Yelp Data Prediction

Preliminary Analysis

Yifan Li, Chenlai Shi, Jianmin Chen

Monday Group 1

2.2 Additional Variable

- **year**: scaled year variable.
- **loc1**: 1 if the restaurant is in the western United States, otherwise 0.
- **loc2**: 1 if the restaurant is in the estern United States, otherwise 0.
- **loc3**: 1 if the restaurant isn't in the United States, otherwise 0.

2.2 Additional Variable

- **S1 $\sim$ S5**: $\text{S1[word]} = \frac{\text{P(this word is included in reviews with 1 star)}}{\text{P(this word is included in reviews with other stars)}}$

Word	Variable	1-star	2-star	3-star	4-star	5-star
refund	frequence	115	15	7	4	2
	probability	0.011	0.002	0	0	0
	S1 $\sim$ S5	34.200	1.080	0.300	0.072	0.025
notdisappoints	frequence	0	2	5	43	110
	probability	0	0	0	0.002	0.003
	S1 $\sim$ S5	0	0.116	0.188	0.917	3.870
and	frequence	9196	8691	12851	25604	32071
	probability	0.859	0.886	0.877	0.895	0.886
	S1 $\sim$ S5	0.968	1.000	0.991	1.020	1.000

4 Compare RMSE with other method

RMSE

Feature\ Model	GLM	LM	SVM	NB	LSTM	NN
frequence	0.864	0	0	0	0	0
tf-idf	0.836	0	0	0	0	0
vector	0	0	0	0	0	0
ad	0	0	0	0	0	0
vector + ad	0	0	0	0	0	0