Determining Future Success of Yelp Restaurants Using Machine Learning

Pranav Dahiya, Pratyush Dwivedi, Siddharth Singh
1710110249, 1710110255, 1710110xxx

The Problem Statement

- Investors need information regarding future success of restaurants in order to make investment decisions
- This can be modelled as a binary classification problem with restaurants that remain open for another year being one class and restaurants that close in the next year being the other

Existing Work

- Previous work of two different papers was analyzed.
- Previous papers used data from 2013 and 2016-17 Yelp DataSet respectively.
- Both Text and Non-text features were used
- Only a linear logistic regression model was used
- Their results were limited to the accuracy of 65% and 67.46% with a baseline of 50% respectively

Features Extracted

- **Stars :** Average Stars / Age
- **Review Count :** Number of Reviews for a restaurant / Age
- **Stars 2017 :** Average Stars of Reviews for a restaurant written in 2017
- **Review Count 2017 :** Total number of Reviews for a restaurant written in 2017
- **Chain :** Number of Data Set in the restaurant with the same name
- **Tips :** Total Number of Tips for a restaurant / Age
- **Tips 2017 :** Number of Tips for a restaurant in 2017
- **Check-Ins :** Total Number of Check-ins for a restaurant / Age
- **Check-Ins 2017 :** Number of Check-ins for a restaurant in 2017

Features Extracted (Contd.)

- **Age :** 2019 - year of earliest Review/Tip/Check-in
- **Density :** Number of restaurants in a 2Km radius
- **Category Density :** Number of restaurants of the same category in a 2Km radius / Density

Relative Values of all features except density and category density were computed with respect to corresponding values of restaurants in a 2Km radius.

No text-based features were used as the previous literature highlighted the inefficiency in doing so.

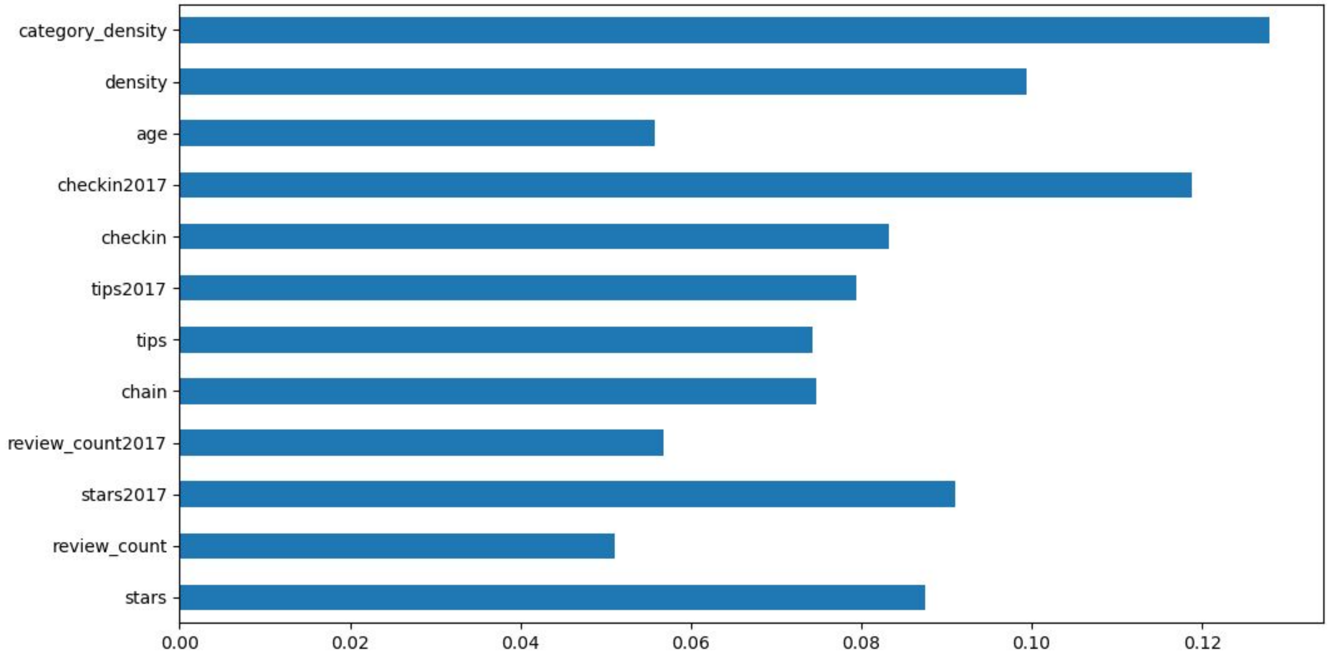
Preprocessing

All features were Normalized by subtracting the mean and dividing by the standard deviation

Results

	accuracy		precision		f1		recall	
Perceptron	56.87	5.23	57.86	5.64	55.39	6.28	54.43	10.56
LogisticRegression	63.33	5.10	63.61	5.14	63.66	3.34	63.92	2.13
SVM	63.27	4.59	63.23	4.64	63.89	3.12	64.75	2.45
SVM with polynomial kernel	64.90	4.94	66.75	5.83	63.50	3.51	60.76	2.32
SVM with gaussian kernel	67.07	4.55	67.55	4.95	67.04	3.26	66.70	2.10
DecisionTree	58.74	3.33	59.04	3.47	58.54	2.43	58.22	2.87
NeuralNetwork	67.23	4.60	67.39	5.09	67.59	3.30	68.06	3.23

Feature Importance (From Decision Tree)



Conclusion

- Achieved comparable results to existing state of the art (State of the art achieves 67.47%, we have achieved 67.23%)
- Extracted more important features than state of the art (chain was most important feature in existing work)
- Most important feature was category_density.