INSY 670 - Assignment 1 - Report

Yichen Wang, Yingxin (Corrine) Jiang, Diwei Zhu, Kejia (Alice) Liu, Kexin Wang

Part I:

1. Which factors are the best predictors of influence? Are there any surprises here? How can a business use your model/results

   we use random forest to do the feature selection and select the top 5 most important features:'A_B_listed_count','A_B_network_feature_1','A_B_mentions_received','A_B_follower_count', and 'A_B_retweets_received'. And then we use logistic, random forest and gradient boosting to train our model. After comparing the accuracy, we choose to use Gradient boosting as our main model which has the highest accuracy score at 0.77. By using our models, we can provide a relatively reliable suggestion on which influencer has more influence power with an accuracy of around 77%

2. What is the boost in expected net profit from using your analytic model (versus not using analytics)? Show all calculations. What is the boost in net profit from using a perfect analytic model (versus not using analytics)?

   (all the detailed calculations within Jupyter Notebook)

   From using our analytics model, the net profit is $6162.55, and the boost in expected net profit is $861.

   From using the perfect analytics model, the net profit is $7972.22, and the boost in expected net profit is $2670.74

Part II:

We scraped 5,000 Tweets with Keyword=”Zelda”. Based on the scraped data, we counted tweets, retweets and mentions of users. For all users that posted or were mentioned, we collected user features (list_counts, followers_counts, etc) that are identified to be informative by the model of Part I.

We use the predictive feature importance of the model in Part I as the initial weights of the features and set degree centrality as the network feature 1. Then, we balanced the weights with ratio (making sure that $w_1+w_2+w_3+w_4=1$) and subjective ±0.5 adjustment.

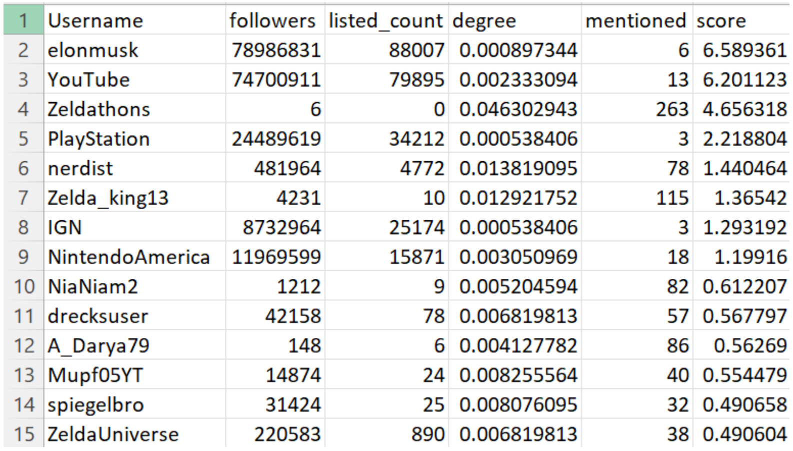
Feature	Initial weight	adjustment	final weight
A_B_listed_count	0.182173	*ratio+adjust	0.392383
A_B_network_feature_1	0.123855	*ratio	0.232778
A_B_mentions_received	0.114590	*ratio-adjust	0.209473
A_B_follower_count	0.111455	*ratio	0.165365

With the assigned weights and the normalized user information, we identified the top 100 influencers of the keyword with the scores calculated through:

Score = $w_1 \times listed\_count + w_2 \times degree\ centrality + w_3 \times mentions + w_4 \times \#followers,$

where $w_1+w_2+w_3+w_4=1.$

The top 15 influencers are shown below. For the full list, please check the code.



1	Username	followers	listed_count	degree	mentioned	score
2	elonmusk	78986831	88007	0.000897344	6	6.589361
3	YouTube	74700911	79895	0.002333094	13	6.201123
4	Zeldathons	6	0	0.046302943	263	4.656318
5	PlayStation	24489619	34212	0.000538406	3	2.218804
6	nerdist	481964	4772	0.013819095	78	1.440464
7	Zelda_king13	4231	10	0.012921752	115	1.36542
8	IGN	8732964	25174	0.000538406	3	1.293192
9	NintendoAmerica	11969599	15871	0.003050969	18	1.19916
10	NiaNiam2	1212	9	0.005204594	82	0.612207
11	drecksuser	42158	78	0.006819813	57	0.567797
12	A_Darya79	148	6	0.004127782	86	0.56269
13	Mupf05YT	14874	24	0.008255564	40	0.554479
14	spiegelbro	31424	25	0.008076095	32	0.490658
15	ZeldaUniverse	220583	890	0.006819813	38	0.490604

With Gephi, we generated the network visual of the top 100 influencers we selected: