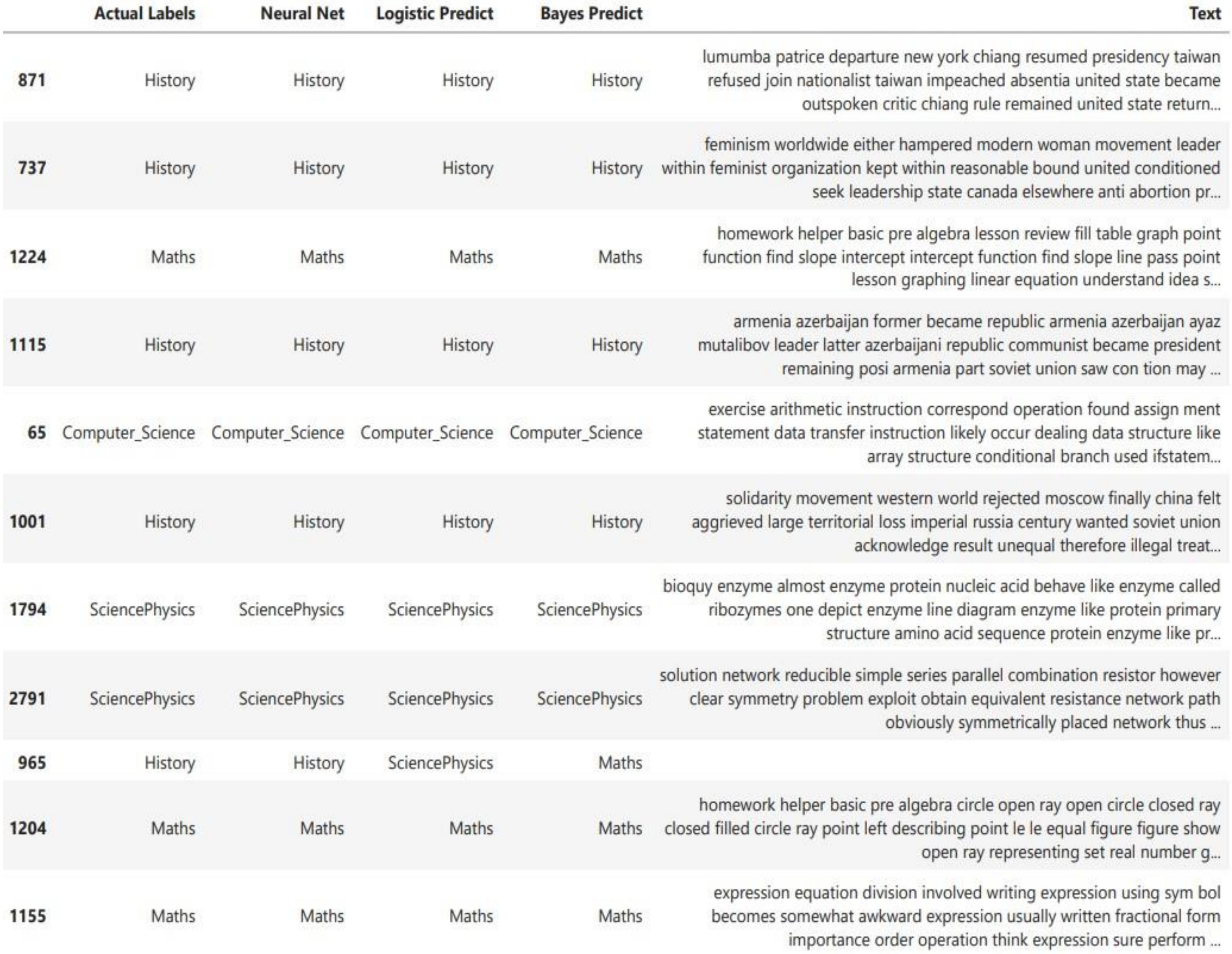
Text Models:

MyGitHub

Text data is everywhere and text data is informative. Text data is so incredibly useful that statistical models can quickly learn how humans communicate. Customized statistical algorithms that can parse over millions of texts and do superb classification in a matter of minutes. Customized, state of the art text models can help firms leverage informative text data for predictive/administrative purposes. It can make firms be more agile to new information/sentiment. For example, in last analyst role, I had to categorized hundreds of claims by their text descriptions. This was laborious and long. Modern text models can do all of this with superb precision. Below I show the results of statistical models! These models/algorithms below were tested on unseen data so test would be unbias. All of them had 90+% accuracy!

1) Subject classification had 98% accuracy
2) Sentiment model 92% accuracy
3) News category classification had 92.2% accuracy
4) Used Cars Prediction R^2 = 90%
5) Online/PDF Table Extractor

Subject classification:

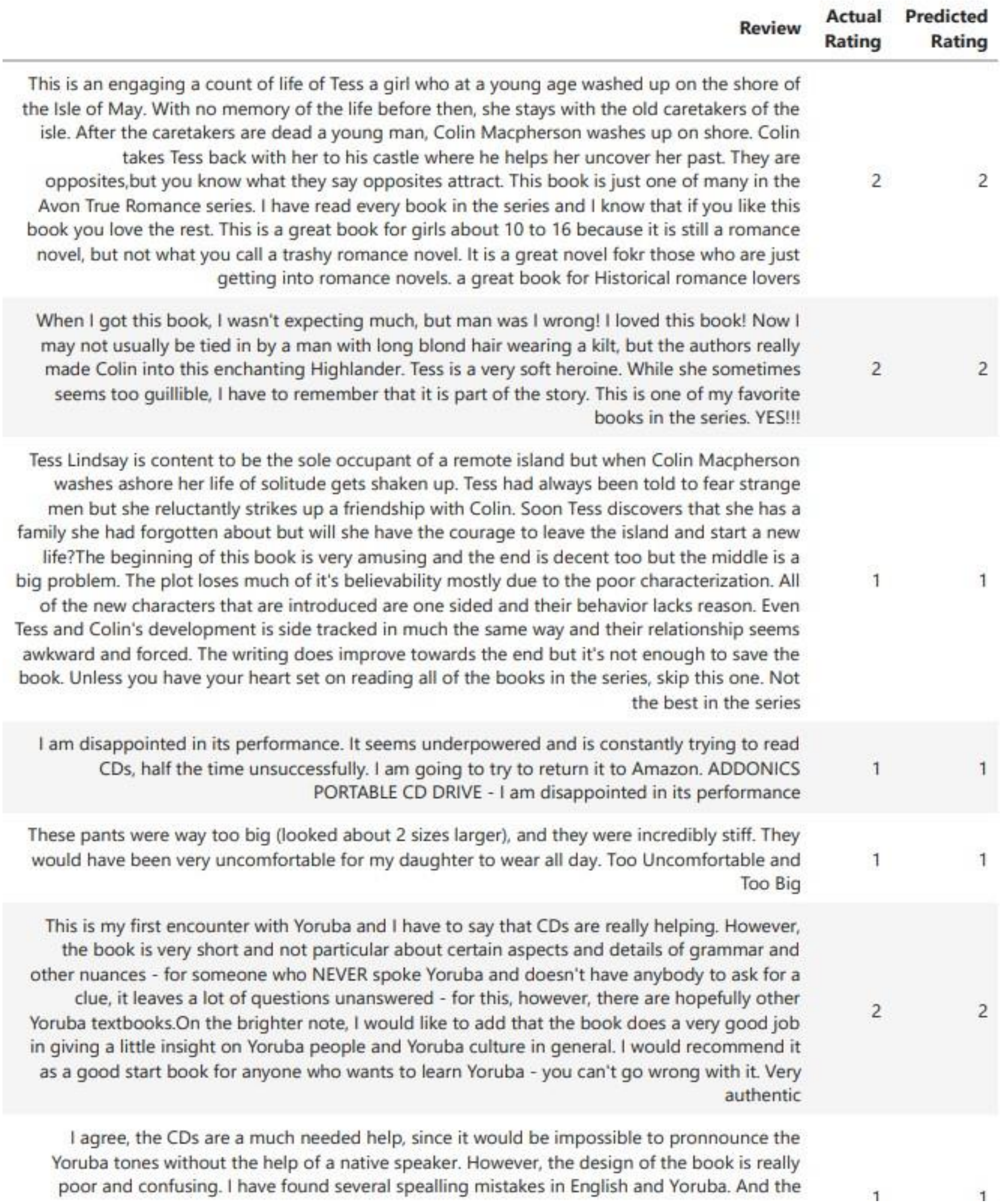


	Actual Labels	Neural Net	Logistic Predict	Bayes Predict	Text
871	History	History	History	History	lumumba patrice departure new york chiang resumed presidency taiwan refused join nationalist taiwan impeached absentia united state became outspoken critic chiang rule remained united state return...
737	History	History	History	History	feminism worldwide either hampered modern woman movement leader within feminist organization kept within reasonable bound united conditioned seek leadership state canada elsewhere anti abortion pr...
1224	Maths	Maths	Maths	Maths	homework helper basic pre algebra lesson review fill table graph point function find slope intercept intercept function find slope line pass point lesson graphing linear equation understand idea s...
1115	History	History	History	History	armenia azerbaijan former became republic armenia azerbaijan ayaz mutalibov leader latter azerbaijani republic communist became president remaining posi armenia part soviet union saw con tion may ...
65	Computer_Science	Computer_Science	Computer_Science	Computer_Science	exercise arithmetic instruction correspond operation found assign ment statement data transfer instruction likely occur dealing data structure like array structure conditional branch used ifstatem...
1001	History	History	History	History	solidarity movement western world rejected moscow finally china felt aggrieved large territorial loss imperial russia century wanted soviet union acknowledge result unequal therefore illegal treat...
1794	SciencePhysics	SciencePhysics	SciencePhysics	SciencePhysics	bioquy enzyme almost enzyme protein nucleic acid behave like enzyme called ribozymes one depict enzyme line diagram enzyme like protein primary structure amino acid sequence protein enzyme like pr...
2791	SciencePhysics	SciencePhysics	SciencePhysics	SciencePhysics	solution network reducible simple series parallel combination resistor however clear symmetry problem exploit obtain equivalent resistance network path obviously symmetrically placed network thus ...
965	History	History	SciencePhysics	Maths	
1204	Maths	Maths	Maths	Maths	homework helper basic pre algebra circle open ray open circle closed ray closed filled circle ray point left describing point le le equal figure figure show open ray representing set real number g...
1155	Maths	Maths	Maths	Maths	expression equation division involved writing expression using sym bol becomes somewhat awkward expression usually written fractional form importance order operation think expression sure perform ...

Sentimental Analysis:

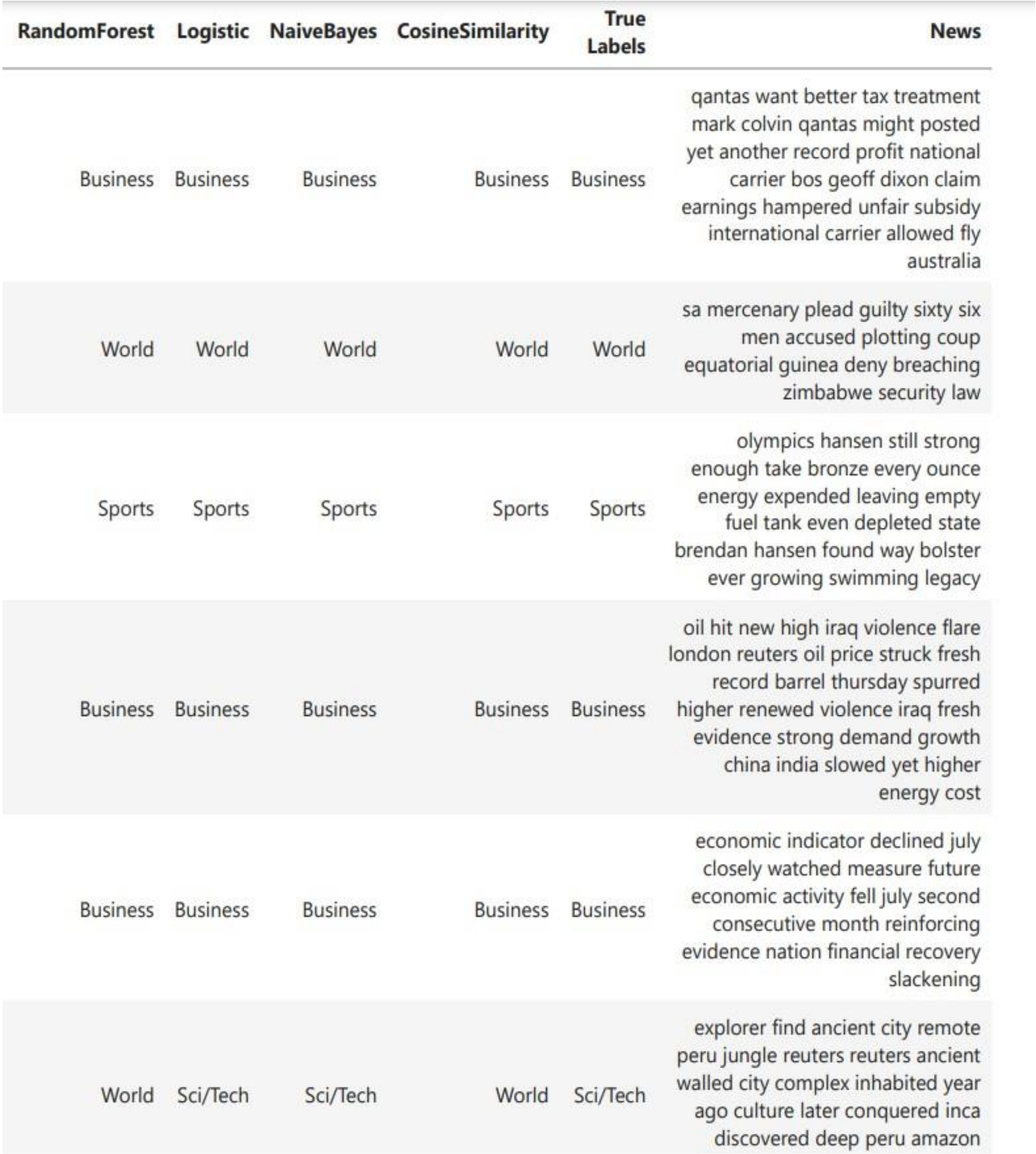
2 = positive

1 = negative



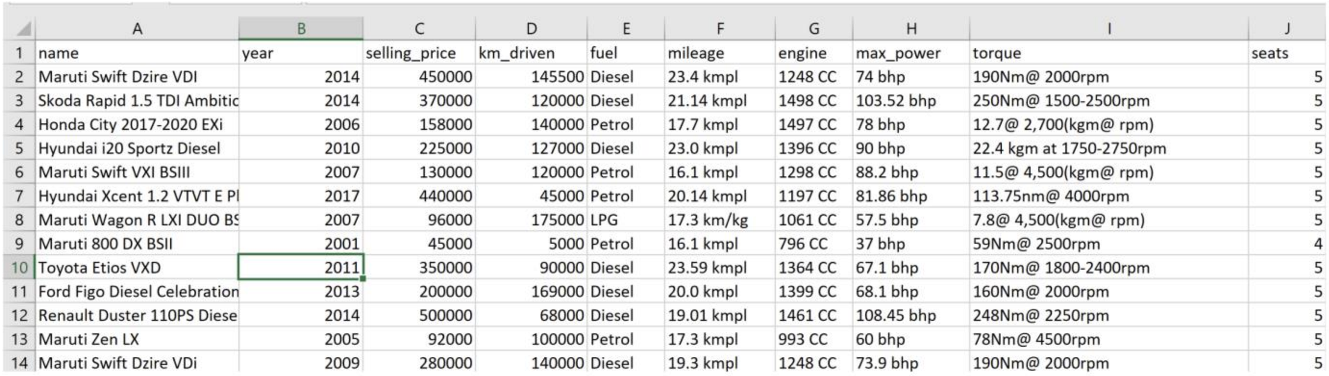
Review	Actual Rating	Predicted Rating
This is an engaging a count of life of Tess a girl who at a young age washed up on the shore of the Isle of May. With no memory of the life before then, she stays with the old caretakers of the isle. After the caretakers are dead a young man, Colin Macpherson washes up on shore. Colin takes Tess back with her to his castle where he helps her uncover her past. They are opposites,but you know what they say opposites attract. This book is just one of many in the Avon True Romance series. I have read every book in the series and I know that if you like this book you love the rest. This is a great book for girls about 10 to 16 because it is still a romance novel, but not what you call a trashy romance novel. It is a great novel fokr those who are just getting into romance novels. a great book for Historical romance lovers	2	2
When I got this book, I wasn't expecting much, but man was I wrong! I loved this book! Now I may not usually be tied in by a man with long blond hair wearing a kilt, but the authors really made Colin into this enchanting Highlander. Tess is a very soft heroine. While she sometimes seems too guillible, I have to remember that it is part of the story. This is one of my favorite books in the series. YES!!!	2	2
Tess Lindsay is content to be the sole occupant of a remote island but when Colin Macpherson washes ashore her life of solitude gets shaken up. Tess had always been told to fear strange men but she reluctantly strikes up a friendship with Colin. Soon Tess discovers that she has a family she had forgotten about but will she have the courage to leave the island and start a new life?The beginning of this book is very amusing and the end is decent too but the middle is a big problem. The plot loses much of it's believability mostly due to the poor characterization. All of the new characters that are introduced are one sided and their behavior lacks reason. Even Tess and Colin's development is side tracked in much the same way and their relationship seems awkward and forced. The writing does improve towards the end but it's not enough to save the book. Unless you have your heart set on reading all of the books in the series, skip this one. Not the best in the series	1	1
I am disappointed in its performance. It seems underpowered and is constantly trying to read CDs, half the time unsuccessfully. I am going to try to return it to Amazon. ADDONICS PORTABLE CD DRIVE - I am disappointed in its performance	1	1
These pants were way too big (looked about 2 sizes larger), and they were incredibly stiff. They would have been very uncomfortable for my daughter to wear all day. Too Uncomfortable and Too Big	1	1
This is my first encounter with Yoruba and I have to say that CDs are really helping. However, the book is very short and not particular about certain aspects and details of grammar and other nuances - for someone who NEVER spoke Yoruba and doesn't have anybody to ask for a clue, it leaves a lot of questions unanswered - for this, however, there are hopefully other Yoruba textbooks.On the brighter note, I would like to add that the book does a very good job in giving a little insight on Yoruba people and Yoruba culture in general. I would recommend it as a good start book for anyone who wants to learn Yoruba - you can't go wrong with it. Very authentic	2	2
I agree, the CDs are a much needed help, since it would be impossible to pronnounce the Yoruba tones without the help of a native speaker. However, the design of the book is really poor and confusing. I have found several spealling mistakes in English and Yoruba. And the	1	1

News Classification:



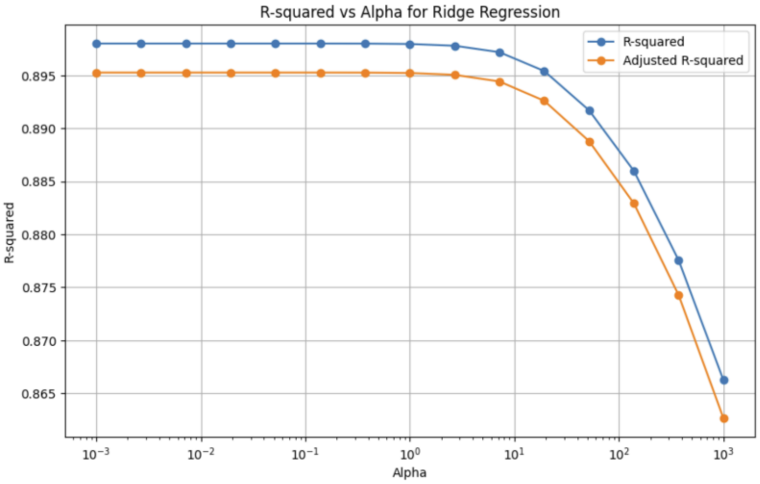
RandomForest	Logistic	NaiveBayes	CosineSimilarity	True Labels	News
Business	Business	Business	Business	Business	qantas want better tax treatment mark colvin qantas might posted yet another record profit national carrier bos geoff dixon claim earnings hampered unfair subsidy international carrier allowed fly australia
World	World	World	World	World	sa mercenary plead guilty sixty six men accused plotting coup equatorial guinea deny breaching zimbabwe security law
Sports	Sports	Sports	Sports	Sports	olympics hansen still strong enough take bronze every ounce energy expended leaving empty fuel tank even depleted state brendan hansen found way bolster ever growing swimming legacy
Business	Business	Business	Business	Business	oil hit new high iraq violence flare london reuters oil price struck fresh record barrel thursday spurred higher renewed violence iraq fresh evidence strong demand growth china india slowed yet higher energy cost
Business	Business	Business	Business	Business	economic indicator declined july closely watched measure future economic activity fell july second consecutive month reinforcing evidence nation financial recovery slackening
World	Sci/Tech	Sci/Tech	World	Sci/Tech	explorer find ancient city remote peru jungle reuters reuters ancient walled city complex inhabited year ago culture later conquered inca discovered deep peru amazon

Used Cars Prediction With Text Data

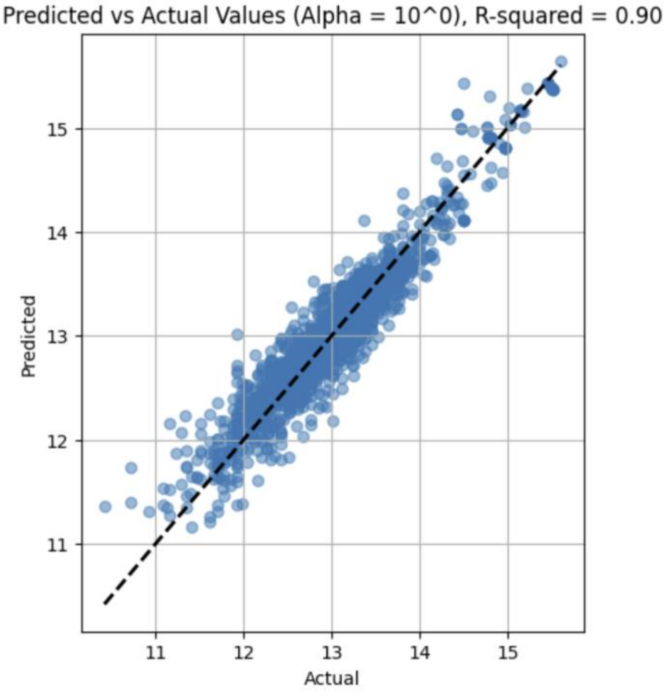


	A	B	C	D	E	F	G	H	I	J
1	name	year	selling_price	km_driven	fuel	mileage	engine	max_power	torque	seats
2	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm	5
3	Skoda Rapid 1.5 TDI Ambitic	2014	370000	120000	Diesel	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm	5
4	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)	5
5	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm	5
6	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm@ rpm)	5
7	Hyundai Xcent 1.2 VTVT E Pl	2017	440000	45000	Petrol	20.14 kmpl	1197 CC	81.86 bhp	113.75nm@ 4000rpm	5
8	Maruti Wagon R LXI DUO BS	2007	96000	175000	LPG	17.3 km/kg	1061 CC	57.5 bhp	7.8@ 4,500(kgm@ rpm)	5
9	Maruti 800 DX BSII	2001	45000	5000	Petrol	16.1 kmpl	796 CC	37 bhp	59Nm@ 2500rpm	4
10	Toyota Etios VXD	2011	350000	90000	Diesel	23.59 kmpl	1364 CC	67.1 bhp	170Nm@ 1800-2400rpm	5
11	Ford Figo Diesel Celebration	2013	200000	169000	Diesel	20.0 kmpl	1399 CC	68.1 bhp	160Nm@ 2000rpm	5
12	Renault Duster 110PS Diese	2014	500000	68000	Diesel	19.01 kmpl	1461 CC	108.45 bhp	248Nm@ 2250rpm	5
13	Maruti Zen LX	2005	92000	100000	Petrol	17.3 kmpl	993 CC	60 bhp	78Nm@ 4500rpm	5
14	Maruti Swift Dzire VDi	2009	280000	140000	Diesel	19.3 kmpl	1248 CC	73.9 bhp	190Nm@ 2000rpm	5

Above you can see the used cars dataset that I had the opportunity to work with. I used cleaning tools to convert mileage, engine, and max_power & torque into numbers that statistical models could understand. I used statistical techniques to maximize R^2 to 90%. Below shows the optimal parameter for alpha!



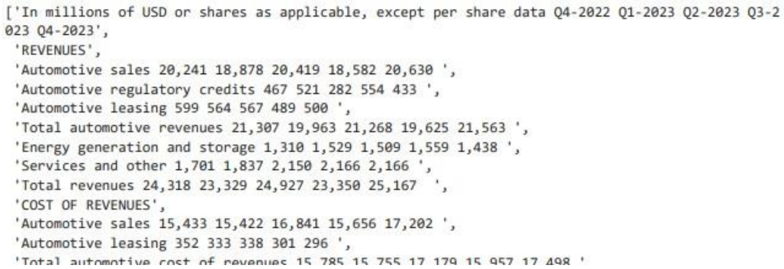
This was the result after statistical learning. The closer the dots are to diagonal the better the model



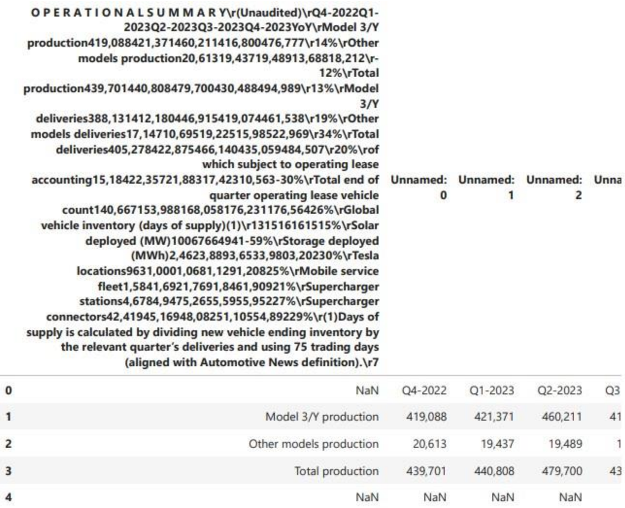
Online/PDF Table Extractor

There is a plethora of information online and In PDFs I've developed tools to clean and extract online/PDF information. My tools below convert messy data into clean readable tables.

Turn messy pdf/web data:

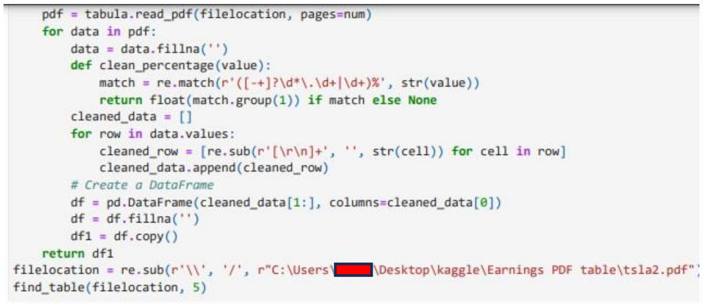


```
['In millions of USD or shares as applicable, except per share data Q4-2022 Q1-2023 Q2-2023 Q3-2
023 Q4-2023',
 'REVENUES',
 'Automotive sales 20,241 18,878 20,419 18,582 20,630 ',
 'Automotive regulatory credits 467 521 282 554 433 ',
 'Automotive leasing 599 564 567 489 500 ',
 'Total automotive revenues 21,307 19,963 21,268 19,625 21,563 ',
 'Energy generation and storage 1,310 1,529 1,509 1,559 1,438 ',
 'Services and other 1,701 1,837 2,150 2,166 2,166 ',
 'Total revenues 24,318 23,329 24,927 23,350 25,167  ',
 'COST OF REVENUES',
 'Automotive sales 15,433 15,422 16,841 15,656 17,202 ',
 'Automotive leasing 352 333 338 301 296 ',
 'Total automotive cost of revenues 15,785 15,755 17,179 15,957 17,498 '
```

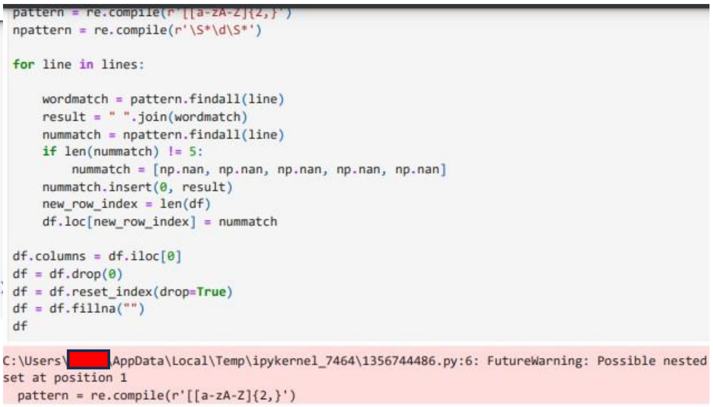


	O P E R A T I O N A L S U M M A R Y\r(Unaudited)\rQ4-2022Q1-2023Q2-2023Q3-2023Q4-2023YoY\rModel 3/Y production419,088421,371460,211416,800476,777\r14%\rOther models production20,61319,43719,48913,68818,212\r-12%\rTotal production439,701440,808479,700430,488494,989\r13%\rModel 3/Y deliveries388,131412,180446,915419,074461,538\r19%\rOther models deliveries17,14710,69519,22515,98522,969\r34%\rTotal deliveries405,278422,875466,140435,059484,507\r20%\rof which subject to operating lease accounting15,18422,35721,88317,42310,563-30%\rTotal end of quarter operating lease vehicle count140,667153,988168,058176,231176,56426%\rGlobal vehicle inventory (days of supply)(1)\r131516161515%\rSolar deployed (MW)10067664941-59%\rStorage deployed (MWh)2,4623,8893,6533,9803,20230%\rTesla locations9631,0001,0681,1291,20825%\rMobile service fleet1,5841,6921,7691,8461,90921%\rSupercharger stations4,6784,9475,2655,5955,95227%\rSupercharger connectors42,41945,16948,08251,10554,89229%\r(1)Days of supply is calculated by dividing new vehicle ending inventory by the relevant quarter's deliveries and using 75 trading days (aligned with Automotive News definition).\r7	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unna
0	NaN	Q4-2022	Q1-2023	Q2-2023	Q3
1	Model 3/Y production	419,088	421,371	460,211	41
2	Other models production	20,613	19,437	19,489	1
3	Total production	439,701	440,808	479,700	43
4	NaN	NaN	NaN	NaN	

Into these clean tables for analytics:

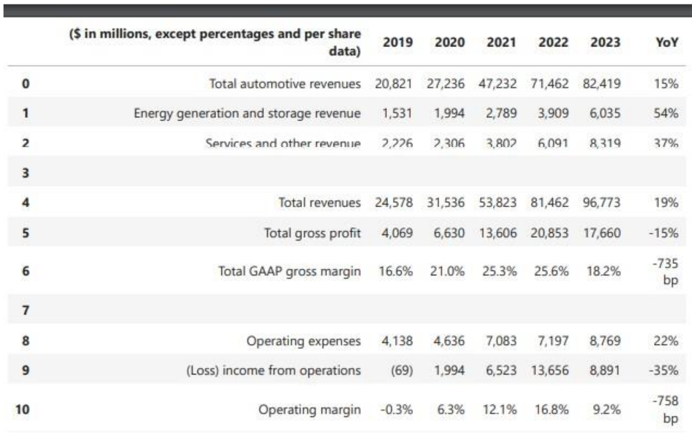


```
    pdf = tabula.read_pdf(filelocation, pages=num)
    for data in pdf:
        data = data.fillna('')
        def clean_percentage(value):
            match = re.match(r'([-+]?\d*\.\d+|\d+)%', str(value))
            return float(match.group(1)) if match else None
        cleaned_data = []
        for row in data.values:
            cleaned_row = [re.sub(r'[\r\n]+', '', str(cell)) for cell in row]
            cleaned_data.append(cleaned_row)
        # Create a DataFrame
        df = pd.DataFrame(cleaned_data[1:], columns=cleaned_data[0])
        df = df.fillna('')
        df1 = df.copy()
    return df1
filelocation = re.sub(r'\\', '/', r"C:\Users\[redacted]\Desktop\kaggle\Earnings PDF table\tsla2.pdf"
find_table(filelocation, 5)
```

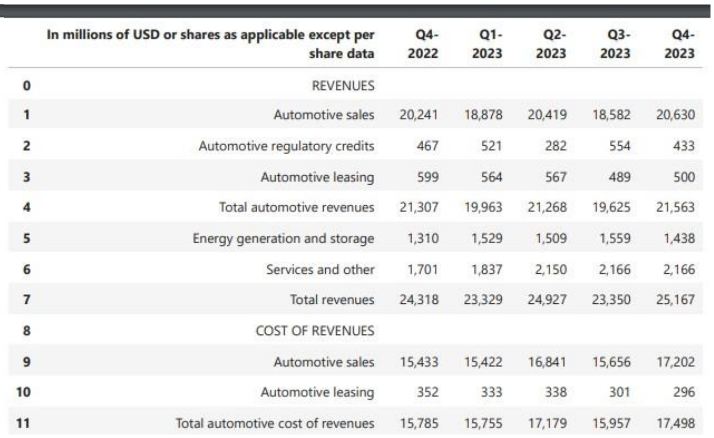


```
pattern = re.compile(r'[[a-zA-Z]{2,}')
npattern = re.compile(r'\S*\d\S*')

for line in lines:

    wordmatch = pattern.findall(line)
    result = " ".join(wordmatch)
    nummatch = npattern.findall(line)
    if len(nummatch) != 5:
        nummatch = [np.nan, np.nan, np.nan, np.nan, np.nan]
    nummatch.insert(0, result)
    new_row_index = len(df)
    df.loc[new_row_index] = nummatch

df.columns = df.iloc[0]
df = df.drop(0)
df = df.reset_index(drop=True)
df = df.fillna("")
df
```

```
C:\Users\[redacted]\AppData\Local\Temp\ipykernel_7464\1356744486.py:6: FutureWarning: Possible nested
set at position 1
  pattern = re.compile(r'[[a-zA-Z]{2,}')
```



	($ in millions, except percentages and per share data)	2019	2020	2021	2022	2023	YoY
0	Total automotive revenues	20,821	27,236	47,232	71,462	82,419	15%
1	Energy generation and storage revenue	1,531	1,994	2,789	3,909	6,035	54%
2	Services and other revenue	2,226	2,306	3,802	6,091	8,319	37%
3							
4	Total revenues	24,578	31,536	53,823	81,462	96,773	19%
5	Total gross profit	4,069	6,630	13,606	20,853	17,660	-15%
6	Total GAAP gross margin	16.6%	21.0%	25.3%	25.6%	18.2%	-735 bp
7							
8	Operating expenses	4,138	4,636	7,083	7,197	8,769	22%
9	(Loss) income from operations	(69)	1,994	6,523	13,656	8,891	-35%
10	Operating margin	-0.3%	6.3%	12.1%	16.8%	9.2%	-758 bp



	In millions of USD or shares as applicable except per share data	Q4-2022	Q1-2023	Q2-2023	Q3-2023	Q4-2023
0	REVENUES					
1	Automotive sales	20,241	18,878	20,419	18,582	20,630
2	Automotive regulatory credits	467	521	282	554	433
3	Automotive leasing	599	564	567	489	500
4	Total automotive revenues	21,307	19,963	21,268	19,625	21,563
5	Energy generation and storage	1,310	1,529	1,509	1,559	1,438
6	Services and other	1,701	1,837	2,150	2,166	2,166
7	Total revenues	24,318	23,329	24,927	23,350	25,167
8	COST OF REVENUES					
9	Automotive sales	15,433	15,422	16,841	15,656	17,202
10	Automotive leasing	352	333	338	301	296
11	Total automotive cost of revenues	15,785	15,755	17,179	15,957	17,498