Predicting Restaurant Success With Yelp Data

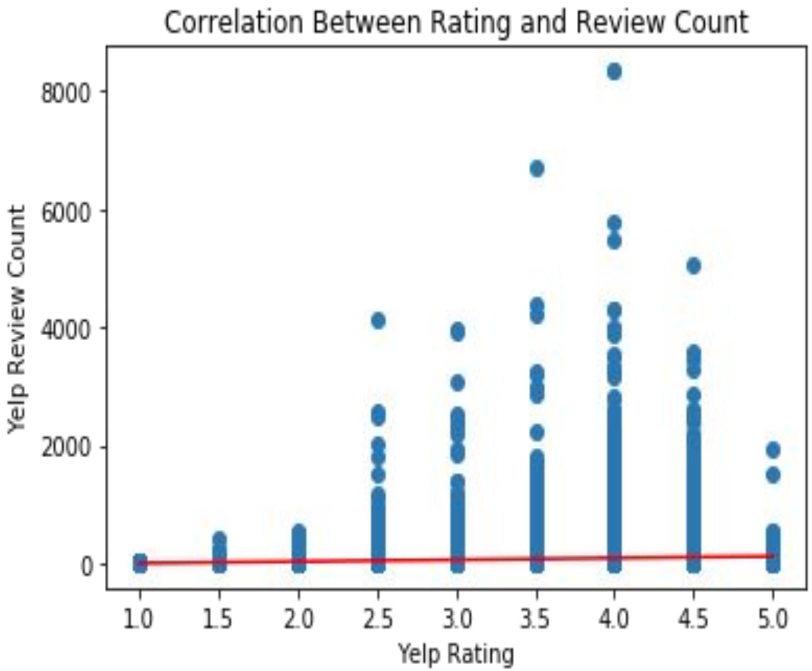
Vaibhav Malik, Siddharth Musale

Increase Chance of Success in an Expensive and Risky Industry

- Research shows that 60% of restaurants fail within the first year and 80% within 5 years with a median investment amount of $350,000
- Analyze dataset provided by Yelp to find relationship between restaurant aspects and success metric: normalized review count per capita. Analysis provides insight on which restaurant features lead to future success
- Success metric is more representative of a restaurant's ability to generate customer volume compared to restaurant rating

High Yelp Rating != High Customer Volume

- Yelp rating and review count had a weak correlation with r=0.13
- Certain restaurant features are peaking customer interest and might be more influential than quality of food and service
- Restaurant owners who have knowledge of which specific features and values of these features lead to higher success metric maximize their opportunity for success



Papers Analyze Different Types of Data and Use Different Definitions of Success

Paper 1

Features

Total count of unigram sentiments, total count of bigram sentiments, and restaurant attributes

Success metric

Whether restaurant remained open between 2016 and 2017

Model

Logistic regression

Paper 2

Features

Total count of negative and positive sentiments for different themes found in reviews

Success metric

Yelp rating

Model

Naive Bayes

Paper 3

Features

Satellite light data, restaurant attributes, road network data, and points of interest information

Success metric

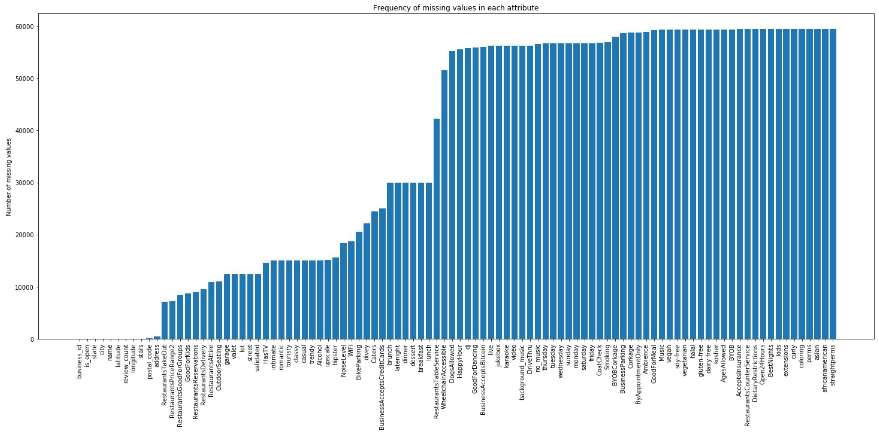
Yelp rating

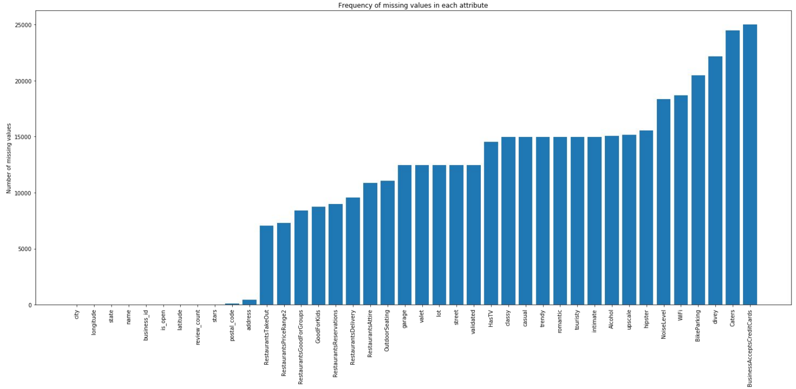
Model

Embedded CNNS

Data Cleaning & Preprocessing Allows Easy Pipelining Into Analytical Models

- Yelp Dataset is in JSON format which was converted into a DataFrame for use in Python and analytical models
- Restaurant Attributes with less than 50% coverage were dropped
- Instance specific attributes were dropped as well

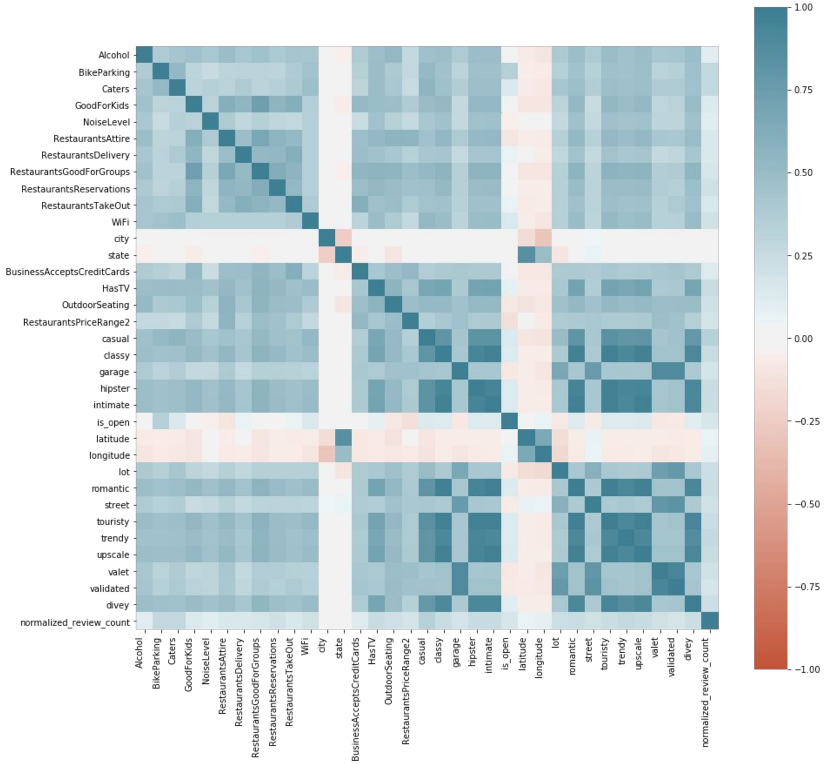




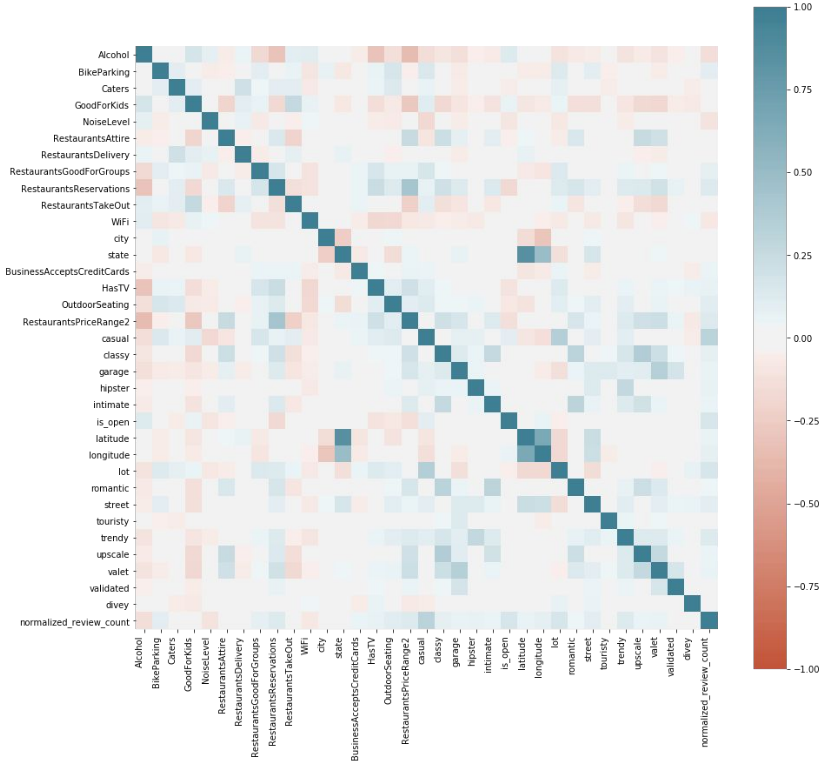
Feature Engineering Determined Which Features to Use in Models

Attribute Name	Type	Values
BikeParking	Nominal	['True', 'False']
Caters	Nominal	['False', 'True']
Alcohol	Nominal	['full_bar', 'none, ,'beer_and_wine']
GoodForKids	Nominal	['True', 'False']
NoiseLevel	Nominal	[‘average' ,'quiet', 'very_loud', 'loud']
RestaurantsAttire	Nominal	[‘casual', ’dressy', ‘formal']
RestaurantsDelivery	Nominal	['False', 'True']
RestaurantsPriceRange2	Numeric	[1, 2, 3, 4]
casual	Numeric	[1, 0]
classy	Numeric	[1 ,0]
garage	Numeric	[1, 0]

Multiple Imputation Techniques Used to Fix Missing Values



Constant “-1” Imputation Correlation Heatmap



1NN Imputation Correlation Heatmap

Success Metric Provides Representation Of How Restaurant Compares to Others

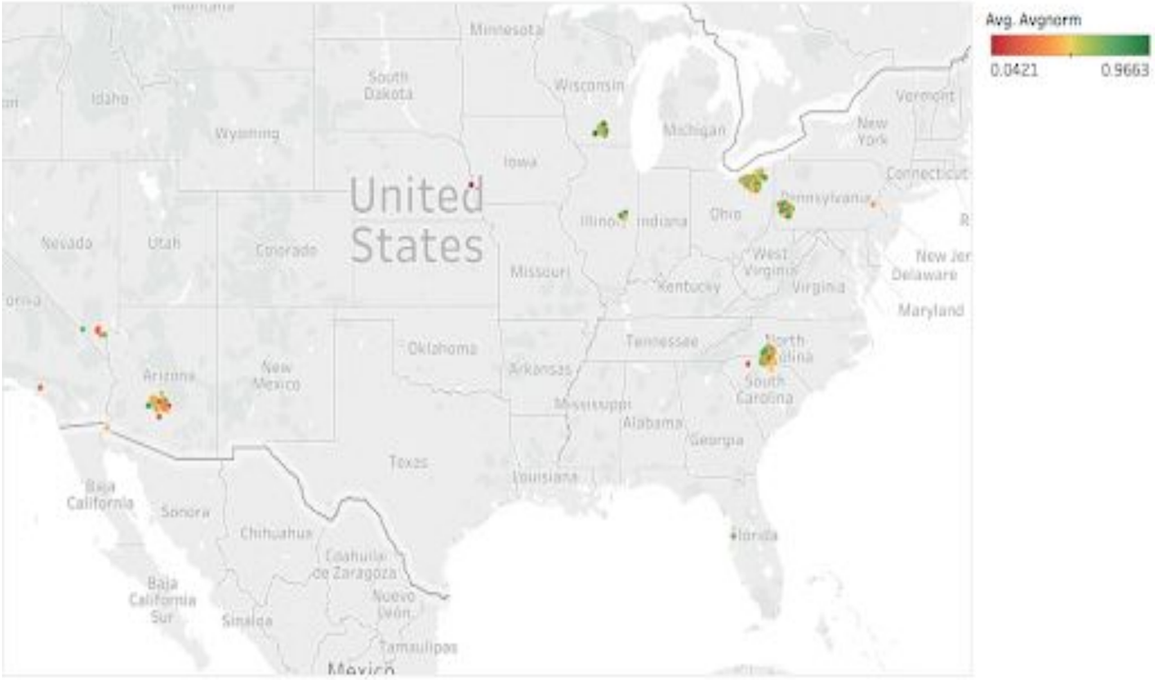
Population of city that restaurant is located in was added to dataset

Review count per capita calculated by dividing review count by city population

Instances with review count per capita >.0003 are filtered out

Review count per capita is normalized by transforming every value to range [0,1]

Average Normalized Review Count for U.S. Cities

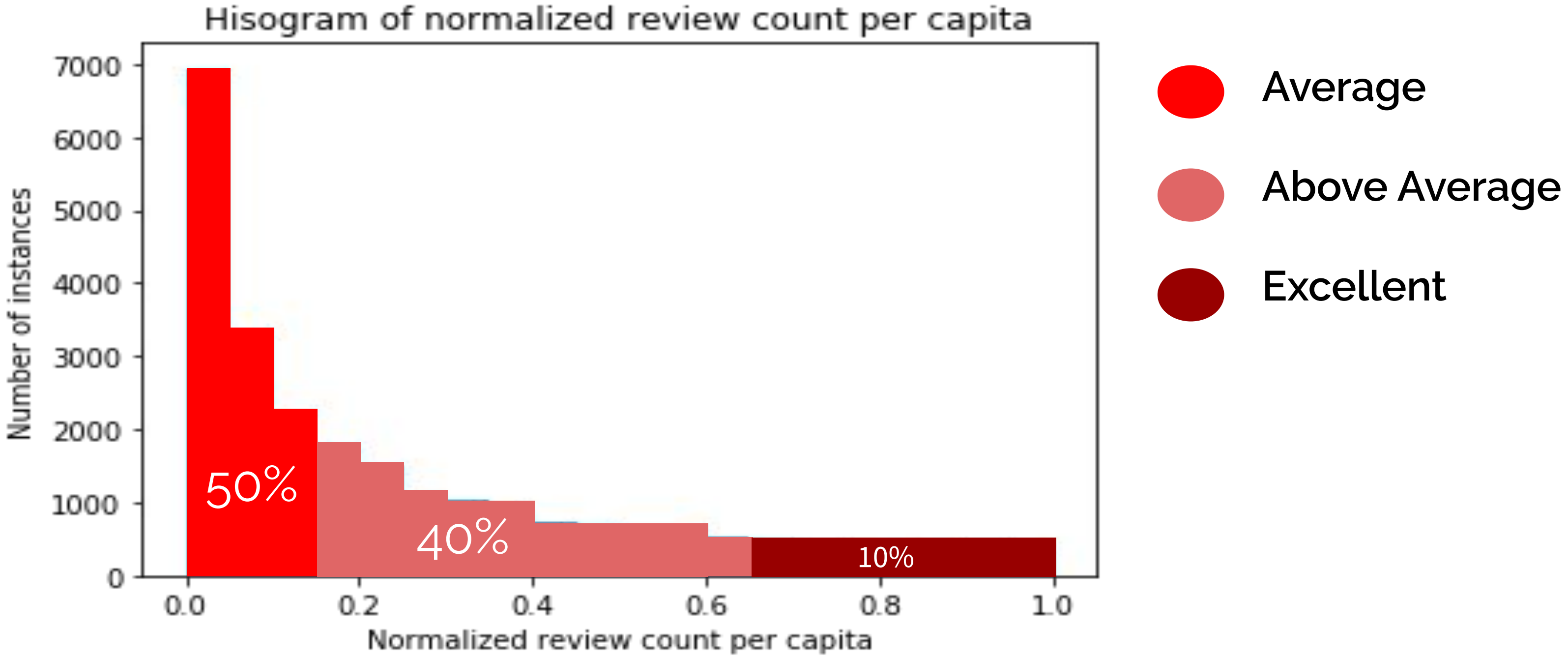


Map based on Longitude and Latitude. Color shows average of Avgnorm.

Regression Models Predict Within 1 Standard Deviation

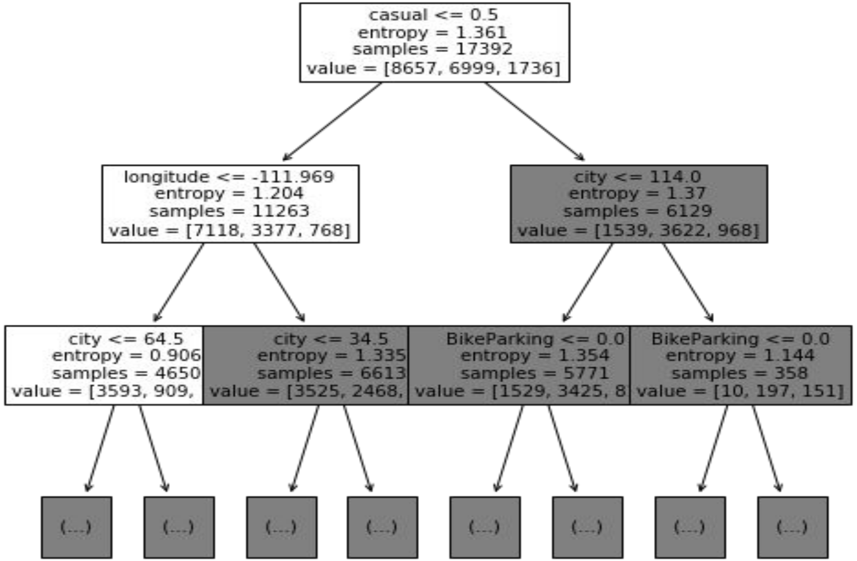
	Decision Tree Regressor	NN Model
RMSE	0.20	.1788
R Squared	0.13	.52
Standard Dev	0.26	.26

Break Up Ranges of Success Metric Values Into Different Classes for Classification



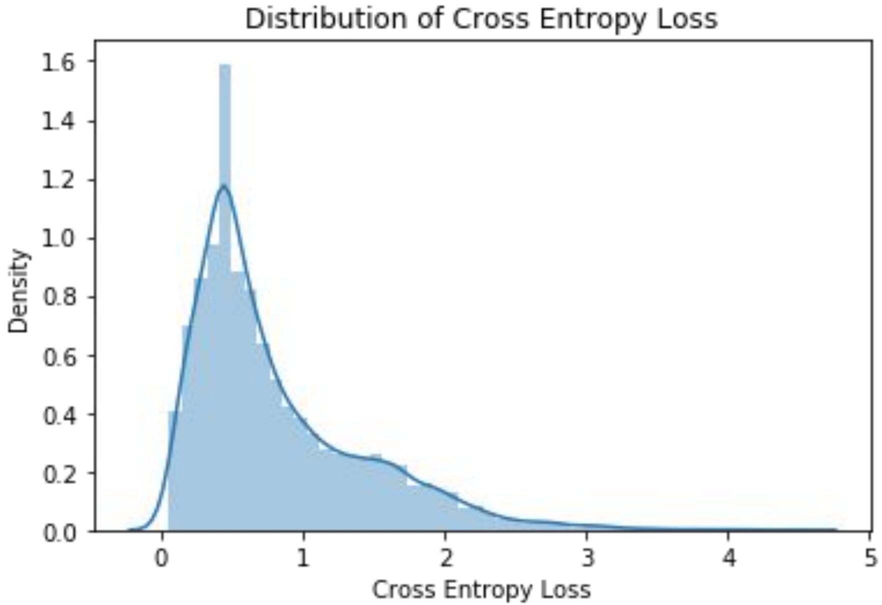
Decision Tree Classifier Honing in on Regional Customer Preferences

Class	Precision	Recall	F1 Score
0	0.81	0.82	0.81
1	0.63	0.75	0.68
2	0.42	0.07	0.12
Accuracy: 0.71			



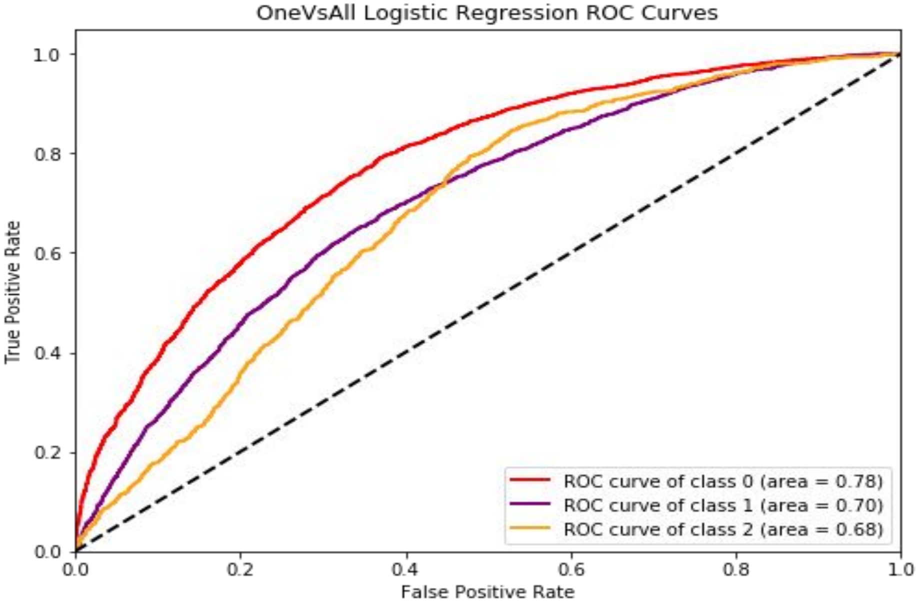
Logistic Regression Unable To Correctly Predict “Excellent” Class

Class	Precision	Recall	F1 Score
0	0.67	0.76	0.71
1	0.57	0.61	0.59
2	0	0	0
Accuracy: 0.63			



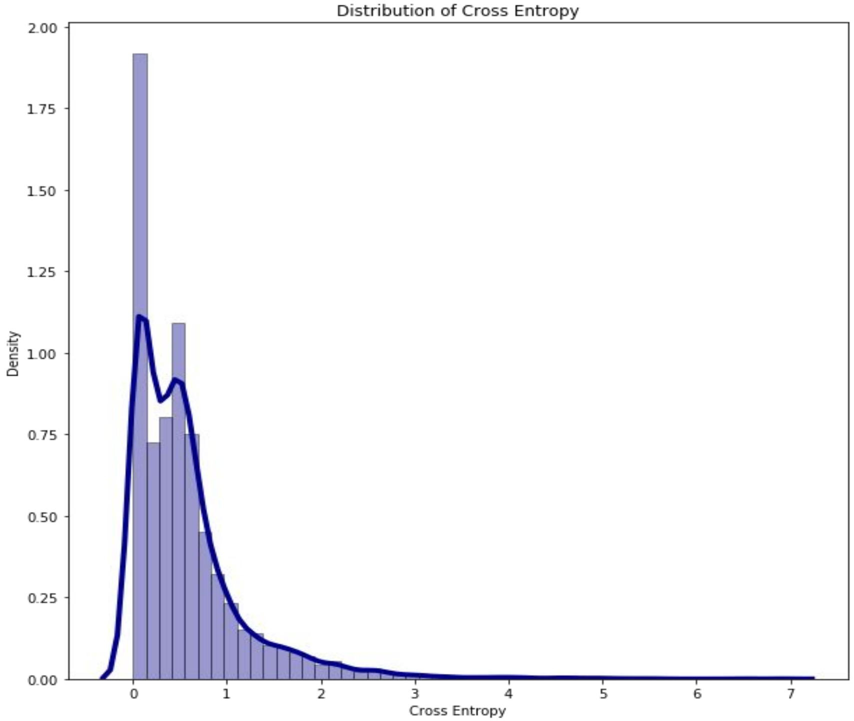
One Vs Rest Classifier Not Considering Relative Probability of Other Classes

Class	Precision	Recall	F1 Score
0	0.70	0.72	0.71
1	0.60	0.48	0.53
2	0	0	0
Accuracy: 0.55			



NN Model Resulted in the Highest F1 Scores Across All Models

Class	Precision	Recall	F1 Score
0	0.84	0.84	0.84
1	0.66	0.79	0.72
2	0.56	0.14	0.23
Accuracy: 0.75			



Difficult To Predict Actual Values, Classes Represent Similarities Between Restaurants

- Numerical models resulted in relatively poor performance because most restaurant attributes are represented by binary and nominal values
- Classification system was able to uncover similarities between restaurants within the same normalized review count range based on attribute values
- Geographical attributes were pivotal to the performance of some classification models and insignificant for others
- Adding state, city, latitude, and longitude attributes resulted in 8% higher accuracy for decision tree classifier, but didn't change accuracy for LR and NN

Many More Questions Asked Than Answered

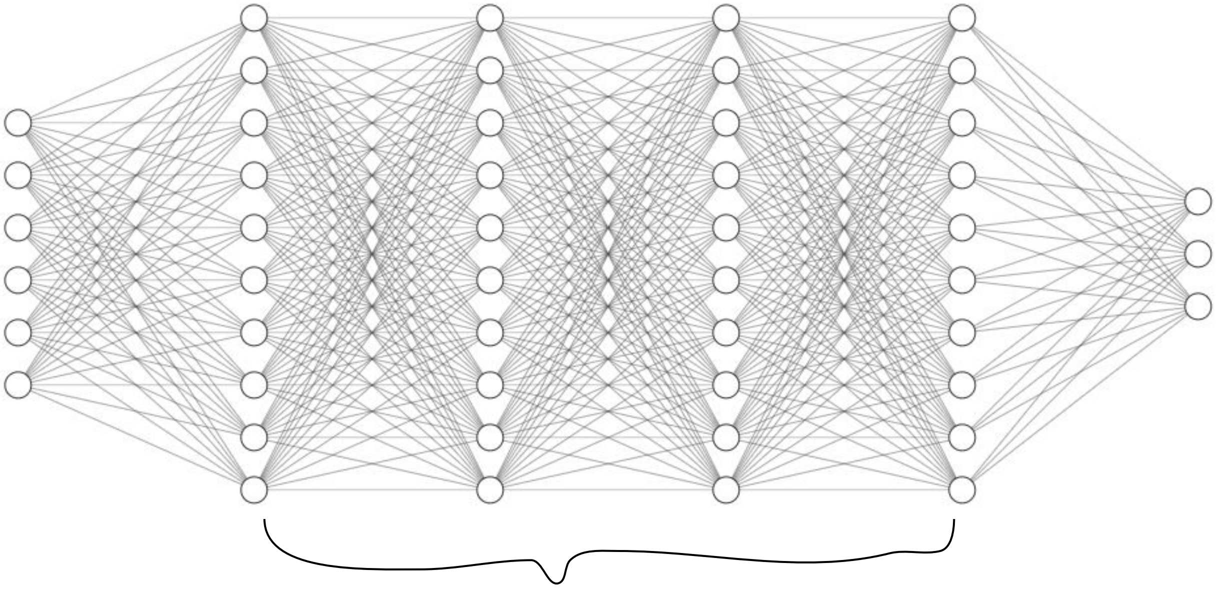
- Does the review count accurately represent how many Yelp users visited the restaurant?
- Foundation of success metric relies on assumption that review count is representative of customer volume. Is the proportion of Yelp reviewers to all customers consistent across restaurants?
- Accuracy of classification models was heavily reliant on the system used to assign classes to instances. In the future, how can we optimize the number of classes and classification conditions to maximize performance?
- Restaurants are essentially competing with local competitors for customers. Would applying model methodology to a specific region or distance radius, i.e. normalize review count per capita according to region, result in higher accuracy?

Questions?

Appendix

NN Architecture (Classification)

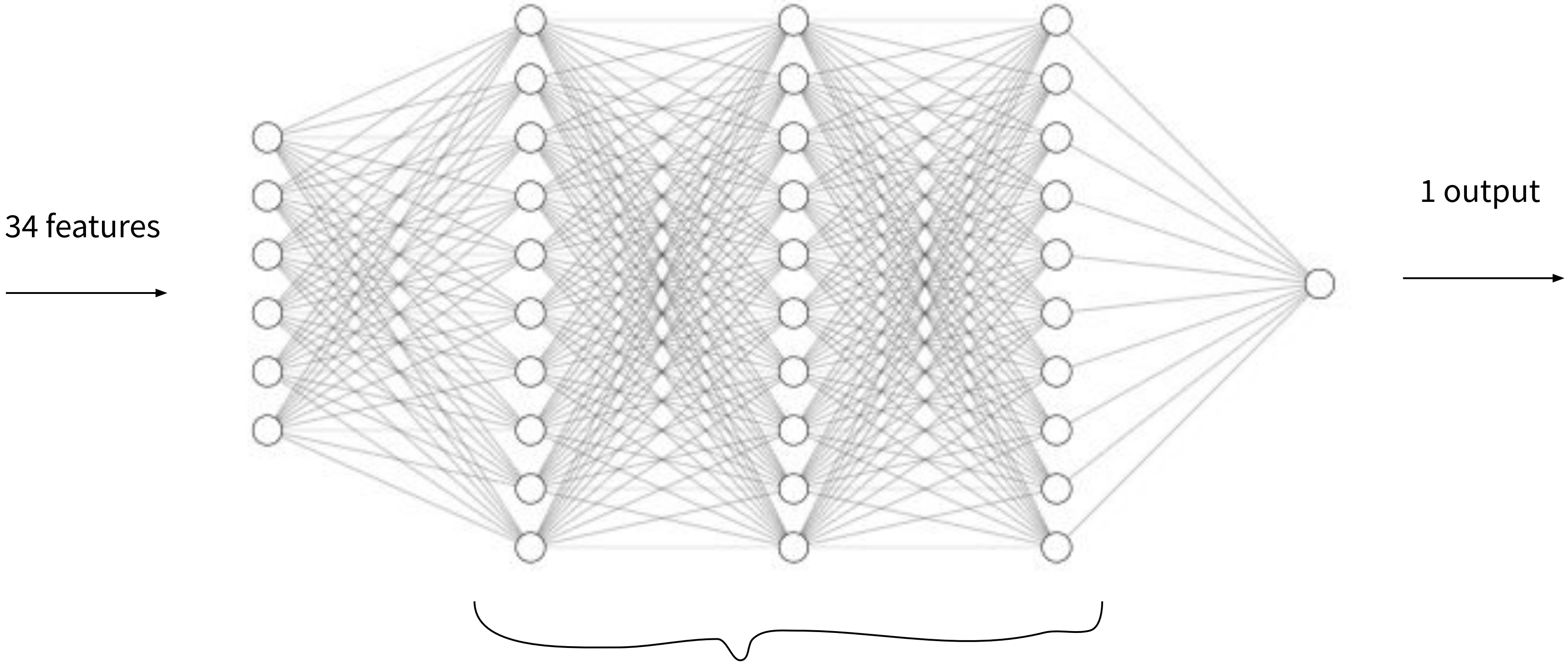
34 features



3 class softmax output

2048 Hidden Units Each Layer

NN Architecture (Regression)



34 features

1 output

1024 Hidden Units Each Layer

Datasets

Yelp Dataset Challenge:

https://www.yelp.com/dataset/documentation/main

U.S. Cities Dataset:

https://simplemaps.com/data/us-zips

Sklearn Machine Learning Models

Decision Tree Regressor:

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

Decision Tree Classification:

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html?highlight=classifier#sklearn.tree.DecisionTreeClassifier

Logistic Regression:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegressionCV.html?highlight=logistic#sklearn.linear_model.LogisticRegressionCV

One vs Rest Classifier:

https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html?highlight=rest#sklearn.multiclass.OneVsRestClassifier

Timelogs

1) Activity Log - Siddharth
2) Activity Log-Vaibhav