Predicting Sucess on yelp

Ramin Ostad and Will Sundstrom

September 24,

Objective

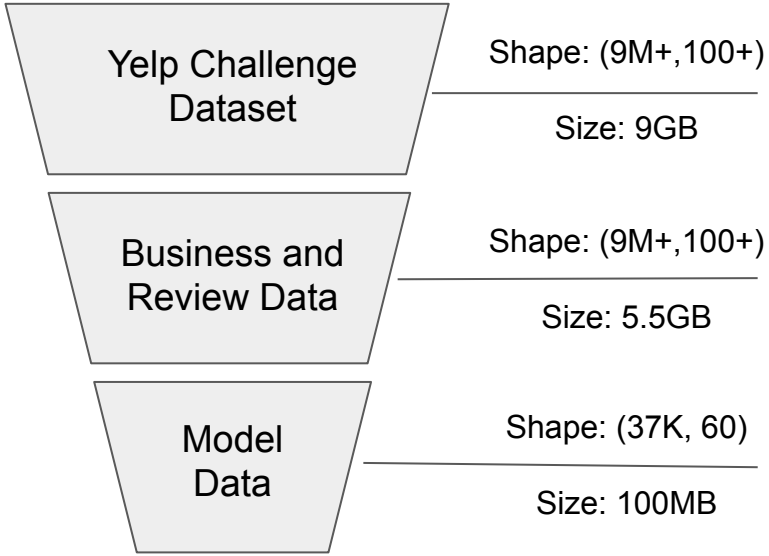
Our aim is to develop a success prediction for new restaurants

Success is defined as 4+ stars

The predictor can help new restaurants evaluate whether they are ‘on track’ to becoming successful restaurants

Data overview

The dataset comes from Yelp and was honed down in Cloud SQL



Data Overview:

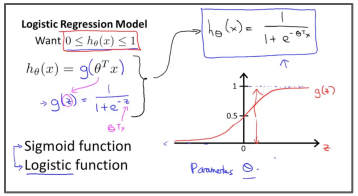
- Data provided through Yelp 2019 Challenge
- Used Google Cloud SQL to store data and run initial cuts
- Types of data:
  - Business information
  - Reviews and tips
  - User information
- Data ultimately used in model:
  - Business information
  - Aggregate characteristics of reviews for each restaurant

Approach

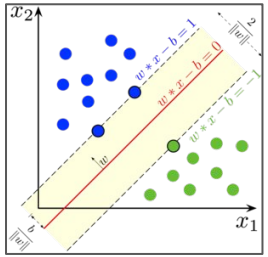
We will attempt 3 models: Logistic Regression, SVM and Random Forest

Models

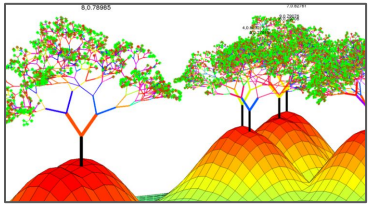
Logistic Regression



Support Vector Machine



Random Forest



Success Metrics:

- Accuracy - we aim to maximize accuracy. False negatives and false positives have relatively similar value in this context. Target classes are balanced.

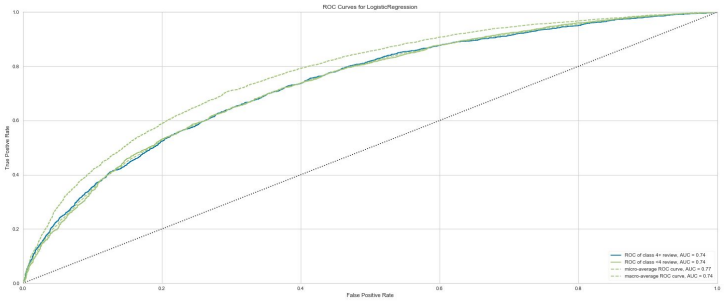
- Efficiency - models should run with our available computing power

Model and Hyperparameter Tuning

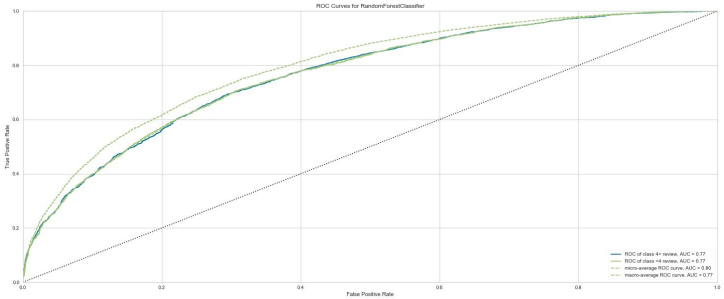
Models, after hyperparameter tuning, resulted in similar results

- Models capped out at ~0.71 accuracy
- We believe that improvement hyperparameter tuning have been exhausted
- Further improvement may be had by adding features from other data in the dataset

ROC Curve: Linear Regression



ROC Curve: Random Forest



Model Selection

Accuracy is close between Random Forest and SVM. We choose Random Forest as our model based on performance and resource requirements

Classifier	Accuracy*	Resource Requirement	Other Consideration
Random Forest	71.7585%		Feature Importance
SVM Nystroem	71.3564%		
SVM Monte Carlo	70.4559%		
Logistic Regression	70.9743%		

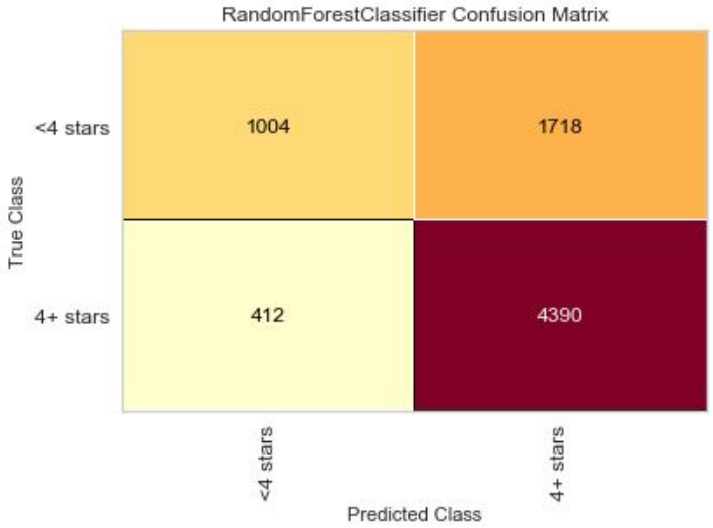
*Mean accuracy in k-fold cross validation

Model Performance

Key features include location as well as aggregate review characteristics

	Importance	Feature
4	0.063159	meanuseful
1	0.062690	latitude
14	0.060382	lowerquartilewordcount
2	0.055452	longitude
12	0.052334	medianwordcount
5	0.048457	avgwordcount
0	0.046001	review_count
13	0.044722	upperquartilewordcount
3	0.041821	meanfunny
10	0.035700	avgusefulwordcount

Key features include location as well as aggregate characteristics from business reviews



Recommendations & Limitations

Further Research and Next Steps

Potential improvement areas:

- Feature addition from reviews
- Reviews
- Changing model to look at initial reviews as a predictor of later success

Thank You!

Slide Template Option

Template Option

Key Point: lorum ipsum etc

Key Point: lorum ipsum etc

Key Point: lorum ipsum etc

Key Point: lorum ipsum etc