

# MSA 8040 DATA MANAGEMENT FOR ANALYTICS: FINAL PROJECT

Houping Xiao, Georgia State University

November 9, 2021

**Requirements (Please read it carefully!):**

- (a) **Please keep it confidential and do not distribute it outside our cohort!**
- (b) **Make a presentation in the final week. (Please check the syllabus accordingly for the specific date). Each group will have at most 10 minutes to (1) present your database models and (2) query demo. [This part is 70%]**
- (c) **Submit a report indicating the database model (such as SQL or NoSQL) and some basic analysis, and the web scraping code. All submission files should be submitted on iCollege before **December 13th, 2021** for full consideration. [This part is 30%]**
- (d) **This is group-based project. Only one group member need to submit the files.**
- (e) **Individual-based peer evaluation form. Each group member need to submit their own evaluation for the other group members in your group.**

## Peer-Evaluation Form

**To better achieve fairness in the class, at the end of the course you will be asked to evaluate yourself and the other members of your group on completing the project. These ratings are used for gauging team members' contributions. The grade you and your group members receive will depend in part on these peer evaluations. Rate each member based on the following criteria: (1) participation in group activities, (2) quality of work, (3) quantity of work, (4) finishing assigned work on time, and (5) ability to work as a team member. Please use the following scale to assign scores:**

5	Exceptional effort, above and beyond the call of duty
4	Above average effort
3	Normal effort (this is the expected score!)
2	Below average effort
1	Unacceptable effort

**Then, submit the following note to the instructor on iCollege:**

Your Name:\_\_\_\_\_ Score:\_\_\_\_\_ Reasons:\_\_\_\_\_  
Team Member #2:\_\_\_\_\_ Score:\_\_\_\_\_ Reasons:\_\_\_\_\_  
Team Member #3:\_\_\_\_\_ Score:\_\_\_\_\_ Reasons:\_\_\_\_\_  
Team Member #4:\_\_\_\_\_ Score:\_\_\_\_\_ Reasons:\_\_\_\_\_  
Team Member #5:\_\_\_\_\_ Score: \_\_\_\_\_ Reasons:\_\_\_\_\_

**Note: Please include a brief reason for any group member. I expect everyone to**

*be thoughtful and diligent in completing this evaluation. Your final project grade will be biased by the peer-evaluation scores from your group members. For instance, you could be graded as ZERO for the project if you receive “1”s from all other group members.*

## 1 Problem Background

Estimize, an open web-based platform, has become an useful source to retrieve financial estimates. It facilitates the aggregation of financial estimates from a diverse community of individuals. As shown in Figure 1, founded in 2011, there are increasing numbers of contributors who join the platform to provide financial estimates for an increasing numbers of companies. Consequently, an increasing numbers of earning forecasts can be used by Estimize for better consensus estimates. As of March 2020, in total there are 97,439 users contributed for 2,220 companies (e.g., stocks) on the platform. Compared with the analysts in IBES, Estimize solicits contributions from a community of individuals with a wide range of occupations. As shown in Figure 1, Estimize covers different kinds of contributors including both professionals, such as sell-side, buy-side, or independent analysts, and nonprofessionals, such as academia, student, Information Technology, and Energy. Because of the contributions of these individuals, who have diverse backgrounds and viewpoints, Estimize consensus can be more accurate than the Wall Street consensus and provides incremental information for forecasting earnings and for measuring the market expectation of earnings.

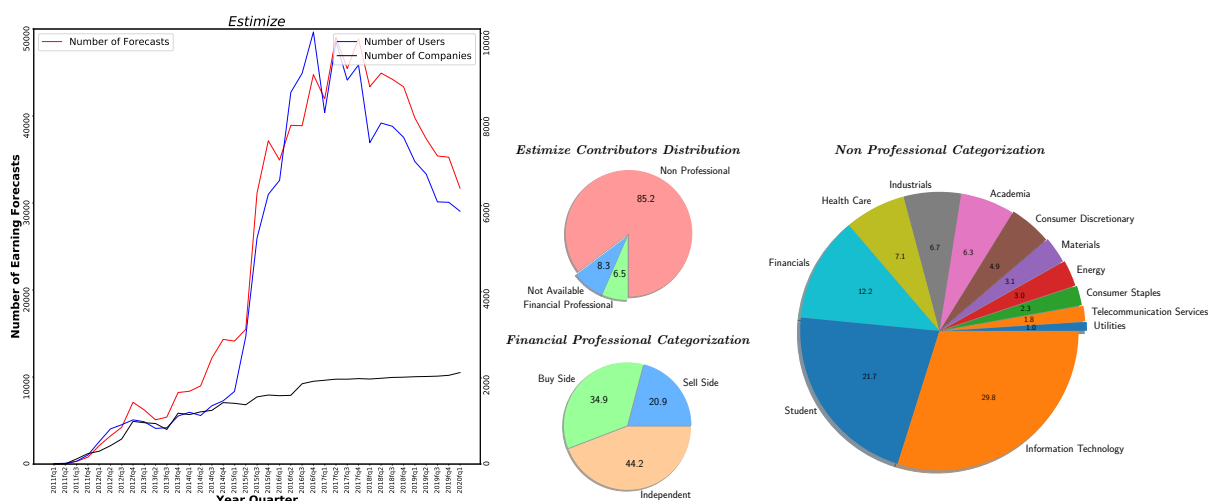


Figure 1: Estimize Sample Coverage and diversity

## 2 Tasks

During your presentation, you should manage to provide and answer the following information. Each section has its points.

- (a) Scrape the data from the Estimize.com for at least one year. [10%]

- (b) Build a dataset to store the EPS estimation information, including:
- (a) The company basic information, such as Ticker, company name, Sectors, Industries, number of followers, number of analysts, as shown in Figure 2. **[6%]**
    - i. ticker: EHTH
    - ii. name: eHealth, inc.
    - iii. Sectors: Financials
    - iv. Industries: Insurance
    - v. number of folowers: 66
    - vi. number of analysts: 99
  - (b) The EPS information, including Reported Earnings, Estimze Consensus, Estimze Mean, Wall Street Consensus, and EPS estimations of all available analysts, as shown in Figure 3. **[10%]**
  - (c) The information about each analyst, including Analyst name, Roles, Join date, Analyst Confidence score, number of estimates (Figure 4), Stocks Covered (Figure 5), pending estimates (Figure 6), Scored estimates (Figure 7). **[14%]**
    - i. name: Steven Halper
    - ii. roles: Financial Professional, Sell Side, and Broker
    - iii. Join Date: Jan 2017
    - iv. Analyst Confidence Score: 5.4
    - v. error rate: 17.3%
    - vi. Accuracy Percentile: 24%
    - vii. points: 16
    - viii. points/Estimate: 1
    - ix. stocks: 20
    - x. pending: 28
    - xi. All companies information in Figures 5, 6 and 7
- (c) Your database should support easy query, extracting the correct information for the questions similar to following ones: **[30%]**
- (a) Given a ticker, how many analysts have made estimations for its EPS? Rank them by their confidence score, total points, error rate or accuracy percentile? **[10%]**
  - (b) Given a industry, how many companies are covered, the average number of analysts, the average bias between the Estimze Consensus and the Reported Earnings? **[10%]**
  - (c) Which company have the largest number of analysts with confidence score greater than 7? **[10%]**
  - (d) Who has the largest number of followers? **[10%]**

You will be asked to answer three questions. Each qustion is 10%.

(d) **Bonus:** [20%]

- (a) Regression analysis is the basic method to find the important independent variables that will affect the dependent variable. So, given all the features constructed and scraped, can you find some important independent variables that affect accuracy of prediction? [10%]
- (b) Can you come up with some novel method to construct a better EPS estimation compared with the Estimate Consensus and Wall Street Consensus? [10%]

**Note:** To be fully consideration, both the presentation and the report should contain the analysis!




Figure 2: Company basic information

EPS Analysts: FQ3 '21

Showing 5/5 estimates [View all-time analyst rankings for EHTH](#) ☐ Show only my followed analysts

Chart	☆	Analyst	Rank	▲	Points	Value	Confidence	Last Revised
		<b>EHTH Reported Earnings</b> eHealth, Inc.				-1.78		11/08/21
		<b>Estimize Consensus</b> 5 estimates weighted				-0.91		11/08/21
		<b>Estimize Mean</b> 5 estimates averaged				-0.89		11/08/21
		<b>Wall Street Consensus</b>				-1.38		11/08/21
	☆	 <b>Steven Halper</b> Cantor_13	1		-3	-1.03	3.6	07/29/21
	☆	 <b>Tobey Sommer</b> Suntrust_86	2		-3	-1.00	3.9	08/18/21
	☆	 <b>Vitaly</b> Vitaly	3		-3	-0.93	5.4	11/08/21
	☆	 <b>Bill</b> BillB1210	4		-4	-0.89	5.5	11/07/21
	☆	 <b>Tobey Sommer</b> Suntrust_5	5		-5	-0.60	1.0	07/23/20

Figure 3: All available analysts



☆

Steven Halper

Cantor\_13

Financial Professional - Sell Side - Broker

Member since Jan 2017

Analyst Confidence

5.4

based on 362 estimates

Equities

Economics

THIS SEASON FALL 2021

17.3%

Error Rate

14.8% all time

24%

Accuracy Percentile

39% all time

16

Points

1.667 all time

1

Points/Estimate

4.6 all time

20

Stocks

27 all time

28

Pending Estimates

362 total

League Memberships

Steven Halper is not a member of any leagues yet.

[Browse all leagues >](#)

Followed Analysts

Steven Halper doesn't follow any analysts.

[Browse all analysts >](#)

Absolute Error Rate Over Time

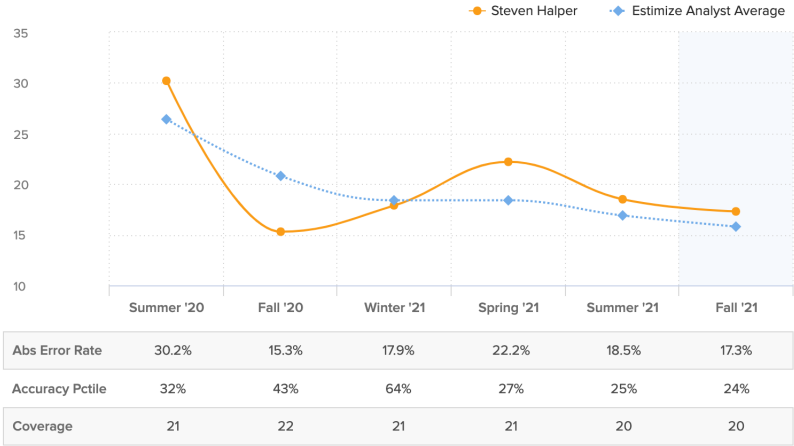


Figure 4: An analyst basic information

Stocks Covered

Showing 10 / 22 stocks covered

Ticker	Reports	Quarters	Points	Pts/Est	Error Rate	Accuracy
VCRA	Feb 10	20	104	5.2	29.2%	34%
MDRX	Feb 17	20	269	13.5	9.3%	51%
OMCL	Feb 3	20	59	3	11.6%	33%
HSTM	Feb 22	20	149	7.5	26.5%	35%
TVTY	Feb 24	19	96	4.9	14.4%	37%
CPSI	Nov 9	19	-84	-4.4	13.7%	27%
CERN	Feb 8	19	319	16.4	1.5%	58%
EVH	Mar 1	19	196	10.1	27.9%	55%
TDOC	Mar 2	18	45	2.4	10.1%	54%
EHTH	Feb 24	18	182	9.8	32.2%	54%

Show 12 more

Figure 5: All covered stock estimates by the analyst

#### Pending Estimates 18 stale

Showing 5/28 pending estimates




Ticker	Quarter	Reports	Published	EPS	Revenue
<a href="#">CPSI</a>	Q3 2021	Nov 9, 2021 <a href="#">AMC</a>	Aug 3, 2021  97 days ago	0.68	70.90
<a href="#">HGY</a>	Q3 2022	Dec 7, 2021 <a href="#">AMC</a>	Oct 4, 2021	0.34	184.00
<a href="#">UNH</a>	Q4 2021	Jan 18, 2022 <a href="#">BMO</a>	Oct 14, 2021	4.31	71.60
<a href="#">ANTM</a>	Q4 2021	Jan 26, 2022 <a href="#">BMO</a>	Oct 20, 2021  19 days ago	5.01	3710
<a href="#">NXGN</a>	Q3 2022	Jan 27, 2022 <a href="#">AMC</a>	Oct 28, 2021  11 days ago	0.24	144.90
Show 20 more					

Figure 6: All pending stock estimates by the analyst

#### Scored Estimates

Showing 5/392 scored estimates

Ticker	Quarter	Reported	Rank	EPS Points	Revenue Points	Total Points
<a href="#">EHTH</a>	Q3 2021	Nov 8, 2021	1 / 5	-3	10	7
<a href="#">MDRX</a>	Q3 2021	Nov 4, 2021	4 / 6	-6	13	7
<a href="#">CI</a>	Q3 2021	Nov 4, 2021	9 / 9	-4	-25	-29
<a href="#">EVH</a>	Q3 2021	Nov 3, 2021	1 / 7	25	21	46
<a href="#">BNFT</a>	Q3 2021	Nov 3, 2021	3 / 3	-12	-6	-18
Show 20 more						

Figure 7: All scored stock estimates by the analyst