CS 110 Final Project

Stephanie Gillow

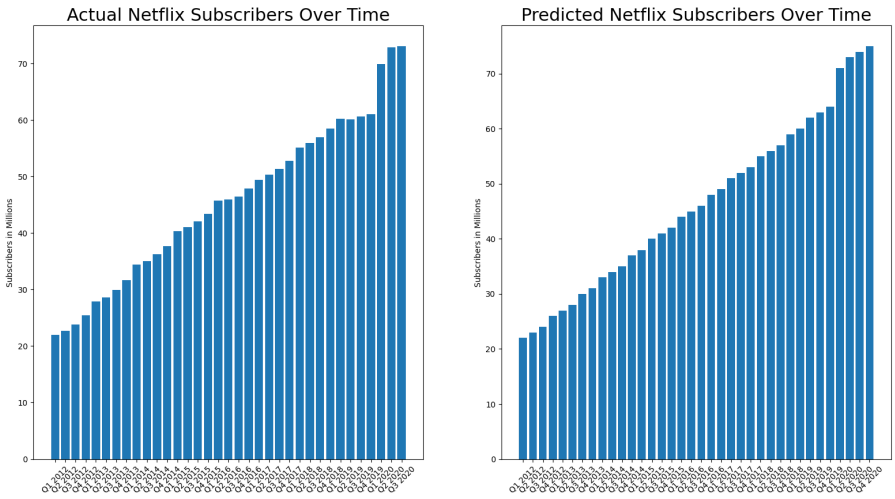
December 2020

1 Overview and Summary of Project

My project is a linear regression model that takes in a CSV file containing existing statistics of a streaming service's subscriber growth, and outputs predictions of future growth, both as string representations and side-by-side histograms. Sample input:

	A	B
1	1	22.02
2	2	22.69
3	3	23.8
4	4	25.47
5	5	27.91
6	6	28.62
7	7	29.93
8	8	31.71
9	9	34.38
10	10	35.09
11	11	36.27
12	12	37.7
13	13	40.32
14	14	41.06
15	15	42.07
16	16	43.4
17	17	45.71
18	18	46
19	19	46.48
20	20	47.91
21	21	49.38
22	22	50.32
23	23	51.35
24	24	52.81
25	25	55.09
26	26	55.96
27	27	56.96
28	28	58.49
29	29	60.23
30	30	60.1
31	31	60.62
32	32	61.04
33	33	69.97
34	34	72.9
35	35	73.08

Sample output:



2 Target Audience

My target audiences are investors of companies that provide any form of streaming services, whether private, or just people interested in buying stocks, as well as the companies themselves - they would likely be interested in figuring out what their future growth will potentially look like, based on a rather solid algorithm. This could also be a useful tool for influencers, or game streamers, or YouTubers; it could track followers on social media just as well, so long as there's prior data available. The more data available, the more accurate it becomes.

3 Specific Programming Techniques Used

For the machine learning portion of this program, I used Scikit-Learn's linear regression model. I trained and tested it using the Netflix dataset I got from Statista.

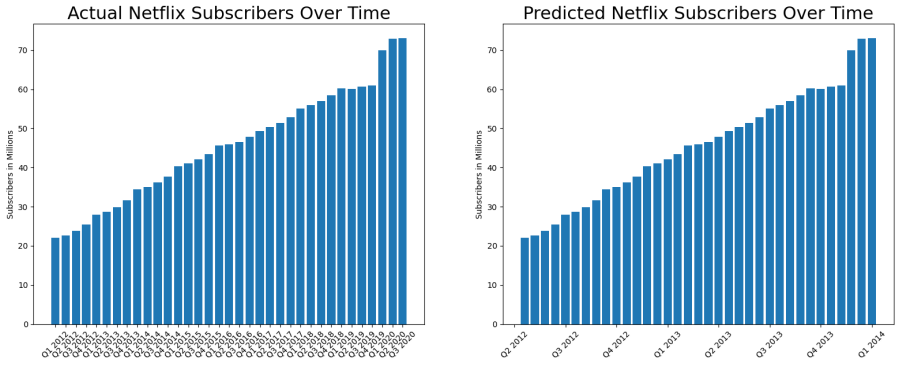
For the string representations, as well as writing the predictions to a CSV, I ran into the issue of having to automate labeling for labels that change often, eg. "Q4 2019" to "Q1 2020". To change the financial quarter labels, I modded the counter by 4, and then wrote a function that changes it to "4" in the event of a "0". For the year, I incremented the latter 2 numbers by .25 each pass, and cast it to an integer - that way, it only changed to the next year once four quarters had passed. Every pass, I also had the loops write to a new CSV file

that's later passed to Pandas and Matplotlib to create the histograms. I created a one-row, two-column subplot to display both the real data and predicted data side-by-side.

4 Challenges

Initially, I wanted to use Tensorflow. However - I could not get it working, no matter what I tried. At first, pip did not even recognize that Tensorflow existed and would say no such package exists. I changed PyCharm to 64-bit, which was weird because 64 is the default, so I must have purposefully changed it at some point. I attempted to download it via direct link. I reinstalled Python, pip, and all dependencies again. I created a Manjaro virtual machine. Eventually, after fixing everything possible, three hours later - I realized all I was dealing with now was a network error. I then attempted to create a linear regression model, but honestly, staring at Tensorflow was annoying me at this point, so I switched to Scikit-Learn.

My first attempt at creating subplots resulted in one functional graph in one window, and two other blank graphs in separate windows. That took a lot of troubleshooting to fix. Then, I had issues with the labels not corresponding correctly for the second graph whatsoever, pictured here:



I emailed Professor Ryan about it, because I was, frankly, stumped. Nothing was actually wrong in my syntax or formatting. I ignored it to finish other aspects, and it eventually went away. Weird.

5 Future Extensions

In the future, this could be applied to multiple datasets at once. This would allow comparison and prediction as to which companies, or individuals, are most likely to experience the most growth over a period of time. As far as new features, I think it would be interesting to try to automate said comparison with

another algorithm. Or, using multiple algorithms to get the predictions, then averaging those, might yield even more accurate results.