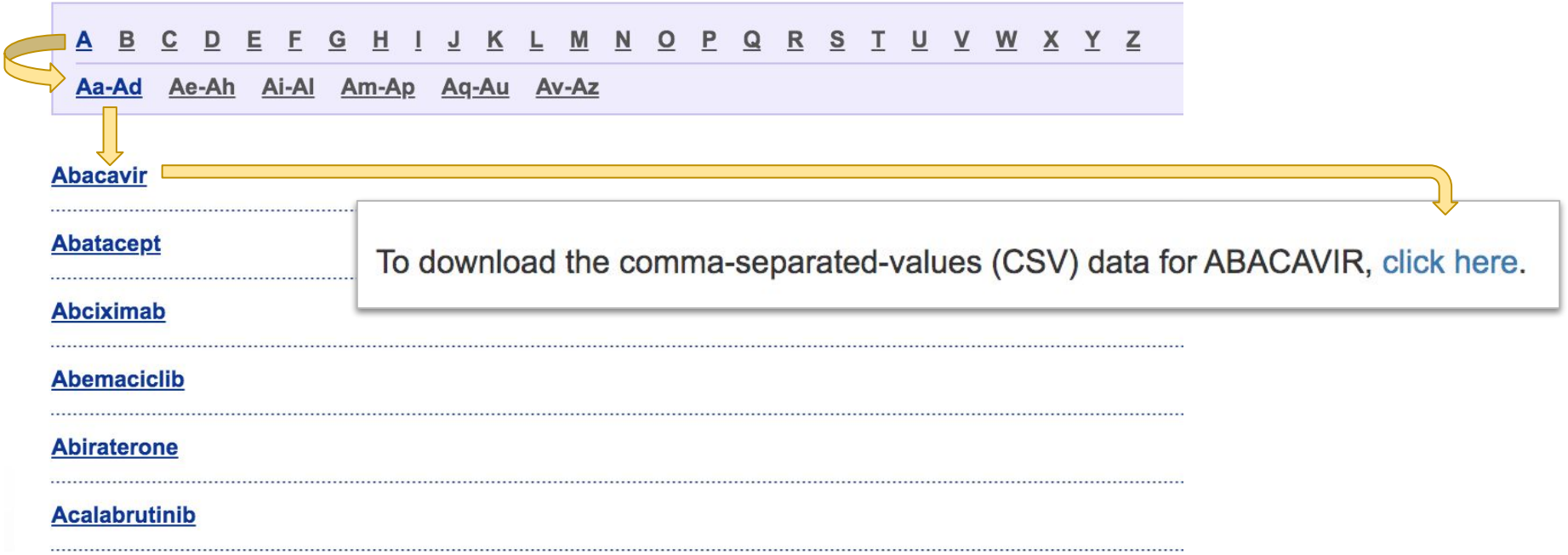
Yellow Card Scheme

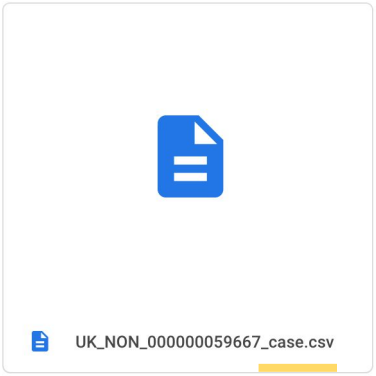
- Run by the MHRA
- The UK system for collecting and monitoring information on safety concerns
  - e.g. suspected side effects or adverse incidents
  - Includes medicines and medical devices
- Relies on voluntary reporting
- Reported by health professionals and the public (patients, carers, parents)

Data Source 1:
Yellow Card Interactive Drug Analysis Profiles

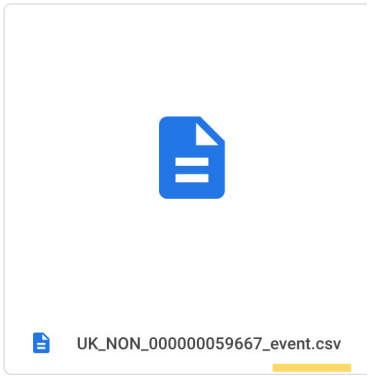
Web Scrape - Selenium

1. Click onto each alphabet
2. Click onto each subset of alphabet
3. Click onto each drug
4. Get unique yellow card ID for each drug
5. Download zip file for each drug
6. Unzip all folders

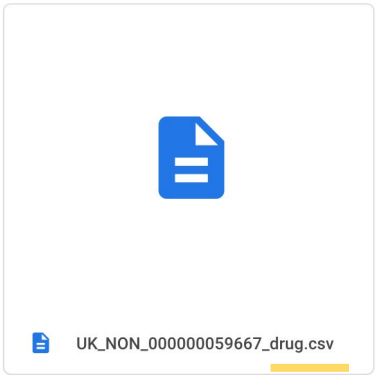




Summary of each report for the drug, includes patient demographics



Adverse reactions within each report for the drug (multiple events per case)



(Not used)
Method of administration for the drug in each report

Yellow Card
Making medicines safer

Create dataset - SQL:

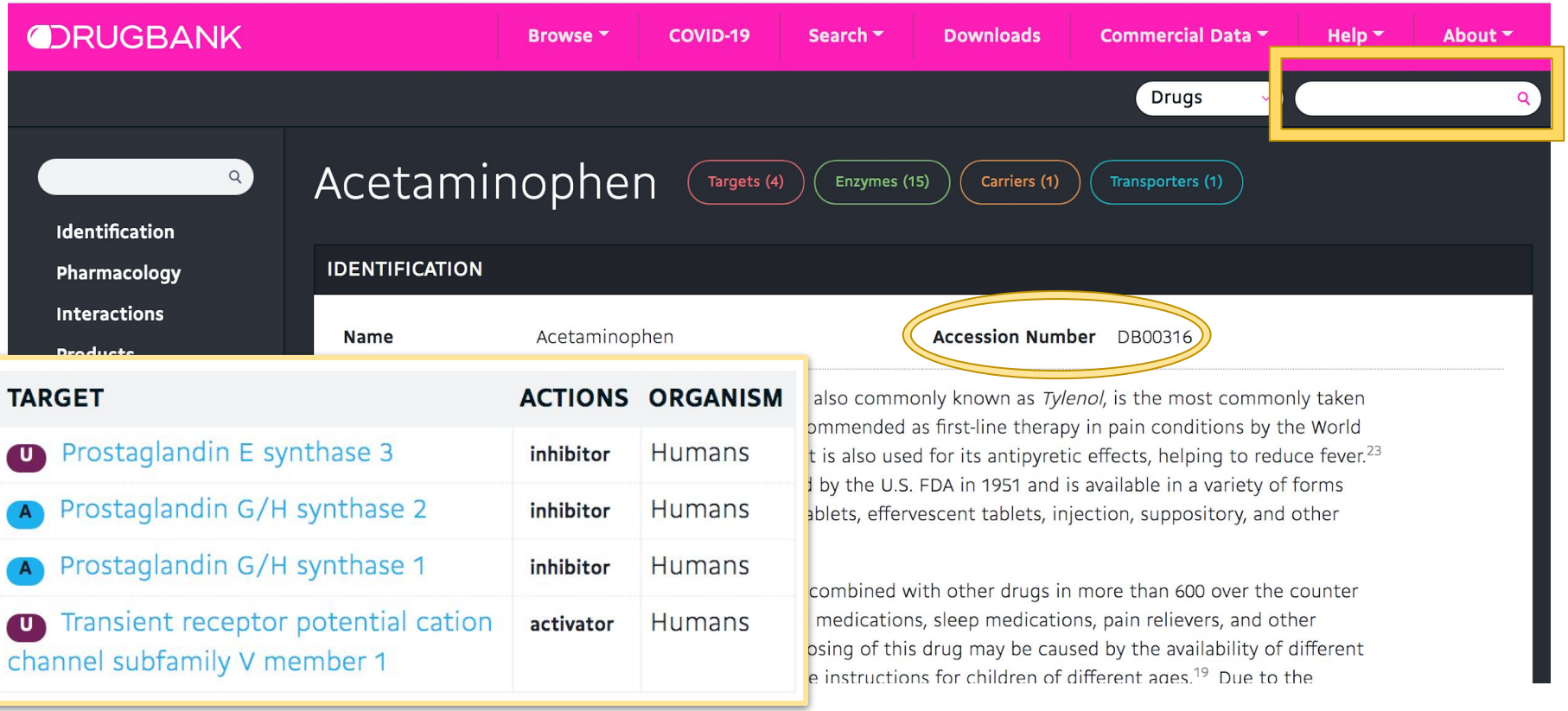
1. Join case to events for each drug on report number
2. Add yellow card id to each row
3. Append combined dataframe to a new SQL table
4. Repeat for all drugs

ADR	PT	SOC_ABBREV	FATAL_YN	SEX	AGE_10	RECVD_YEAR	SENDER_TYPE	CONSUMER_YN	HCP_YN	NONSERIOUS_SERIOUS_FATAL_NSF
1	Chills	Genrl	N	Female	40.0	2003	Indirect	N	Y	S
1	Malaise	Genrl	N	Female	40.0	2003	Indirect	N	Y	S
1	Vomiting	Gastr	N	Female	40.0	2003	Indirect	N	Y	S
2	Oral candidiasis	Infec	N	Female	50.0	2003	Direct	N	Y	N
3	Photosensitivity reaction	Skin	N	Female	60.0	2003	Direct	N	Y	S

Dataframe 1: Yellow Card

Data Source 2: Drugbank

1. Search each drug
2. Obtain accession number and extra information on drug targets



```
Int64Index: 1987 entries, 0 to 2328
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   drug      1987 non-null   object
 1   yc_id     1987 non-null   int64
 2   target    1568 non-null   object
 3   drug_cat  675 non-null    object
 4   db_id     1987 non-null   object
dtypes: int64(1), object(4)
memory usage: 93.1+ KB
```

Dropped drug_cat column due to insufficient information

Dataframe 2: DrugBank

	drug	yc_id	target	db_id
0	abacavir	40046536	['Reverse transcriptase/RNaseH', 'HLA class I ...	DB01048
1	abatacept	561378321	['T-lymphocyte activation antigen CD80', 'T-ly...	DB01281
2	abciximab	231911819	['Integrin beta-3', 'Integrin alpha-IIb', 'Low...	DB00054
3	abemaciclib	369408139	['Cyclin-dependent kinase 4', 'Cyclin-dependen...	DB12001
4	abiraterone	968368347	['Steroid 17-alpha-hydroxylase/17,20 lyase']	DB05812

Data Source 3:
UCI dataset - Drugs.com reviews

Web scrape - Selenium

1. Get unique drugNames (brand names)
2. Find active ingredients of each product

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

	drugName	condition	review	rating	date	usefulCount
163740	Mirtazapine	Depression	"I've tried a fe	10	February 28, 2012	22
206473	Mesalamine	Crohn's Disease, Mai	"My son has Crohn&#	8	May 17, 2009	17
159672	Bactrim	Urinary Tract Infectio	"Quick reduction of s	9	September 29, 2017	3
3929	Contrave	Weight Loss	"Contrave combines	9	March 5, 2017	35
9 763	Cyclafem 1 / 35	Bi h Control	"I have been on this	9	October 22, 2015	4
208087	Zyt	Keratosis	"4 days in on first 2 w	4	July 3, 2014	13
215892	Copper	Birth Control	"I've had the c	6	June 6, 2016	1
169852	Amitriptyline	Migraine Prevention	"This has been great	9	April 21, 2009	32
23295	Methadone	Opiate Withdrawal	"Ive been on Methac	7	October 18, 2016	21
71428	Levora	Birth Control	"I was on this pill for	2	April 16, 2011	3

BRAND name

DRUG name/ active ingredient

Extract necessary information

1. Groupby DrugBank ID
2. Obtain unique conditions
3. Calculate mean rating for each product

	drugName	condition	review	rating	d
0	Mirtazapine	Depression	"I've tried a fe	10	
3	Mesalamine	Crohn's Disease, Mai	"My son has Crohn&#	8	
2	Bactrim	Urinary Tract Infectic	"Quick reduction of s	9	Su
3	Contrave	Weight Loss	"Contrave combines		
8	Cyclafem 1 / 35	Birth Control	"I have been on this		
7	Zyclara	Keratosis	"4 days in on first 2 w		
2	Copper	Birth Control	"I've had the c		
2	Amitriptyline	Migraine Prevention	"This has been great		
5	Methadone	Opiate Withdrawal	"Ive been on Methac		
8	Levora	Birth Control	"I was on this pill for		

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

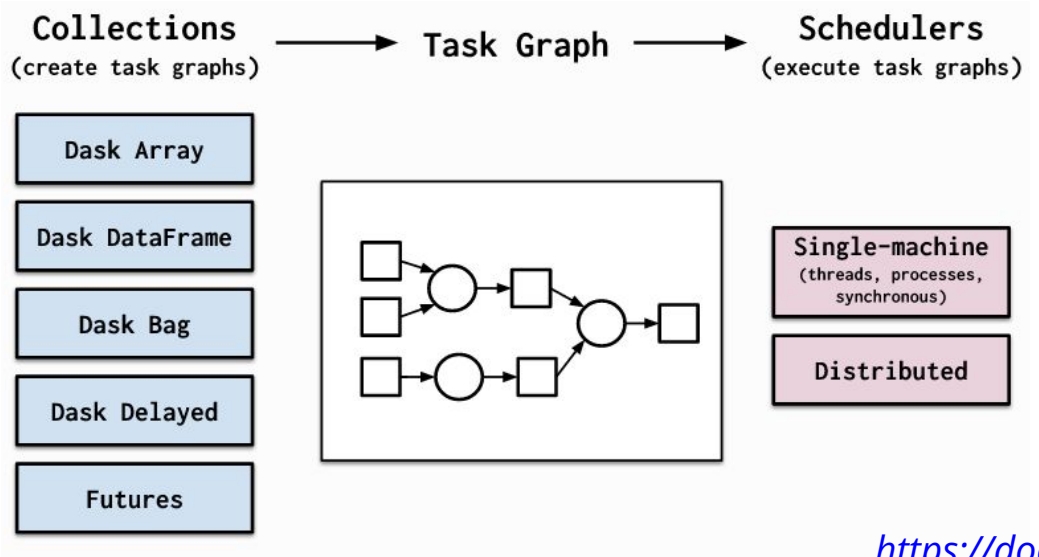
Dataframe 3: Drugs.com reviews

	db_id	drug	condition	review	rating	n_reviews
0	DB00002	['erbitux' 'cetuximab']	['colorectal cancer' 'head and neck cancer' 's...	['"i have stage 4 colon cancer with liver mets...	6.875000	16
1	DB00003	['dornase alfa' 'pulmozyme']	['cystic fibrosis']	['"my outcome/experience with pulmozyme was go...	10.000000	2
2	DB00005	['etanercept']	['rheumatoid arthritis' 'psoriatic arthritis' ...	['"took enbrel with methotrexate for four yea...	8.246305	203
3	DB00007	['leuprolide' 'lupron depot-ped' 'lupron' 'eli...	['uterine fibroids' 'prostate cancer' 'endomet...	['"i was so skeptical about these shots, scare...	6.179487	273
4	DB00008	['pegasys' 'peginterferon alfa-2a']	['hepatitis c']	['"i was treated with this twice on it's ...	6.791667	24

Merging data - Dask

Dask - Introduction

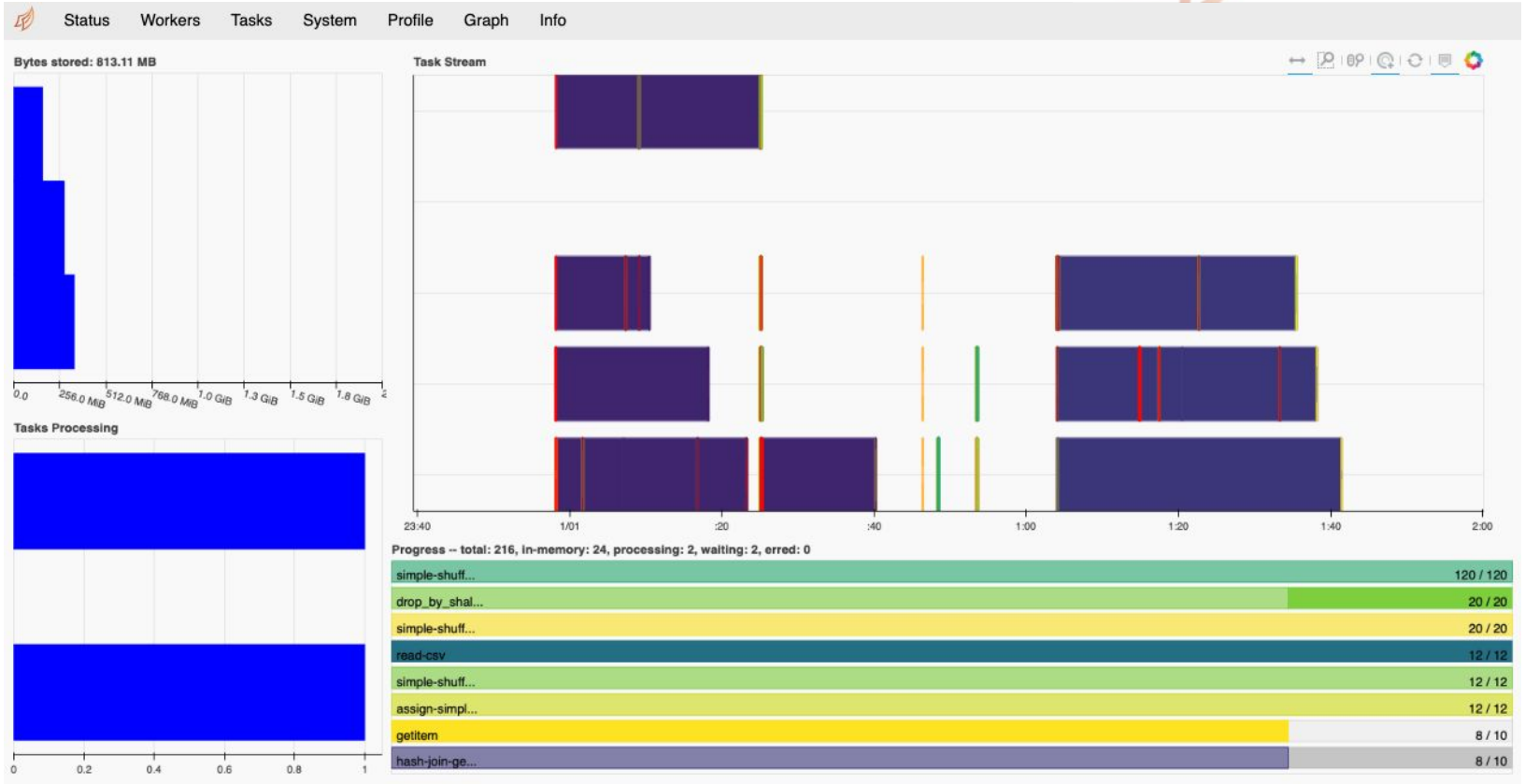
- Composed of two parts:
  - Dynamic task scheduling
  - ‘Big data’ collection - parallel arrays, dataframes etc.



https://docs.dask.org/en/latest/

Dask - Uses and advantages

- Enable efficient parallel computations on single machines
  - Leverages multi-core CPUs and streams data efficiently from disk
  - Allows to swap out the cluster for single-machine schedulers
- Avoids excess memory use:
  - Finds ways to evaluate computations in a low-memory footprint when possible
  - Pulls in chunks of data from disk, does necessary processing, throws away intermediate values asap
- Able to perform computations on moderately large datasets (100GB+) even on relatively low-power laptops.
- Requires no configuration and no setup, installed by default

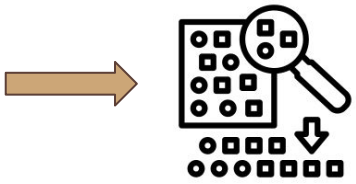

Status
Workers
Tasks
System
Profile
Graph
Info
Bytes stored: 813.11 MB
0.0
256.0 MiB
512.0 MiB
768.0 MiB
1.0 GiB
1.3 GiB
1.5 GiB
1.8 GiB
Task Stream
23:40
1/01
:20
:40
1:00
1:20
1:40
2:00
Tasks Processing
0
0.2
0.4
0.6
0.8
1
Progress -- total: 216, in-memory: 24, processing: 2, waiting: 2, erred: 0
simple-shuff... 120 / 120
drop_by_shal... 20 / 20
simple-shuff... 20 / 20
read-csv 12 / 12
simple-shuff... 12 / 12
assign-simpl... 12 / 12
getitem 8 / 10
hash-join-ge... 8 / 10

Dask - issues

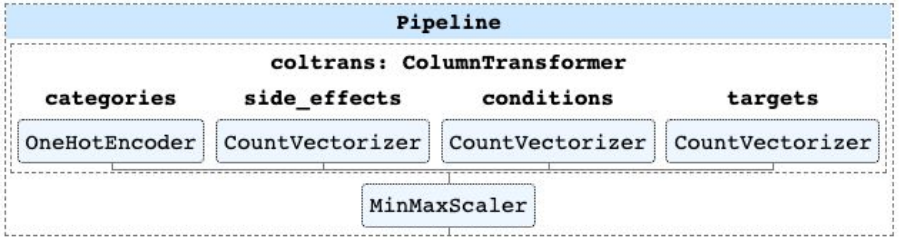
- Did not close down clients (workers) properly
  - Memory leaked
  - Keep running out of application memory when running models
- Still not enough RAM to process dataset after successfully merging
  - Dropped reviews column
  - Instead used the number of reviews each product received as an indicator
  - Sample 10% of dataset to run models

Overall Approach

Dataset

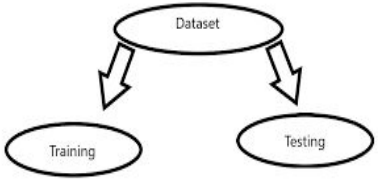


Sample

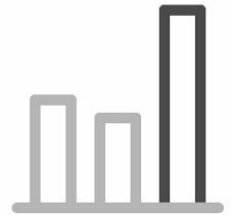


Preprocessing

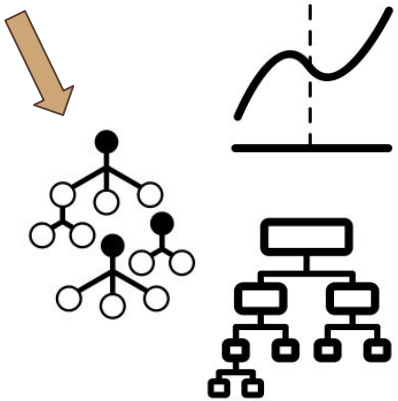
EDA



Train-test split



Under-sample training set



Modeling

Final Dataset

1,755,219

Number of rows

16

Columns
(Prior to processing)

3

Classes within target variable

3

Text columns

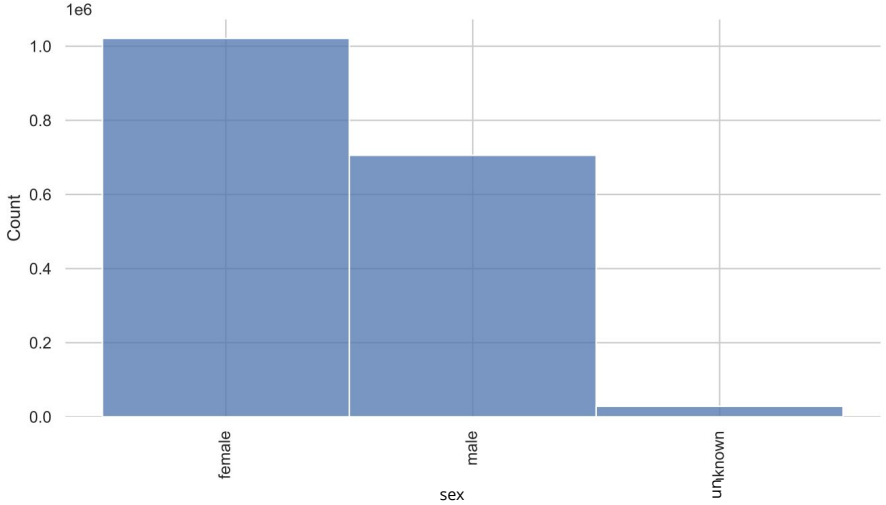
6

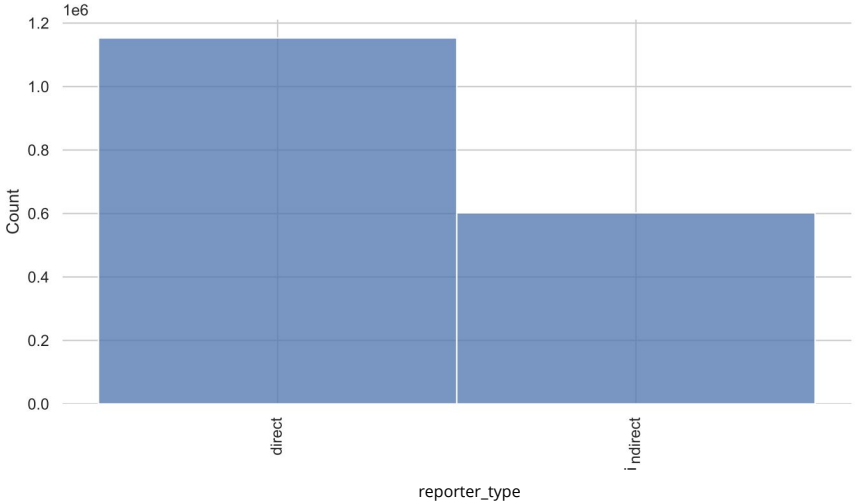
Categorical columns

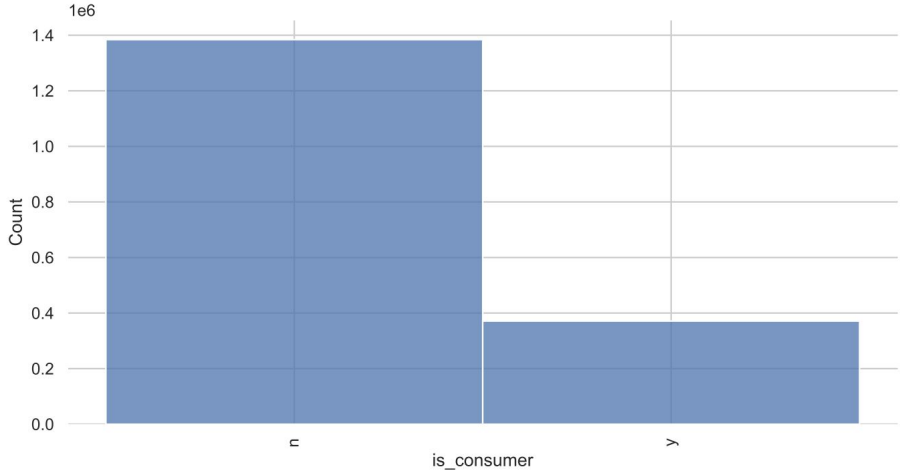
4

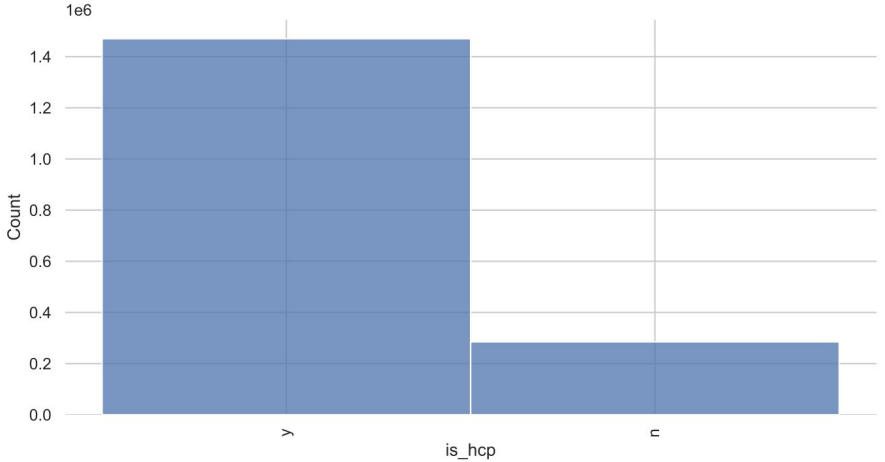
Numeric columns

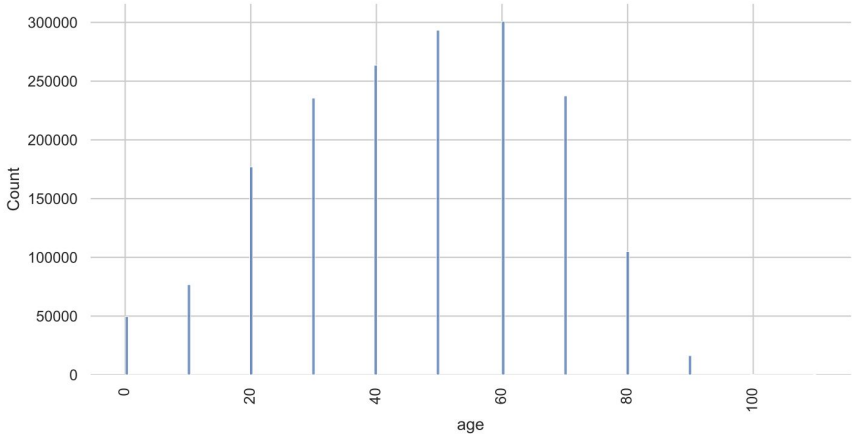
EDA

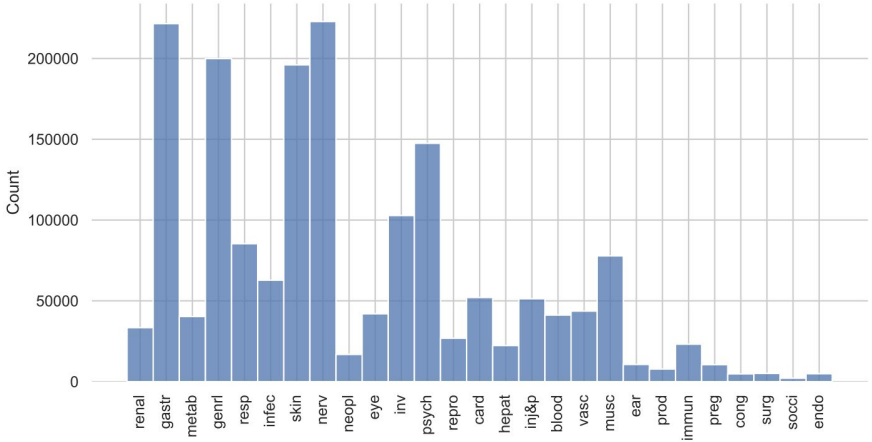

1e6
1.0
0.8
0.6
0.4
0.2
0.0
Count
female
male
unknown
sex

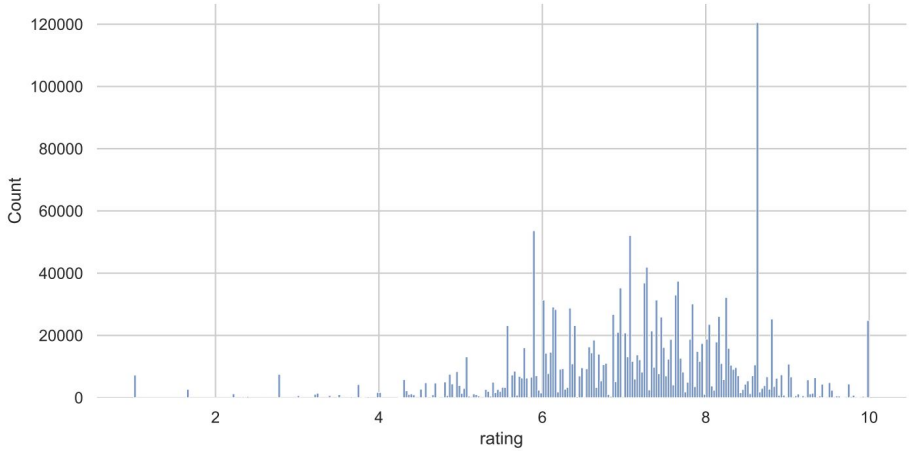

1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count
direct
indirect
reporter_type

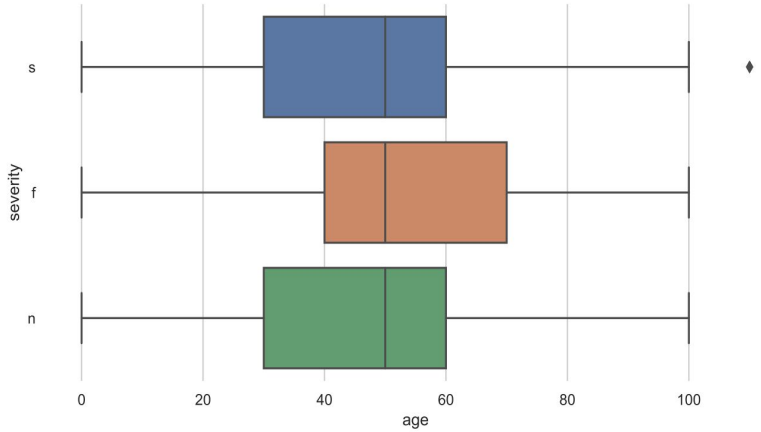

1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count
n
y
is_consumer

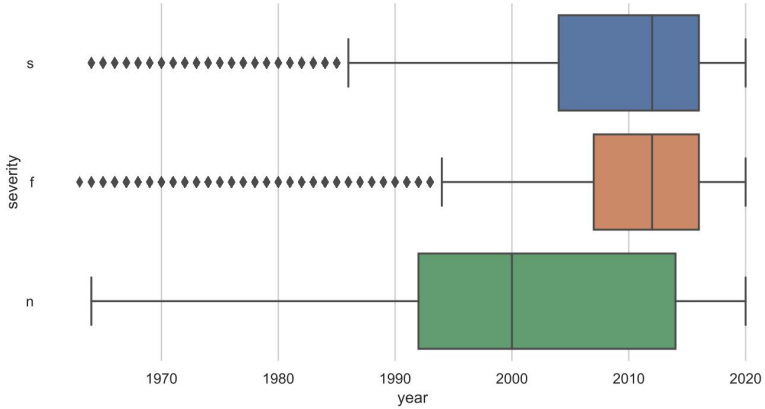

1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count
y
n
is_hcp

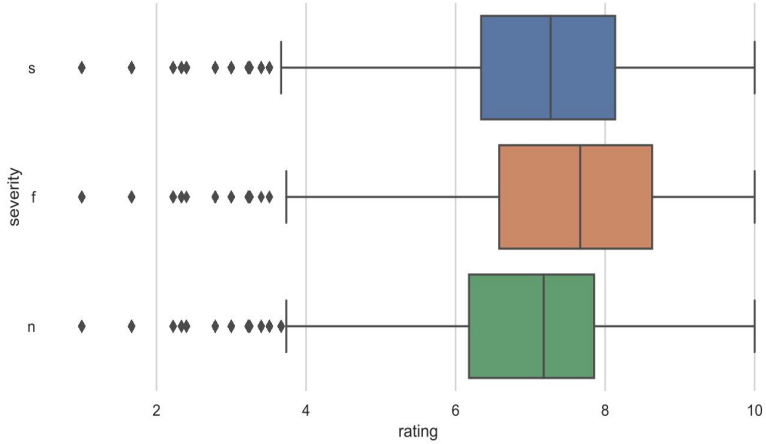

Count
300000
250000
200000
150000
100000
50000
0
0
20
40
60
80
100
age

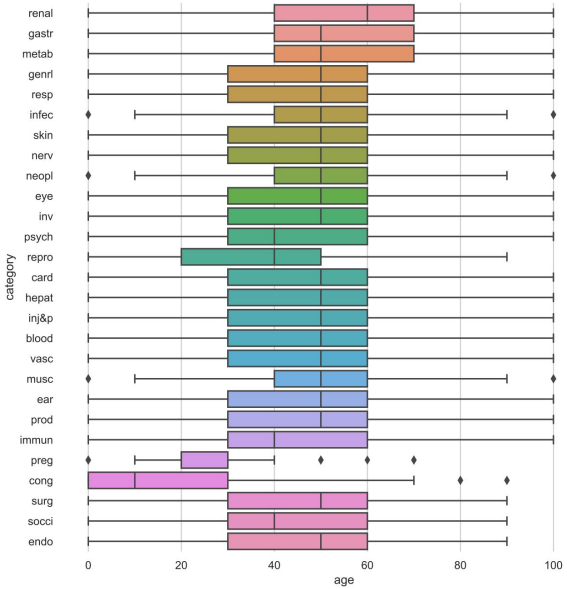

Count
200000
150000
100000
50000
0
renal
gastr
metab
genrl
resp
infec
skin
nerv
neopl
eye
inv
psych
repro
card
hepat
inj&p
blood
vasc
musc
ear
prod
immun
preg
cong
surg
socci
endo

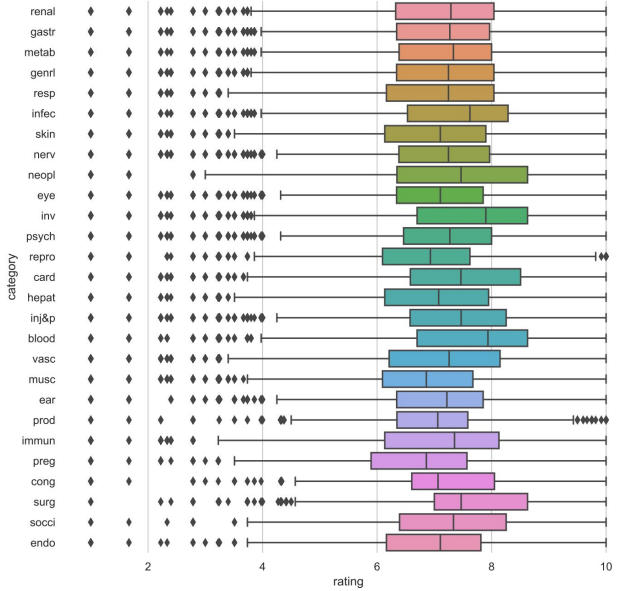

Count
120000
100000
80000
60000
40000
20000
0
2
4
6
8
10
rating

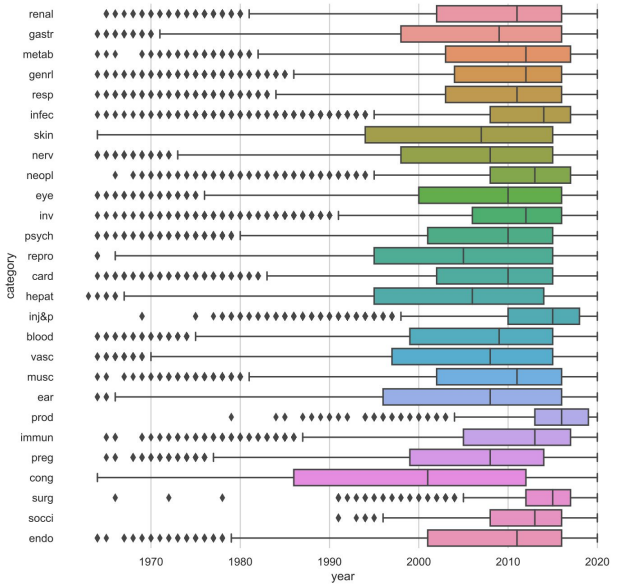

s
f
n
severity
0
20
40
60
80
100
age


s
f
n
severity
1970
1980
1990
2000
2010
2020
year

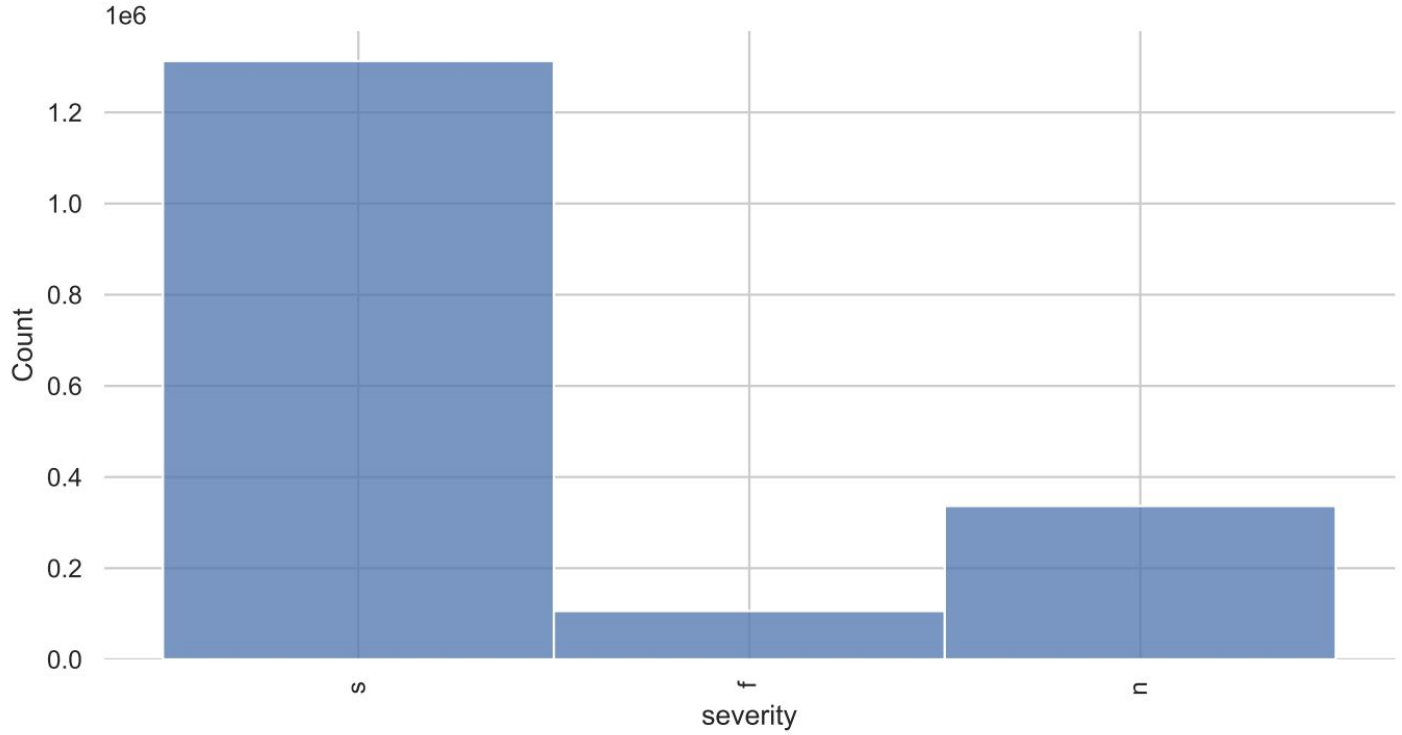

s
f
n
severity
2
4
6
8
10
rating


renal
gastr
metab
genrl
resp
infec
skin
nerv
neopl
eye
inv
psych
repro
card
hepat
inj&p
blood
vasc
musc
ear
prod
immun
preg
cong
surg
socci
endo
category
0
20
40
60
80
100
age


renal
gastr
metab
genrl
resp
infec
skin
nerv
neopl
eye
inv
psych
repro
card
hepat
inj&p
blood
vasc
musc
ear
prod
immun
preg
cong
surg
socci
endo
category
2
4
6
8
10
rating


renal
gastr
metab
genrl
resp
infec
skin
nerv
neopl
eye
inv
psych
repro
card
hepat
inj&p
blood
vasc
musc
ear
prod
immun
preg
cong
surg
socci
endo
category
1970
1980
1990
2000
2010
2020
year

Histogram showing the distribution of target variable classes



NLP

Natural Language Processing

Drug targets

subunit alpha	364
muscarinic acetylcholine	238
adrenergic alpha	158
growth factor	142
binding protein	140
calcium channel	130
channel subunit	130
voltage dependent	127
sodium dependent	120
channel subfamily	117
beta adrenergic	114
penicillin binding	113
protein kinase	111
type calcium	107
subfamily member	107
dependent type	103
voltage gated	101
type opioid	100
prostaglandin synthase	97
aminobutyric acid	96

Drug indications (conditions)

influenza prophylaxis	245
nausea vomiting	119
birth control	97
vomiting pregnancy	83
nausea vomiting pregnancy	83
blood pressure	76
high blood pressure	76
high blood	76
rheumatoid arthritis	73
diabetes type	59
tract infection	58
allergic rhinitis	49
nasal congestion	46
uterine bleeding	46
abnormal uterine	45
abnormal uterine bleeding	45
urinary tract	44
urinary tract infection	39
crohn disease	37
bipolar disorder	35

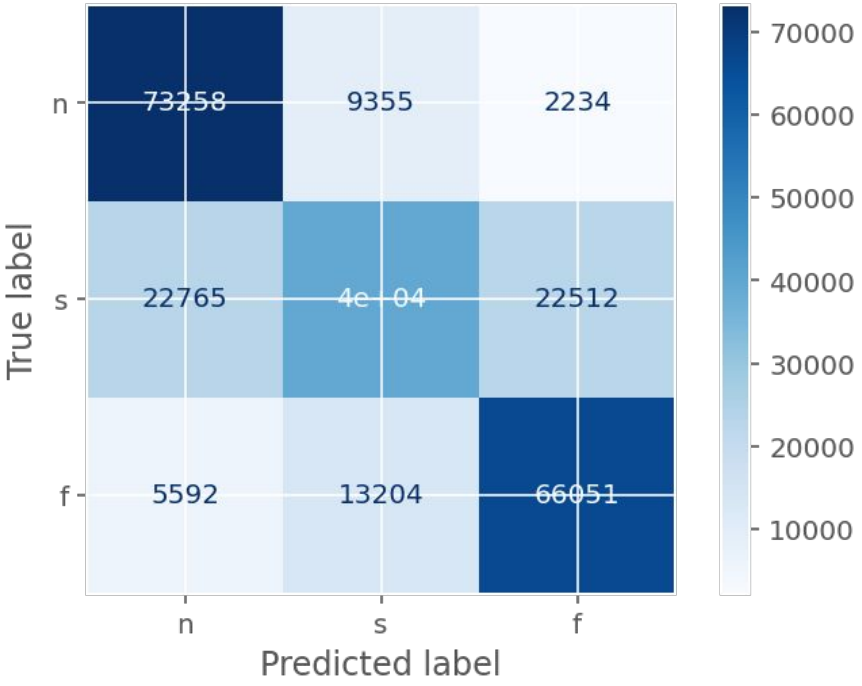
Drugs.com reviews

blood pressure	63.007784
years ago	59.812564
started taking	59.187928
birth control	50.596942
twice day	43.769524
years old	43.570576
times day	41.860535
weight gain	39.980575
year old	38.192683
works great	35.673650
doctor prescribed	34.085749
feel better	33.624177
months ago	32.096512
dry mouth	30.421098
highly recommend	29.078865
stopped taking	28.799457
mood swings	27.933365
taking medication	27.421454
worked great	27.337887
went away	25.870941

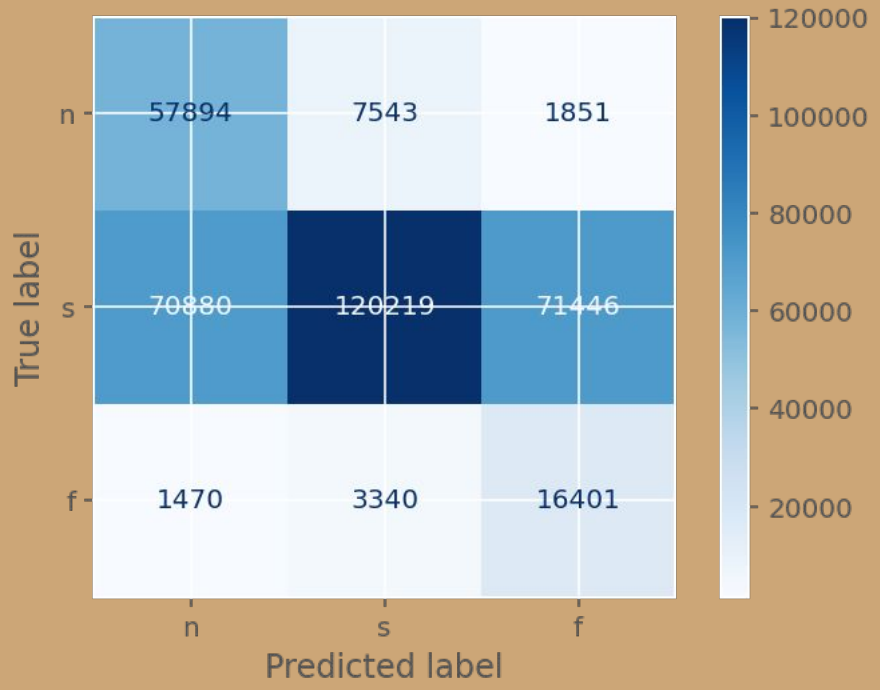
Data Modelling

Model	Best estimators/ parameters	Best Score	Training Score	Test Score
Random Forest	max_depth = 20, max_features = 90, n_estimators = 500	0.693	0.729	0.507
Decision Tree Classifier	max_depth = 20, max_features = 70	0.639	0.622	0.524
AdaBoost	max_depth = 3, max_features = 0.3, n_estimators = 100	0.695	0.699	0.544
Hist Gradient Boosting Classifier	max_depth = 3	0.682	0.699	0.544
Naive Bayes (Multinomial)	alpha = 0.0001	0.617	0.616	0.456
Logistic Regression	'C': 100.0, 'multi_class': 'ovr', 'penalty': 'l2', 'solver': 'saga'	**0.699**	0.703	0.554
Baseline accuracy: 74.8% (0.747897)				

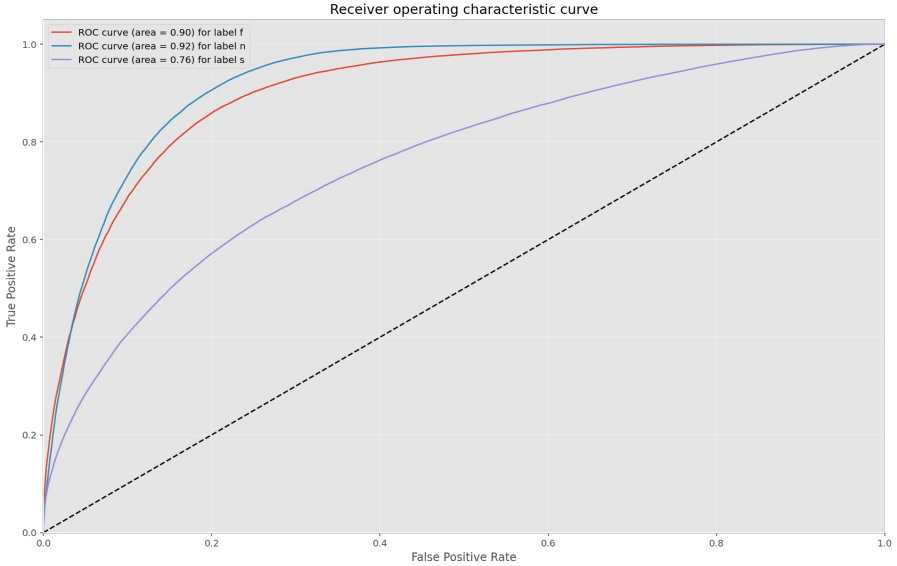
Model:
Logistic Regression

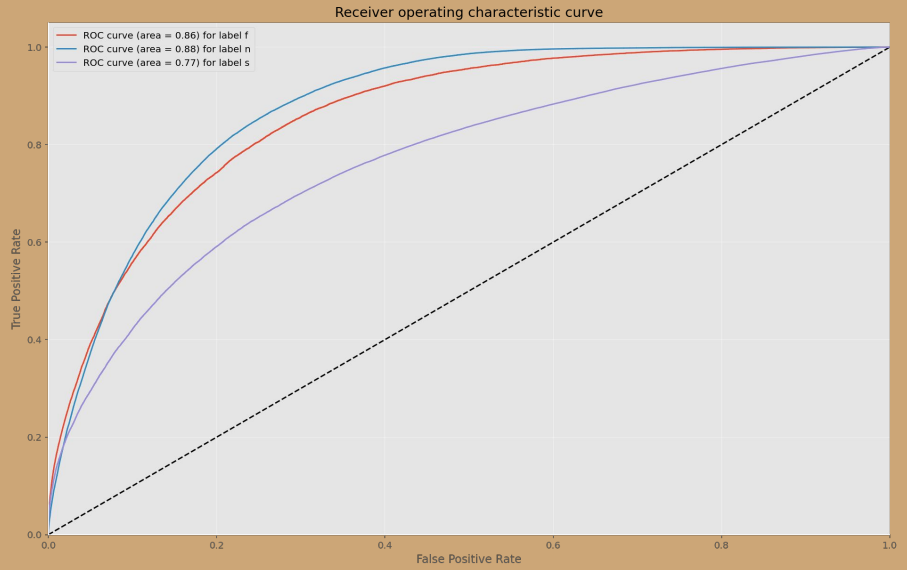


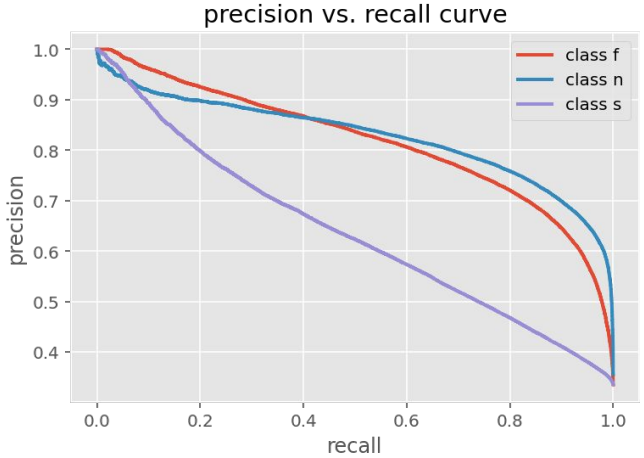
	precision	recall	f1-score	support
f	0.78	0.73	0.75	90797
n	0.86	0.72	0.79	101615
s	0.47	0.64	0.54	62129
accuracy			0.70	254541
macro avg	0.70	0.70	0.69	254541
weighted avg	0.74	0.70	0.71	254541

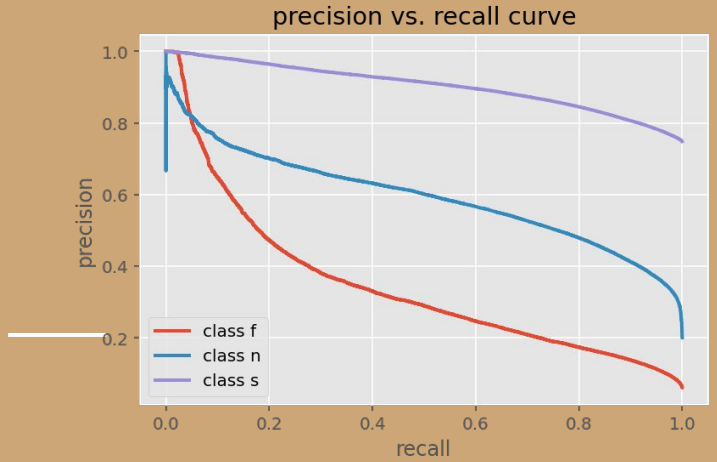


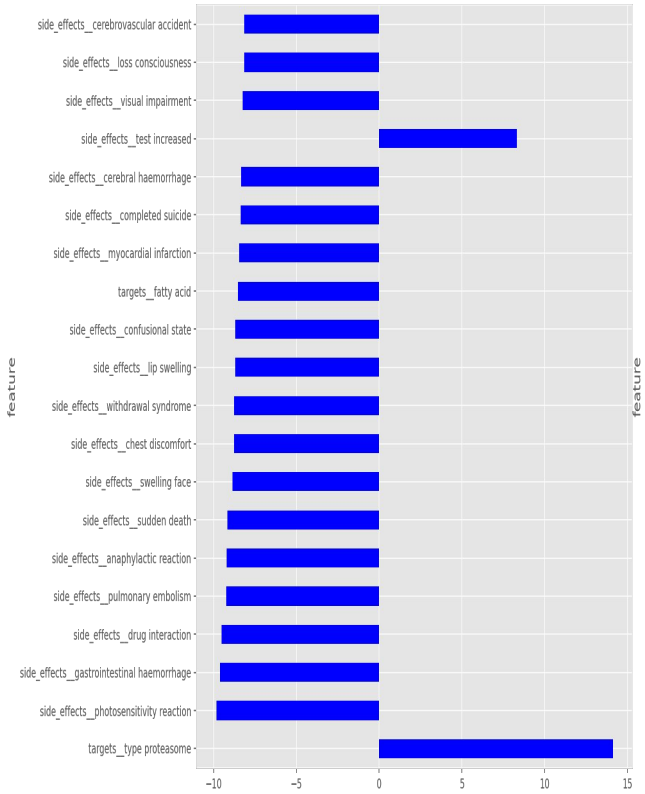
	precision	recall	f1-score	support
f	0.77	0.18	0.30	89698
n	0.86	0.44	0.59	130244
s	0.46	0.92	0.61	131102
accuracy			0.55	351044
macro avg	0.70	0.51	0.50	351044
weighted avg	0.69	0.55	0.52	351044


Receiver operating characteristic curve
ROC curve (area = 0.90) for label f
ROC curve (area = 0.92) for label n
ROC curve (area = 0.76) for label s
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

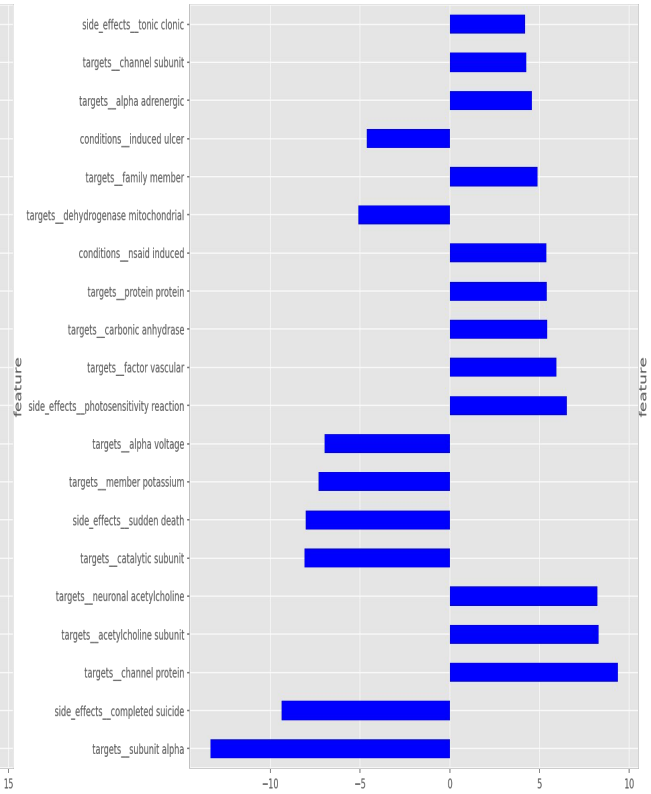

Receiver operating characteristic curve
ROC curve (area = 0.86) for label f
ROC curve (area = 0.88) for label n
ROC curve (area = 0.77) for label s
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


precision vs. recall curve
class f
class n
class s
precision
recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
1.0

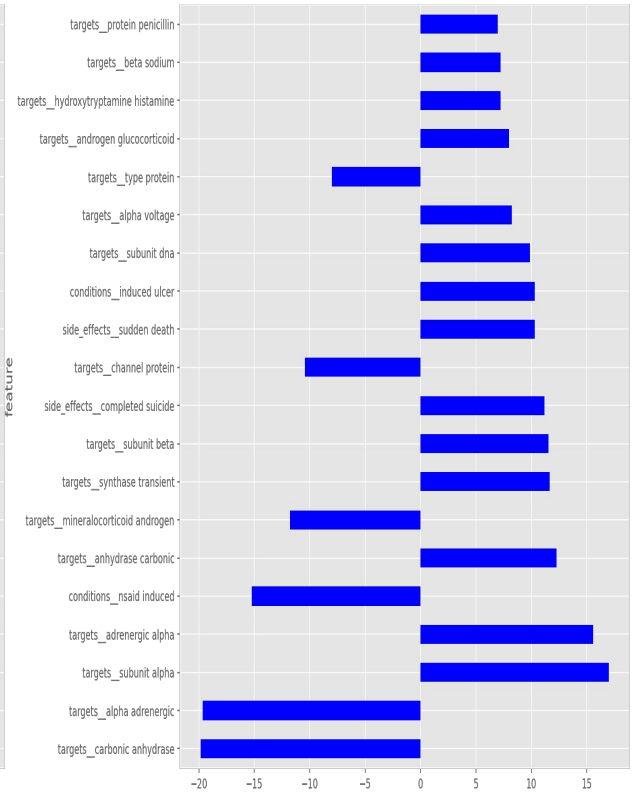

precision vs. recall curve
class f
class n
class s
precision
recall
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0



non-serious



serious



fatal

Limitations

Limitations - Data

Yellow Card reports:

- Relies on voluntary reporting
  - May not truly represent distribution of side effect severity in real life

- Naturally skewed class distribution
  - Low fatal cases:
    - Marketing license approval only if drug safety is established after extensive clinical trials
    - Difficult to pinpoint cause of death - comorbidities/ complications/ actual side effects
  - Low non-serious cases:
    - Less likely to report non-serious side effects

- Subjectivity/ ambiguity in determining severity of side effects
  - Can be serious if reporter considers it serious
  - Criteria provided - only 1 out of 6 options cover the class ‘fatal’

MHRA Yellow Card definitions

"The seriousness criteria… were determined by a working group of the Council for International Organizations of Medical Sciences (CIOMS) and are defined as 6 possible categories… reporters to select one of the following criteria by ticking the appropriate box on the Yellow Card.

(1) patient died due to reaction

(2) life threatening

(3) resulted in hospitalisation or prolonged inpatient hospitalisation

(4) congenital abnormality

(5) involved persistent or significant disability or incapacity

(6) if the reaction was deemed medically significant.

In addition to this, seriousness of reaction terms has also been defined by the MHRA in our medical dictionary. Therefore an ADR report can be serious because the reporter considers the reaction to be serious or because the reaction term itself is considered serious in our medical dictionary."

https://yellowcard.mhra.gov.uk/definitions

Limitations

Drugs.com reviews:

- Reviews only up to 2018, no recent reviews
- US brands

DrugBank:

- Information on US-marketed drugs, not the UK
- May not include all the information for each drug
  - e.g. drug classes and drug targets

General:

- Insufficient computational power/ RAM
- Insufficient time
- Fewer data points after under-sampling training set
  - Insufficient data to train model properly

Future work

Data:

- Incorporate drug classes (e.g. ATC codes from WHO) *(categories based on what conditions the drug is used for)*
- Use ATC codes as unique ID instead of DrugBank number
- Include actual reviews for more NLP/ sentiment analysis
- Dummify drug target column instead of NLP

Modelling:

- Use different under-/over-sampling methods and ratios
- Increase sample size
- Run more extensive models
- Fine tune parameters

Others:

- Utilise AWS