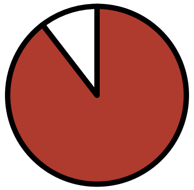
Rating Prediction
Big Data Machine Learning

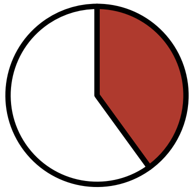
Weili Yuan, Ping Ren, Fengxu Li, Jianbo Wu, Haiyun Wang

Project Overview

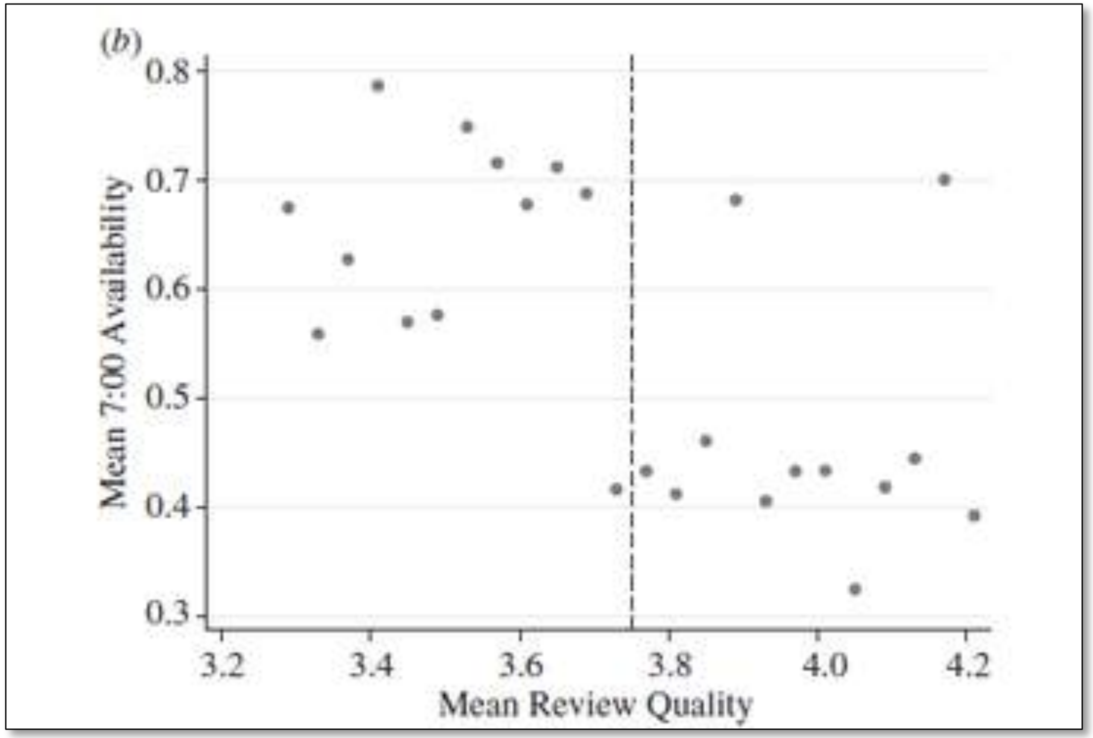
The difference of a half-star in a Yelp rating can make a huge difference



57% of consumers will only use a business if it has 4 or more stars



40% of consumers only take into account reviews written within the past 2 weeks



This project creates value for investors and restaurant managers

What kind of restaurants are more likely to be popular?
What kind of restaurants are more likely to higher returns?

For new restaurants, what is the most important area to pay attention to?
What should be done to improve rating?

business.json

{"business_id":"1SWheh84yJXfytovILXOAQ","name":"Arizona Biltmore Golf Club","address":"2818 E Camino Acequia Drive","city":"Phoenix","state":"AZ","postal_co
{"business_id":"QXAEGFB4oINsVuTFxEYKFQ","name":"Emerald Chinese Restaurant","address":"30 Eglinton Avenue W","city":"Mississauga","state":"ON","postal_code"
{"business_id":"gnKjwL_1w79qoiV3IC_xQQ","name":"Musashi Japanese Restaurant","address":"10110 Johnston Rd, Ste 15","city":"Charlotte","state":"NC","postal_c
{"business_id":"xvX2CttrVhyG2z1dFg_0xw","name":"Farmers Insurance - Paul Lorenz","address":"15655 W Roosevelt St, Ste 237","city":"Goodyear","state":"AZ","p
{"business_id":"HhyxOkGAM07SRYtlQ4wMFQ","name":"Queen City Plumbing","address":"4209 Stuart Andrew Blvd, Ste F","city":"Charlotte","state":"NC","postal_code
{"business_id":"68dUKd8_8liJ7in4aWOSEA","name":"The UPS Store","address":"Credit Valley Town Plaza, F2 - 6045 Creditview Rd","city":"Mississauga","state":"C
{"business_id":"5JucpCfHZltJh5r1JabjDg","name":"Edgeworxx Studio","address":"20 Douglas Woods Drive Southeast","city":"Calgary","state":"AB","postal_code":"
{"business_id":"gbQN7vr_caG_A1ugSmGhWg","name":"Supercuts","address":"4545 E Tropicana Rd Ste 8, Tropicana","city":"Las Vegas","state":"NV","postal_code":"8
{"business_id":"Y6iyemLX_oylRpnr38vgMA","name":"Vita Bella Fine Day Spa","address":"5940 W Union Hills Dr","city":"Glendale","state":"AZ","postal_code":"85
{"business_id":"4GBVPIYRvzGh4K4TkRQ_rw","name":"Options Salon & Spa","address":"21689 Lorain Rd","city":"Fairview Park","state":"OH","postal_code":"44126","
{"business_id":"fcXOEZdXYeZqnQ3lGlOXmg","name":"Nucleus Information Service","address":"1210 8th Street SW, Unit 220","city":"Calgary","state":"AB","postal_
{"business_id":"1Dfx3zM-rW4n-31KeC8sJg","name":"Taco Bell","address":"2450 E Indian School Rd","city":"Phoenix","state":"AZ","postal_code":"85016","latitude
{"business_id":"5t3KVdMnFgAYmSl1wYLhmA","name":"The Kilted Buffalo Langtree","address":"119 Landings Dr, Ste 101","city":"Mooresville","state":"NC","postal_
{"business_id":"fweCYi8FmbJXHCqLnwuk8w","name":"Marco's Pizza","address":"5981 Andrews Rd","city":"Mentor-on-the-Lake","state":"OH","postal_code":"44060","l
{"business_id":"-K4gAv8_vjx8-2BxkVeRkA","name":"Baby Cakes","address":"4145 Erie St","city":"Willoughby","state":"OH","postal_code":"44094","latitude":41.63
{"business_id":"A98xW4qb7vOTguggHFs7Ng","name":"Hot Yoga Wellness","address":"1455 16th Avenue","city":"Richmond Hill","state":"ON","postal_code":"L4B 3G6",
{"business_id":"giC3pVVFxCRR89rApqklyw","name":"Knot Salon","address":"4848 E Cactus Rd, Ste 100","city":"Scottsdale","state":"AZ","postal_code":"85254","la
{"business_id":"PZ-LZzSlhSe9utkQYU8pFg","name":"Carluccio's Tivoli Gardens","address":"1775 E Tropicana Ave, Ste 29","city":"Las Vegas","state":"NV","postal
{"business_id":"nh_kQ16QAoXWwqZ05MPfBQ","name":"Myron Hensel Photography","address":"","city":"Las Vegas","state":"NV","postal_code":"89121","latitude":36.1
{"business_id":"zSpQmEBvRe3IhTUlMSA6HQ","name":"Totum Life Science","address":"445 King Street W, Suite 101","city":"Toronto","state":"ON","postal_code":"M5
{"business_id":"dFMxzHygTy6F873843dHAA","name":"Fremont Arcade","address":"450 Fremont St, Ste 179","city":"Las Vegas","state":"NV","postal_code":"89101","l
{"business_id":"lxnuq9wJiwLOPJ4uZU2ljg","name":"Las Vegas Motorcars","address":"3650 N 5th, Ste 100","city":"North Las Vegas","state":"NV","postal_code":"89
{"business_id":"KWywu2tTEPWmR9JnBc0WyQ","name":"Hunk Mansion","address":"6007 Dean Martin Dr","city":"Las Vegas","state":"NV","postal_code":"89118","latitud
{"business_id":"1RHY4K3BD22FK7Cfftn8Mg","name":"Marathon Diner","address":"Center Core - Food Court, Fl 3, Pittsburgh International Airport","city":"Pittsbu
{"business_id":"BsMdebN4nZySpGTfXr-7yg","name":"Maurices","address":"7981 W Tropical Pkwy, Ste 120","city":"Las Vegas","state":"NV","postal_code":"89149","l
{"business_id":"tstimHoMcYbkSC4eBA1wEg","name":"Maria's Mexican Restaurant & Bakery","address":"6055 E Lake Mead Blvd","city":"Las Vegas","state":"NV","post
{"business_id":"C9oCPomVP0mtKa8z99E3gg","name":"Bakery Gateau","address":"865 York Mills Road, Unit 1","city":"Toronto","state":"ON","postal_code":"M3B 1Y6"
{"business_id":"C9keC4mWuXdl2mYFHZXudQ","name":"Uncle Otis Clothing","address":"26 Bellair St","city":"Toronto","state":"ON","postal_code":"M5R 2C7","latitu
{"business_id":"iojTeSaoPuxm4WeCzDUA6w","name":"AW Collision","address":"3325 W Desert Inn Rd, Ste 101","city":"Las Vegas","state":"NV","postal_code":"89102
{"business_id":"NDuUMJfrWk52RA-H-OtrpA","name":"Bolt Fresh Bar","address":"1170 Queen Street W","city":"Toronto","state":"ON","postal_code":"M6J 1J5","latit
{"business_id":"LB6ZyCfUzeX9OLdunHhnOQ","name":"Ross Dress for Less","address":"8800 W Charleston Blvd","city":"Las Vegas","state":"NV","postal_code":"89117
{"business_id":"mNBp4KI2goFJKDB9VLGP9w","name":"JSE Automotive Services","address":"6875 N 21st Ave, Ste 118","city":"Phoenix","state":"AZ","postal_code":"8
{"business_id":"SP_YXIEwkFPPl_9anCYmpQ","name":"The Steady Cafe & Bar","address":"1051 Bloor Street W","city":"Toronto","state":"ON","postal_code":"M6H 1M4"
{"business_id":"irft4YkdNsww4DNf_Aftew","name":"So Cool Frozen Yogurt","address":"9020 B Albemarle Rd","city":"Charlotte","state":"NC","postal_code":"28227"
{"business_id":"HVPcIcqiJkrpD36xZFGN6g","name":"Bank of America","address":"3150 N Rainbow Blvd","city":"Las Vegas","state":"NV","postal_code":"89108","lati
{"business_id":"BvYU3jvGd0TJ7IyZdfiN2Q","name":"Manzetti's Tavern","address":"6401 Morrison Blvd","city":"Charlotte","state":"NC","postal_code":"28211","lat
{"business_id":"0_ohldeFwysbglrTLSGM4Q","name":"The Lounge Barber Shop","address":"15224 N 59th Ave","city":"Glendale","state":"AZ","postal_code":"85306","l
{"business_id":"M9DM1ktbW-TB7nXu3Z4RDw","name":"Sonoma Village","address":"1318 S Vineyard","city":"Mesa","state":"AZ","postal_code":"85210","latitude":33.3
{"business_id":"n2kOsDur7tCLygSa87glJQ","name":"A Woman's Place","address":"2789 Sunridge Heights Pkwy, Ste 100","city":"Henderson","state":"NV","postal_coc

UTF-8, Line 1, Column 1 Tab Size: 4 JSON

This project deals with two big data set

business.json

192,609 109

59,371 100

- Ambience: Ambience_casual, Ambience_classy, etc.
- Best Nights: BestNights_friday, BestNights_monday, etc.
- Parking: Parking_garage, Parking_lot, Parking_street, etc.
- Dietary: DietaryRestrictions_dairy-free, etc.
- Others: Category, Delivery, Group, Price, Reservation, Open hour, etc.

review.json

1,015,774 9

480,562 4

business_id, stars, text, useful

This project contains 5 main steps

Data Exploration: Data Cleaning, Data Visualization

Feature Engineering: Feature Selection, Feature Aggregation, Feature extension

Model Training and Optimization

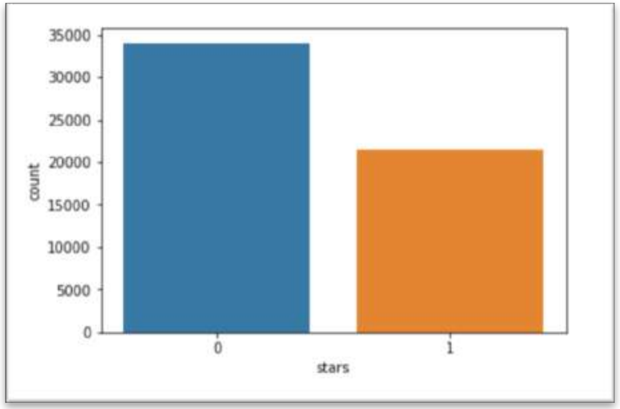
Model Evaluation: Accuracy, Confusion Matrix, ROC_AUC

Conclusion and Looking Forward

Data Exploration

Data Cleaning

1. Transform the json file into csv file
2. Extract the data with “category = restaurant”
3. Deleted the columns which have null > 1000, the rows which have null > 1000
4. The label is relatively balanced, and null values in both positive and negative labels are close. (P: 1690, N: 2207)
5. Tried to fill null with mode， but later we figured out that this method added to much noise. So we deleted all rows and columns with nulls. And our final dataset has 55457 rows.
6. Deal with different data types. Eg: True = 1, False=0
7. Deal with Operating Hours.



AR	AS	AT	AU	AV	AW	AX
hours.Friday	hours.Monday	hours.Saturday	hours.Sunday	hours.Thursday	hours.Tuesday	hours.Wednesday
9:0-19:0	9:0-19:0	10:0-18:0		9:0-19:0	9:0-19:0	9:0-19:0
10:0-21:0	10:0-21:0	10:0-21:0	12:0-18:0	10:0-21:0	10:0-21:0	10:0-21:0
10:0-19:0	10:0-19:0	9:0-17:0	11:0-17:0	10:0-19:0	10:0-19:0	10:0-19:0
8:0-21:0	8:0-21:0	10:0-21:0		8:0-21:0	8:0-21:0	8:0-21:0
10:0-21:0	10:0-21:0	10:0-21:0	11:0-19:0	10:0-21:0	10:0-21:0	10:0-21:0
8:0-17:0	8:0-17:0			8:0-17:0	8:0-17:0	8:0-17:0
5:0-22:0	5:0-22:0	8:0-18:0	9:0-16:0	5:0-22:0	5:0-22:0	5:0-22:0
10:0-3:0	10:0-22:0	10:0-3:0	10:0-22:0	10:0-3:0	10:0-22:0	10:0-3:0

Before

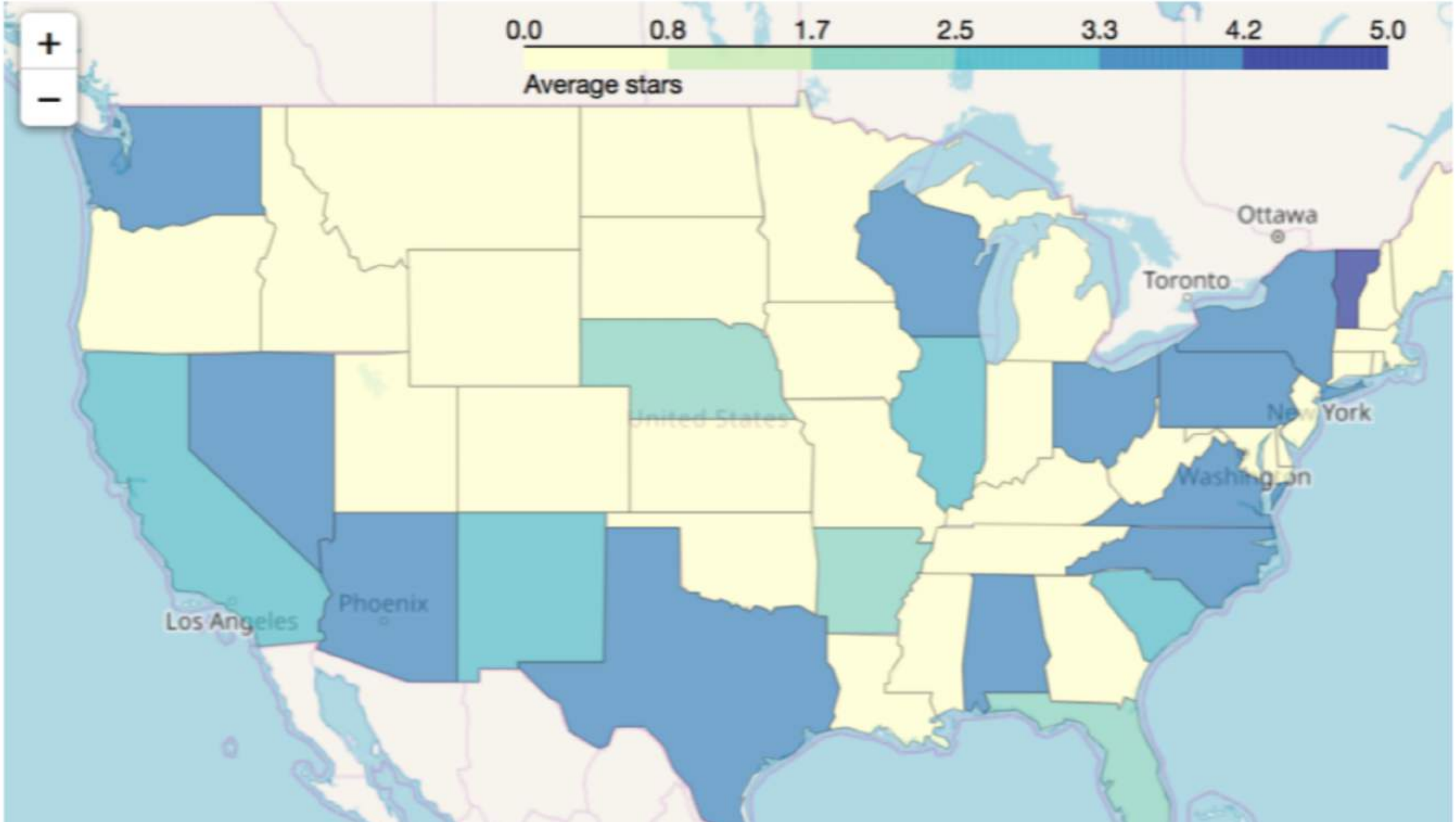
Breakfast	Brunch	Lunch	Dinner
0	0	1	1
0	0	1	1
0	0	1	1
0	0	0	1
1	1	1	1
1	1	1	0
0	1	1	1
1	1	1	1
1	1	1	1
0	0	0	1
0	0	0	1
0	1	1	1

+

Fridayophour	Mondayophour	Tuesdayophour	Wednesdayophour	Thursdayophour
7	0	7	7	7
7	0	7	7	7
10	0	10	10	10
7	13	6	6	7
20	19	19	19	19
7	7	7	7	7
10	10	10	10	10
10	10	10	10	10
18	18	18	18	18

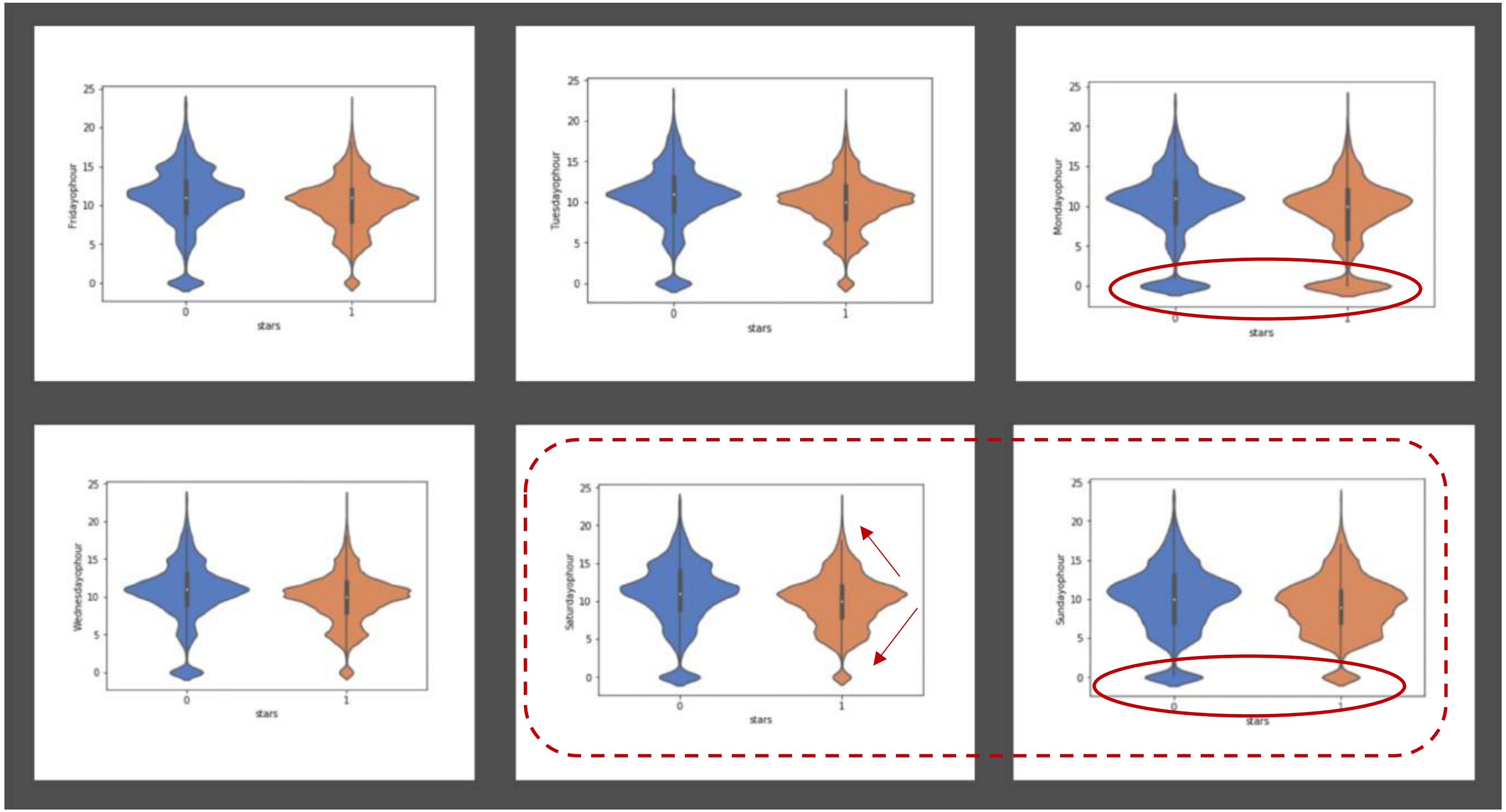
After

Data Visualization - Star Map

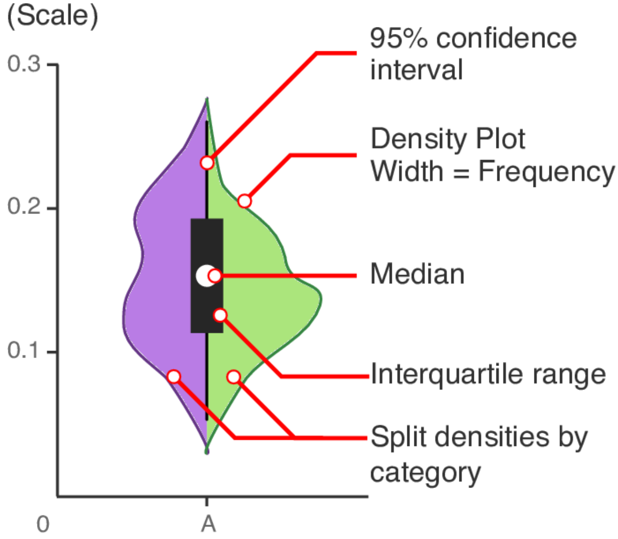


Different location has different average stars.

Data Visualization - Operating Hours



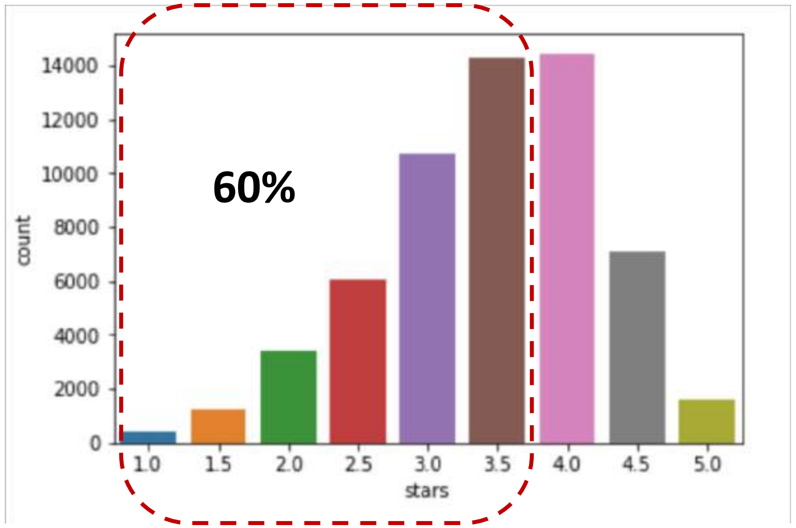
Notes for Violin Plot

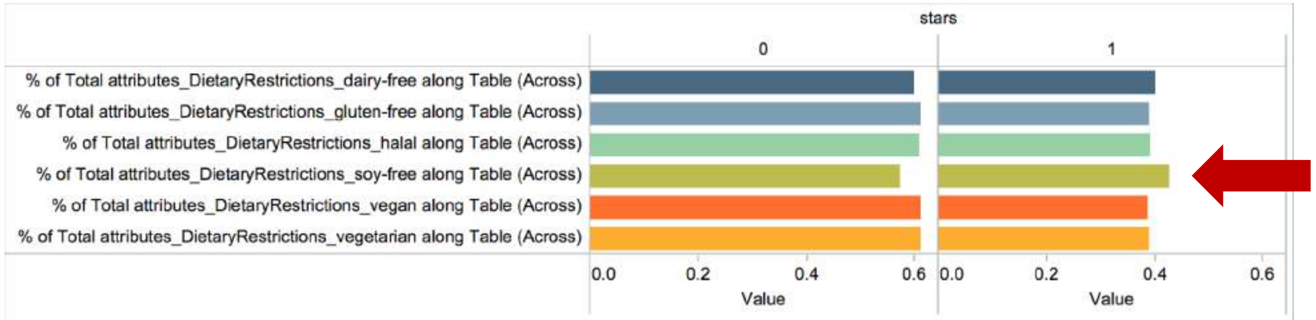


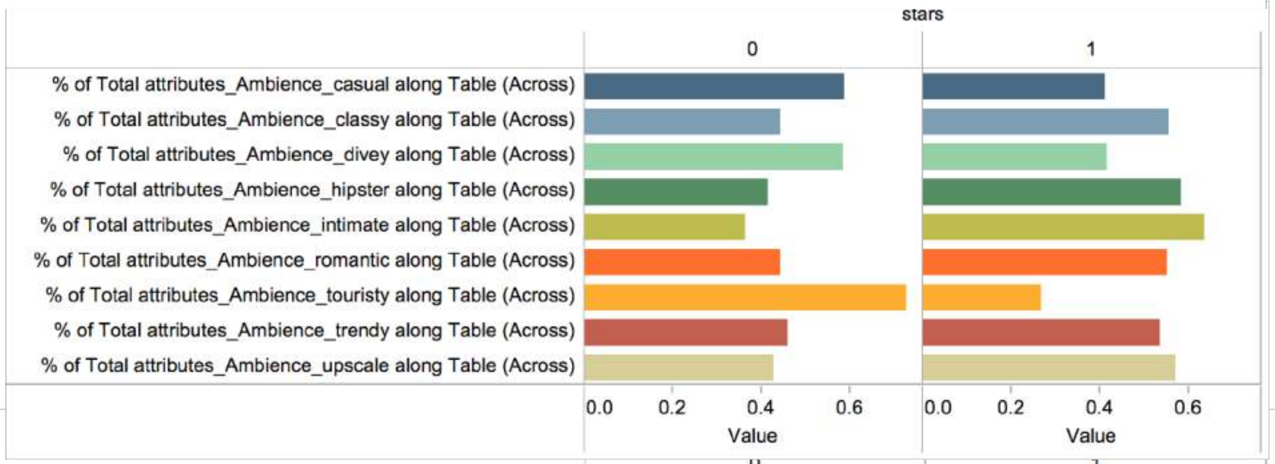
- 0 open hours
- The medium and mode of good restaurant is higher than the bad restaurant
- Difference between weekdays and weekends, Saturday and Sunday is “fat”

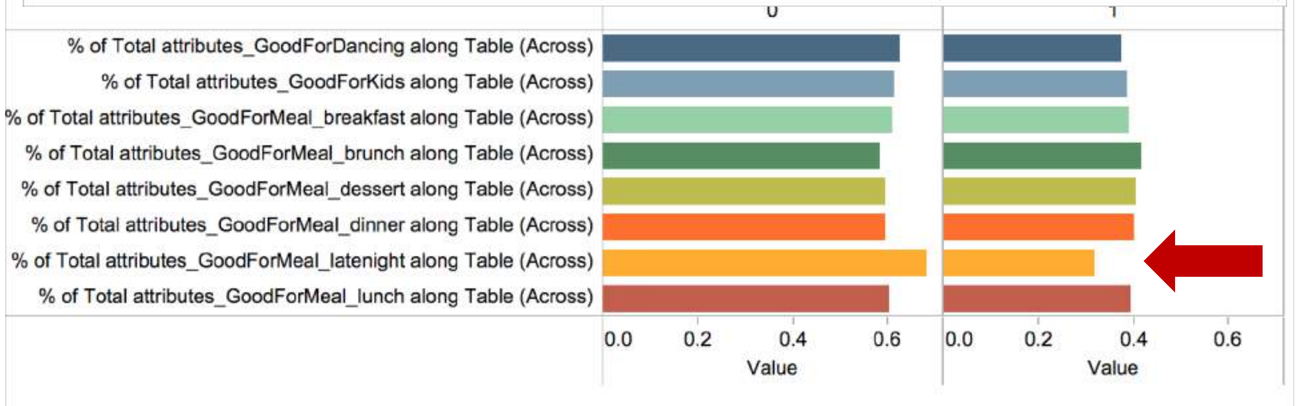
Data Visualization – Group Features

- Three group features:
  - Dietary Restrictions
  - Ambience
  - GoodFor
- “Dietary Restrictions”, ” GoodFor” - no significant influence on the label
- The relationship in the Ambience is not consistence, these group features should be significant









Feature Engineering

Model Performance is extremely low before Feature Engineering

Accuracy

```
Algo: Perceptron            and Score: 0.5555
Algo: LogisticRegression    and Score: 0.6602
Algo: Linear SVM            and Score: 0.6598
Algo: Decision Tree         and Score: 0.6376
Algo: Random Forest         and Score: 0.6885
Algo: Neural Net            and Score: 0.6701
Algo: Naive Bayes           and Score: 0.6273
Algo: Nearest Neighbors     and Score: 0.6072
Algo: XGBoost               and Score: 0.6798
```

AUC

```
In [12]:  from sklearn.metrics import roc_curve
          pipe = Pipeline( [ ('scl', StandardScaler()),
                             ('clf', RandomForestClassifier(max_depth=15, min_samples_leaf=1, max_features=0.3,
                                     min_samples_split=10, n_estimators=200, oob_score=False))])
          pipe.fit(X_train, y_train)
          fpr, tpr, thresholds = roc_curve(y_test, pipe.predict_proba(X_test)[:, 1])

In [13]:  from sklearn.metrics import auc
          auc(fpr, tpr)

Out[13]:  0.7176698801061321
```

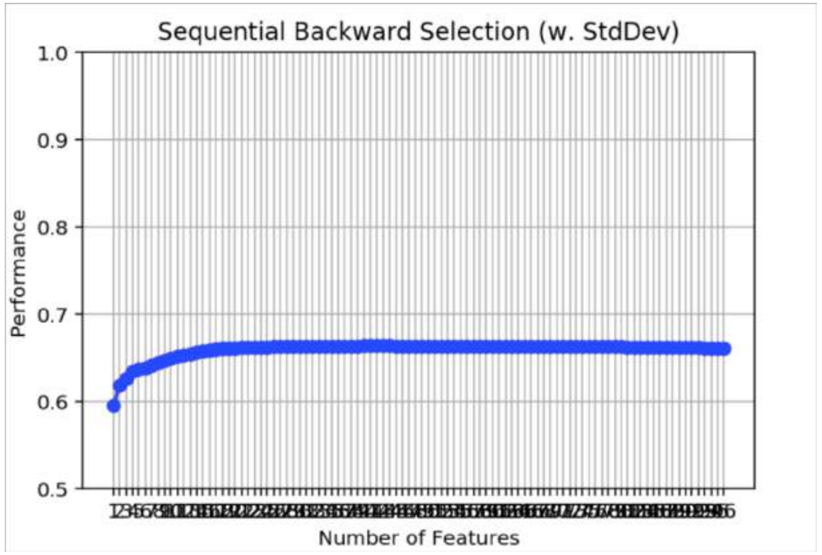
After data cleaning, we have run a rough model training based on all the features. Results show that, even the best model can only achieve an accuracy of **0.6885**, with AUC of **0.7177**.

Step 1: Feature Selection can reduce dimension without ignoring significant features

Wrapper Selection Method (Step Backwards)

Base Algorithms:

- LogisticRegression
- RandomForest

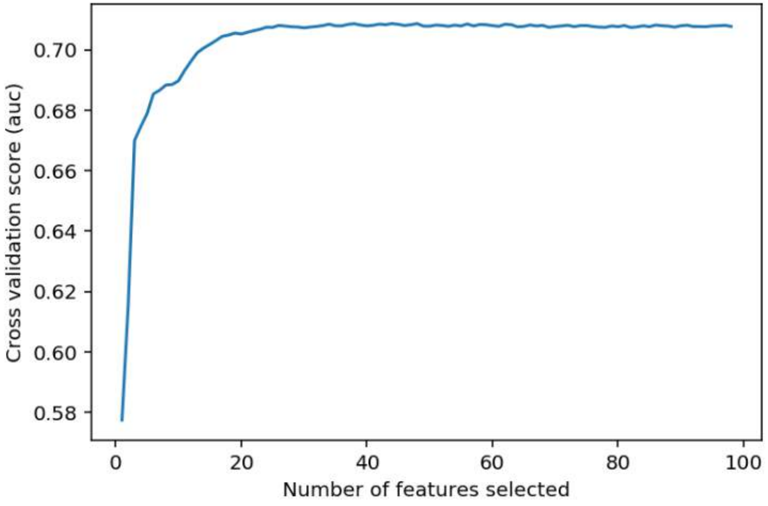


LogisticRegression (wrapper)

RFECV (Recursive Feature Elimination with Cross-Validation)

Base Algorithms:

- RandomForest



RandomForest (RFECV)

We selected 87 features combining all selected features

rf1 - 40	rf2 - 26	lg - 43
AgesAllowed	AgesAllowed	AgesAllowed
Alcohol	Alcohol	Alcohol
Ambience_casual	Ambience_casual	Ambience_casual
Ambience_classy	Ambience_classy	Ambience_classy
Ambience_divey	Ambience_divey	Ambience_divey
Ambience_hipster	Ambience_hipster	Ambience_hipster
Ambience_intimate	Ambience_intimate	Ambience_intimate
Ambience_romantic	Ambience_romantic	Ambience_romantic
Ambience_touristy	Ambience_touristy	Ambience_touristy
Ambience_trendy	Ambience_trendy	Ambience_trendy
Ambience_upscale	Ambience_upscale	Ambience_upscale
BYOB	BYOB	BYOB
BYOBCorkage	BYOBCorkage	BYOBCorkage
BestNights_friday	BestNights_friday	BestNights_friday
BestNights_monday	BestNights_monday	BestNights_monday
BestNights_saturday	BestNights_saturday	BestNights_saturday
BestNights_sunday	BestNights_sunday	BestNights_sunday
BestNights_thursday	BestNights_thursday	BestNights_thursday
BestNights_tuesday	BestNights_tuesday	BestNights_tuesday
BestNights_wednesday	BestNights_wednesday	BestNights_wednesday
BikeParking	BikeParking	BikeParking
AcceptsBitcoin	AcceptsBitcoin	AcceptsBitcoin
AcceptsCreditCards	AcceptsCreditCards	AcceptsCreditCards
Parking_garage	Parking_garage	Parking_garage
Parking_lot	Parking_lot	Parking_lot
Parking_street	Parking_street	Parking_street
Parking_valet	Parking_valet	Parking_valet
Parking_validated	Parking_validated	Parking_validated
Caters	Caters	Caters
CoatCheck	CoatCheck	CoatCheck
Corkage	Corkage	Corkage
DietaryRestrictions_dairy-free	DietaryRestrictions_dairy-free	DietaryRestrictions_dairy-free
DietaryRestrictions_gluten-free	DietaryRestrictions_gluten-free	DietaryRestrictions_gluten-free
DietaryRestrictions_halal	DietaryRestrictions_halal	DietaryRestrictions_halal
DietaryRestrictions_kosher	DietaryRestrictions_kosher	DietaryRestrictions_kosher
DietaryRestrictions_soy-free	DietaryRestrictions_soy-free	DietaryRestrictions_soy-free
DietaryRestrictions_vegan	DietaryRestrictions_vegan	DietaryRestrictions_vegan
DietaryRestrictions_vegetarian	DietaryRestrictions_vegetarian	DietaryRestrictions_vegetarian
DogsAllowed	DogsAllowed	DogsAllowed
DriveThru	DriveThru	DriveThru
GoodForDancing	GoodForDancing	GoodForDancing
GoodForKids	GoodForKids	GoodForKids
GoodForBreakfast	GoodForBreakfast	GoodForBreakfast
GoodForBrunch	GoodForBrunch	GoodForBrunch
GoodForDessert	GoodForDessert	GoodForDessert
GoodForDinner	GoodForDinner	GoodForDinner

Step 2: Feature Aggregation can aggregate features of low significance to generate a more powerful aggregated feature

Logistic Regression was used to aggregate sub-features

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5210	0.031	-16.884	0.000	-0.581	-0.461
DietaryRestrictions_dairy-free	0.0133	0.129	0.103	0.918	-0.239	0.266
DietaryRestrictions_gluten-free	0.0445	0.026	1.686	0.092	-0.007	0.096
DietaryRestrictions_halal	0.0223	0.068	0.326	0.744	-0.112	0.156
DietaryRestrictions_soy-free	0.1803	0.310	0.582	0.561	-0.427	0.787
DietaryRestrictions_vegan	0.0446	0.024	1.845	0.065	-0.003	0.092
DietaryRestrictions_vegetarian	0.0417	0.029	1.431	0.153	-0.015	0.099

In [129]:

```
df['DietaryRestrictions'] = np.exp(0.0445*df['DietaryRestrictions_gluten-free'] + 0.0446*df['DietaryRestrictions_vegan'])
```

Example of 'DietaryRestrictions' sub-features aggregation

We have aggregated 7 sub-features to aggregated features:

- Wifi
- Smoking
- DietaryRestrictions
- BestNights
- Parking
- Alcohol
- Attire

Step 3: Use Yelp Reviews dataset to add *'Topic Score'* features into the feature space

Yelp Reviews Topic Model:
Latent Dirichlet Allocation (LDA)

What are the most important topics customers care the most? How can restaurant manager catch up with the trend?

We obtained the top 10 topics trained from the reviews as followings:

	Key Words	Topic Abstract
Topic0	vegas hotel strip las lamb floor casino stay pool club rooms hummus greek check dance bathroom line staying resort	Hotel_related (upper-level)
Topic1	waitress manager customer left walked waiting guy waiter customers water check owner seated friend looked arrived	Restaurant Service
Topic2	pork fried spicy dish beef shrimp bbq chinese thai bowl tea green curry broth crispy noodle korean tender tofu rolls	Asian Food
Topic3	breakfast cream chocolate ice cake dessert butter fruit desserts sugar waffles apple cup pie strawberry flavors bakery	Disserts
Topic4	tacos buffet chips mexican line taco salsa decent places prices reviews star half extra pay burrito items tasted fast	Mexican Food
Topic5	pizza happy beer wings tables party group friends patio large music inside wine fun door game friend friday decided	Group Activity & party
Topic6	fries burger bread steak sandwich cooked wine dish served pasta burgers italian grilled bacon dessert garlic appetizer	American Cusine
Topic7	vegan com kids drive use shop room card old kid need located gluten counter home parking park cash mall local	Resturant Feature & Extra Amenity
Topic8	sushi roll fish crab ramen rolls tuna salmon japanese lobster sashimi oysters seafood miso poke shrimp foie legs pour	Seafood & Japanese Food
Topic9	feel location atmosphere selection visit awesome excellent quite enjoy places high perfect inside open town dining	Environment

Step 3: Use Yelp Reviews dataset to add *'Topic Score'* features into the feature space

Topic$_i$ Score for a restaurant: $$TopicScore_i = \frac{sum(Positive) - sum(Negative)}{sum(Topic\ Related\ reviews)}$$

Example: *First Watch*

	business_id	text	useful	stars
711	clnZkUSckKwxCqAR7s2ETw	I've been here twice to this same location. Th...	3.0	4.0
712	clnZkUSckKwxCqAR7s2ETw	First Watch is my first encounter with a chain...	3.0	3.0
713	clnZkUSckKwxCqAR7s2ETw	First Watch was so delicious and much needed a...	4.0	4.0
714	clnZkUSckKwxCqAR7s2ETw	Super healthy breakfast foods. Super tasty to...	5.0	4.0
715	clnZkUSckKwxCqAR7s2ETw	This place rules!\nI have never had even a sli...	3.0	5.0

715 row's review is about *Topic 9: Environment*, it talks about clean place and good environment and no other reviews about the environment. Therefore, this restaurant's environment Topic score is calculated as (1 – 0) / 1 = 1, shown as below:

	business_id	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9
18652	clnZkUSckKwxCqAR7s2ETw	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0

65 Features have been selected as the final feature space

Ambience_casual	Ambience_classy	Ambience_divey	Video
Ambience_hipster	Ambience_intimate	Ambience_romantic	Price
Ambience_touristy	Ambience_trendy	Ambience_upscale	review_count
BikeParking	AcceptsCreditCards	Caters	Dinner
DietaryRestrictions_kosher	DriveThru	GoodForKids	Thursdayophour
GoodForBreakfast	GoodForBrunch	GoodForDinner	Fridayophour
GoodForLatenight	GoodForLunch	HasTV	Smoking_no
Outdoor	Delivery	Group	Parking
TakeOut	latitude	longitude	DietaryRestrictions
Breakfast	Brunch	Lunch	Noisy
Mondayophour	Tuesdayophour	Wednesdayophour	Reservation
Saturdayophour	Sundayophour	Alcohol_beer_and_wine	
Alcohol_full_bar	Attire_casual	Attire_dressy	Earlynight
Smoking_outdoor	Wifi_free	Wifi_no	TableService
Topic0	Topic1	Topic2	Topic3
Topic4	Topic5	Topic6	Topic7
Topic8	Topic9		

Model Training and Optimization

Model Training

Train-test-split: **70% train; 30% test.**

10 Models tested, 5 Models Used:

- **Logistic Regression(77.23%)**
- **Linear SVM(77.57%)**
- **Random Forest(78.10%)**
- **Neural Net(78.05%)**
- **XGBoost (with tree booster) (77.83%)**

Algo: Perceptron	and Score: 0.6853
Algo: LogisticRegression	and Score: 0.7723
Algo: Linear SVM	and Score: 0.7757
Algo: Decision Tree	and Score: 0.7356
Algo: Random Forest	and Score: 0.7810
Algo: RBF SVM	and Score: 0.6117
Algo: Neural Net	and Score: 0.7805
Algo: Naive Bayes	and Score: 0.6817
Algo: Nearest Neighbors	and Score: 0.6881
Algo: XGBoost	and Score: 0.7783

Model Training – Application of PCA

With PCA(n=55)

```
Algo: LogisticRegression    and Score: 0.7727
Algo: Linear SVM            and Score: 0.7719
Algo: Random Forest         and Score: 0.7589
Algo: Neural Net            and Score: 0.7767
Algo: XGBoost               and Score: 0.7685
```

Without PCA

```
Algo: LogisticRegression    and Score: 0.7723
Algo: Linear SVM            and Score: 0.7714
Algo: Random Forest         and Score: 0.7830
Algo: Neural Net            and Score: 0.7763
Algo: XGBoost               and Score: 0.7783
```

According to the performance on the test set, we can see that PCA seems to slightly improved performance of logistic regression and linear SVM, but random forest, neural network, and XGBoost perform much better without PCA. Therefore, we decided to forgo PCA.

Model Optimization-GridSearch

Random forest (78.30% to 78.46%):

```
params = {
            'clf__max_depth': list(range(10, 21)),
            'clf__n_estimators' : [200, 500],
            'clf__max_features': ["sqrt", 0.4, 0.3, 0.2],
            'clf__oob_score': [False],
            'clf__min_samples_split':[5, 10, 15],
            'clf__min_samples_leaf': [1, 2, 5]
        }
```

```
{'clf__max_depth': 15, 'clf__max_features': 0.3, 'clf__min_samples_leaf': 1, 'clf__min_
samples_split': 10, 'clf__n_estimators': 200, 'clf__oob_score': False}
```

Model Optimization-GridSearch

XGBoost (77.83% to 78.87%):

```
params = {'clf__eta'       : [0.005, 0.01, 0.015],
          'clf__max_depth ': [2, 5, 10, 15],
          'clf__gamma ': [0, 2, 5],
          'clf__subsample ': [0.3, 0.5, 0.6, 0.8, 1],
          'clf__n_estimators': [500, 1000]
}
```

```
{'clf__eta': 0.005, 'clf__gamma ': 0, 'clf__max_depth ': 2, 'clf__n_estimators': 1000,
'clf__subsample ': 0.3}
```

Optimized model

XGBoost: performance improved from 67.98%(before feature engineering) to 78.87%

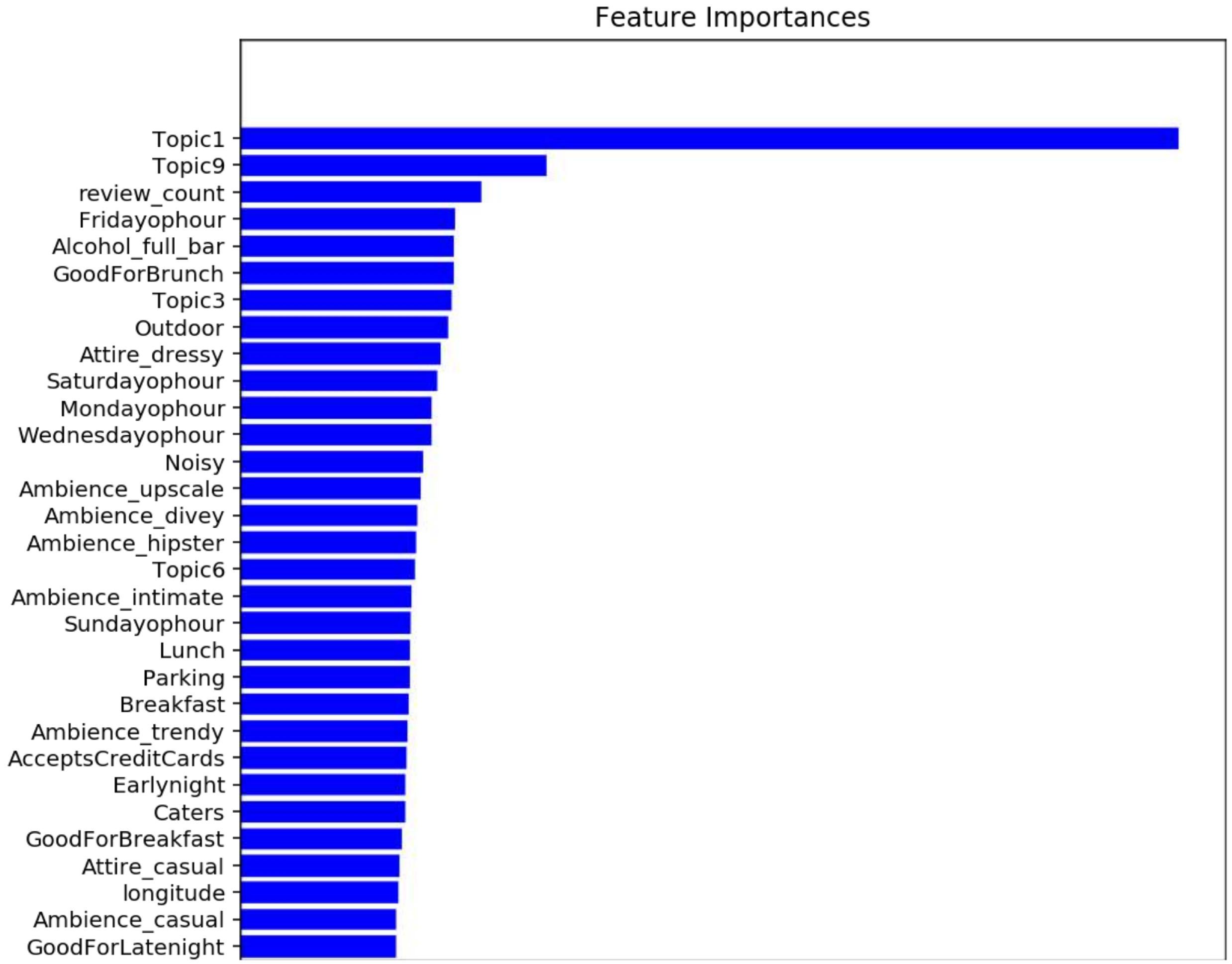
{'clf__eta': 0.005, 'clf__gamma ': 0, 'clf__max_depth ': 2, 'clf__n_estimators': 1000,
'clf__subsample ': 0.3}

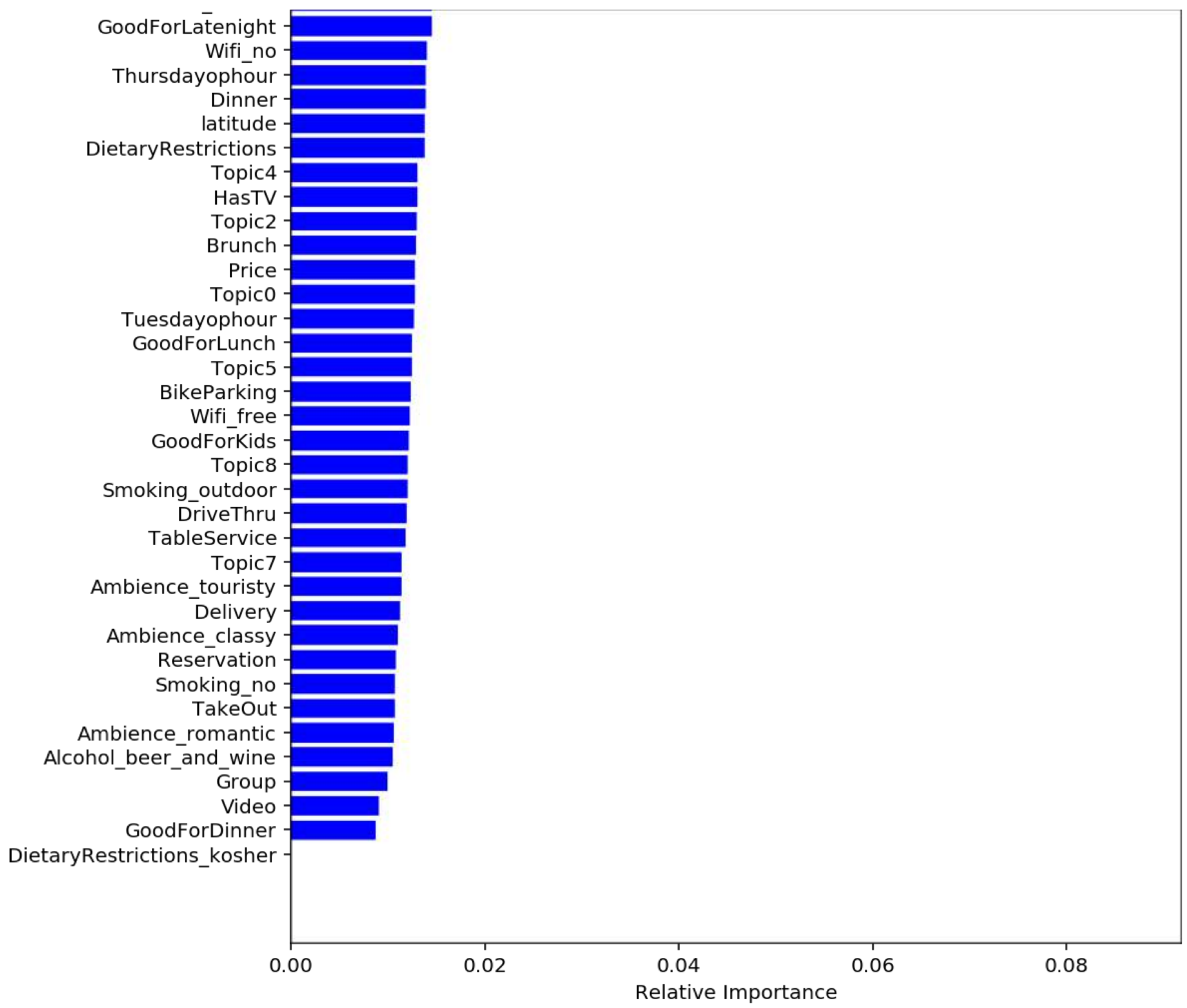
Model Optimization-ensemble learning

Plurality voting (78.93%)
Stacking with random forest(78.77%)
(models used: random forest, neural net, xgboost)

Bagging(78.86%)
(model used: XGBoost)

Optimized model – feature importances





Model Evaluation

Model Evaluation – Accuracy & Model Running Time

Accuracy Score

```
Algo: Random Forest         and Accuracy Score: 0.7880
Algo: Neural Net            and Accuracy Score: 0.7821
Algo: XGBoost               and Accuracy Score: 0.7887
```

Running Time

```
import datetime

for name, clf in zip(names, classifiers):
    pipe = Pipeline( [ ('scl', StandardScaler()),
                       ('clf', clf)])
    last_time = datetime.datetime.now()
    pipe.fit(X_train, y_train)
    pipe.score(X_test, y_test)
    current_time = datetime.datetime.now()
    print("Algo: {0:<20s} and Running time: {1}s".format(name, (current_time-last_time).seconds))
```

```
Algo: Random Forest         and Running time: 17s
Algo: Neural Net            and Running time: 22s
Algo: XGBoost               and Running time: 12s
```

Model Evaluation – Confusion Matrix

Random Forest	Positive (Predicted)	Negative (Predicted)
Positive (Actual)	8867	1389
Negative (Actual)	2143	4239

+ = 13,106

Neural Net	Positive (Predicted)	Negative (Predicted)
Positive (Actual)	7811	2445
Negative (Actual)	1528	4854

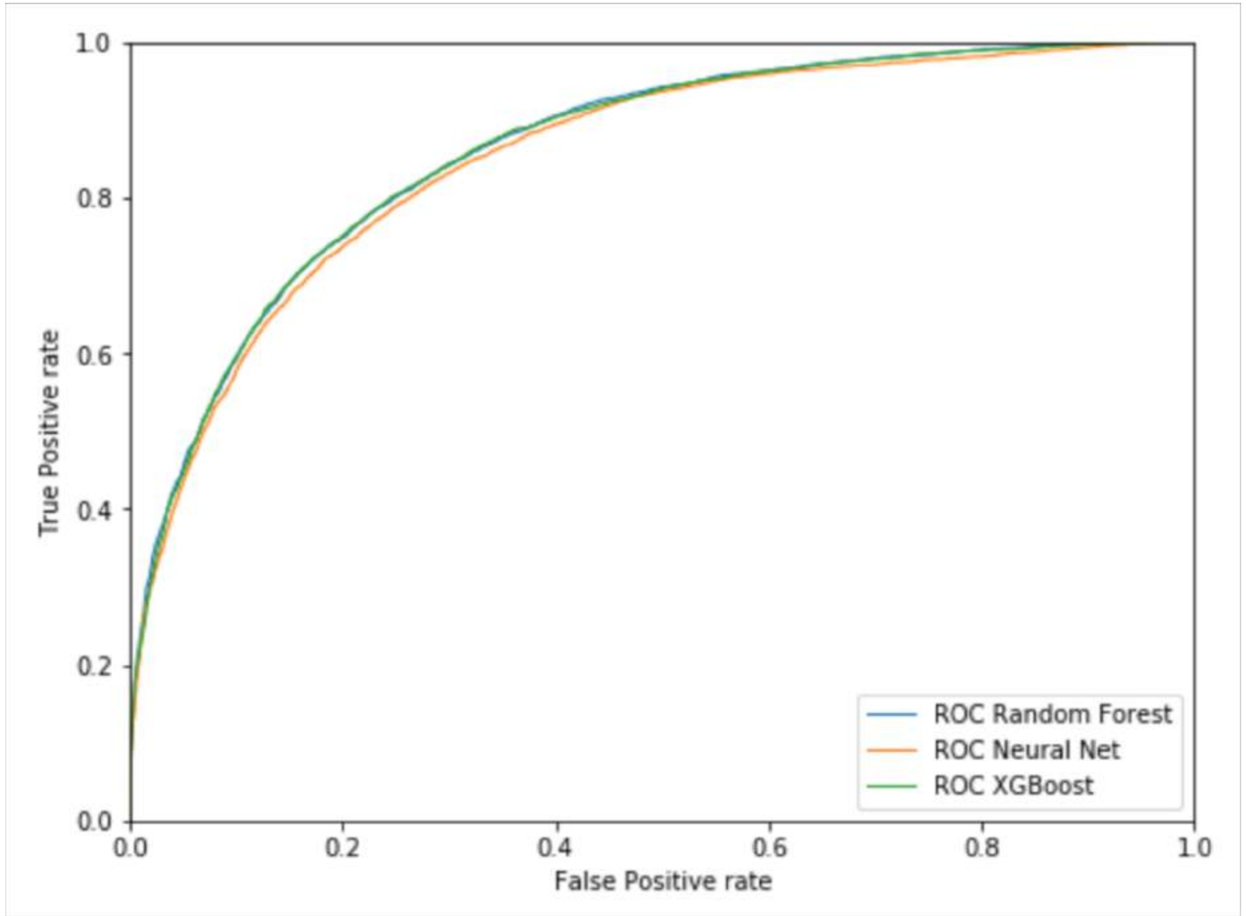
+ = 12,665

XG Boost	Positive (Predicted)	Negative (Predicted)
Positive (Actual)	8709	1547
Negative (Actual)	1970	4412

+ = 13,121

Model Evaluation – ROC_AUC

ROC Curve



AUC Score

Random Forest: 0.8612

Neural Net: 0.8520

XG Boost: 0.8605

Insights and Limitations

Insights & Limitations

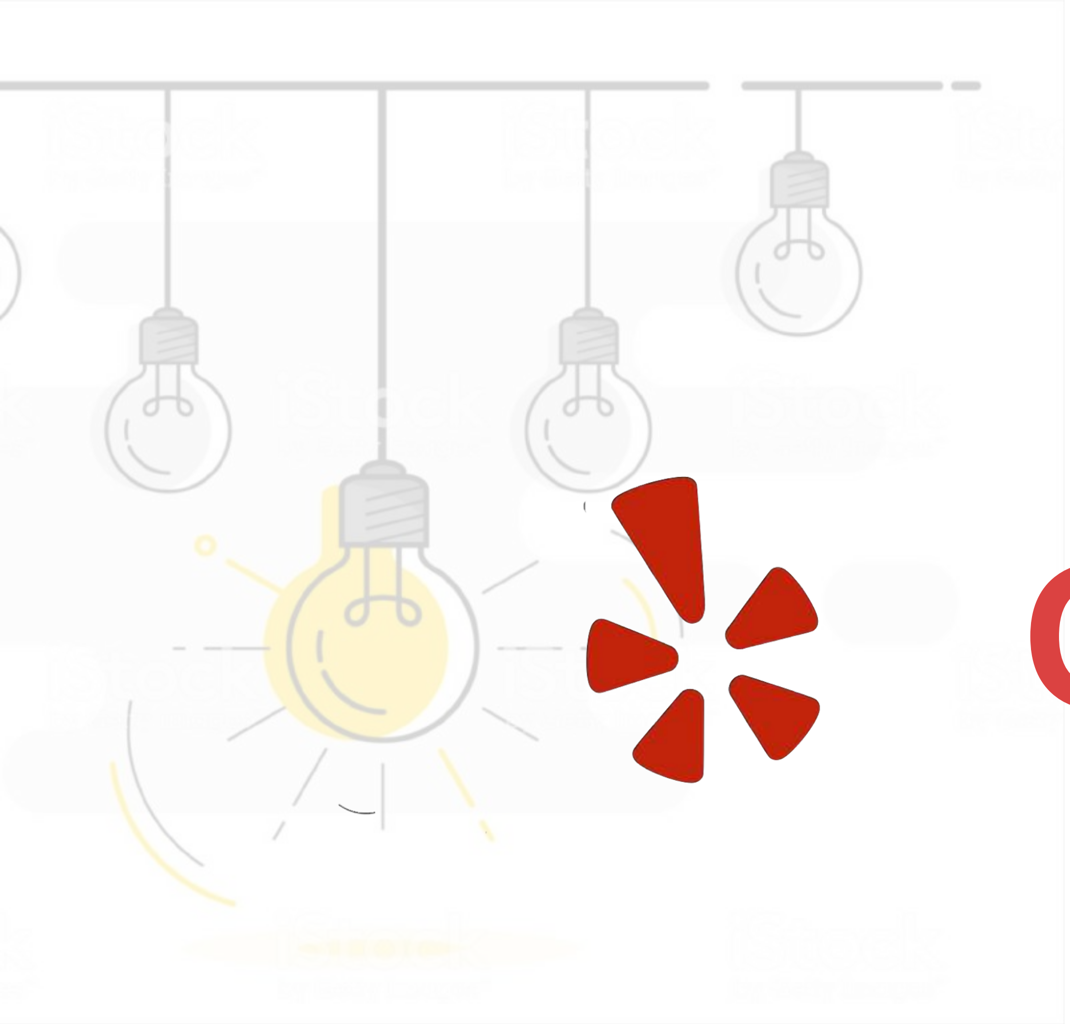
Insights

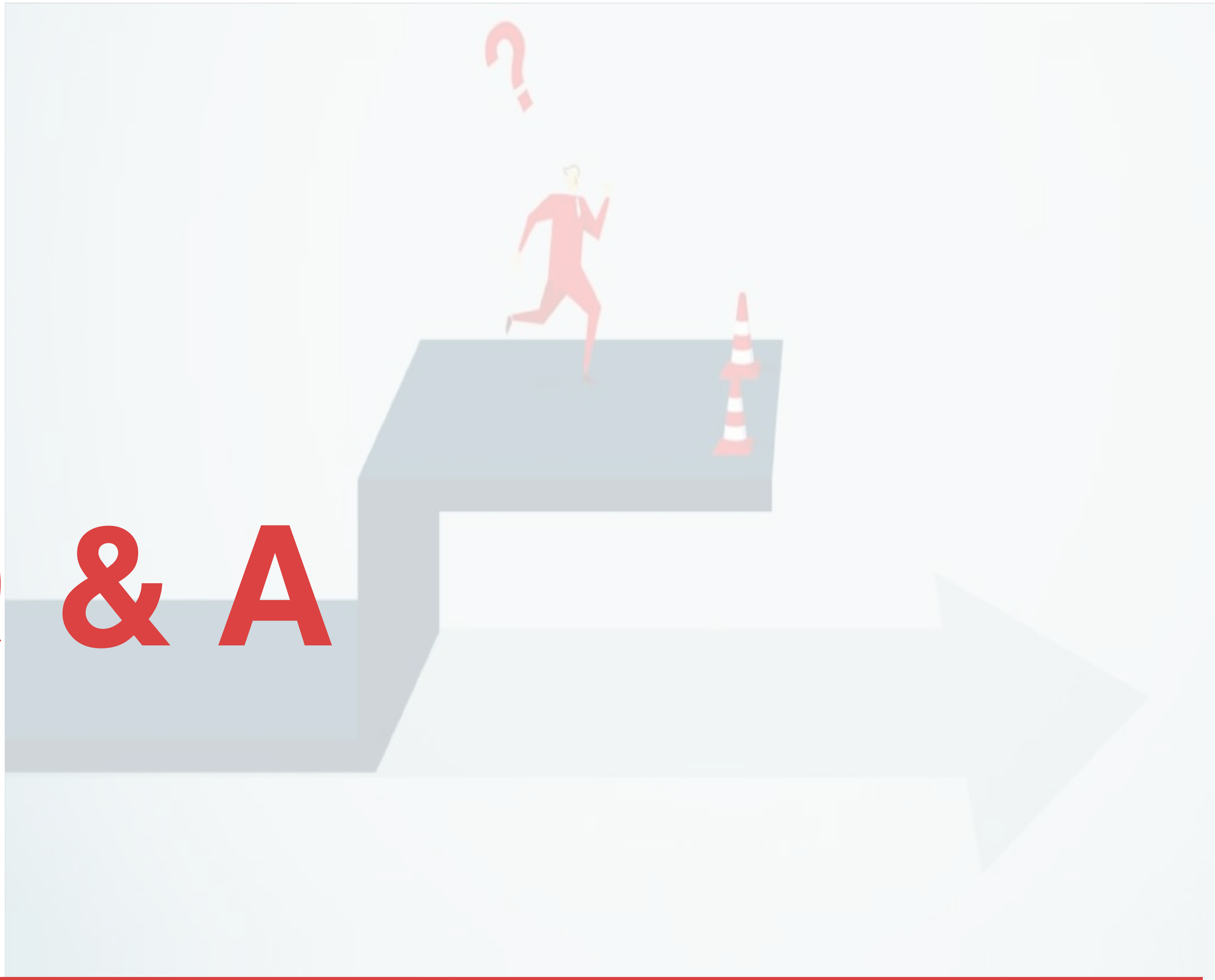
- Restaurant owners can be more cost-effective by focusing on the important features first
- Top important features: Environment, Service, American Cuisine, Review counts, Asian food, Mexican food, Saturday op hours, Monday & Friday hours, Longitude, and Group activity space

Limitations

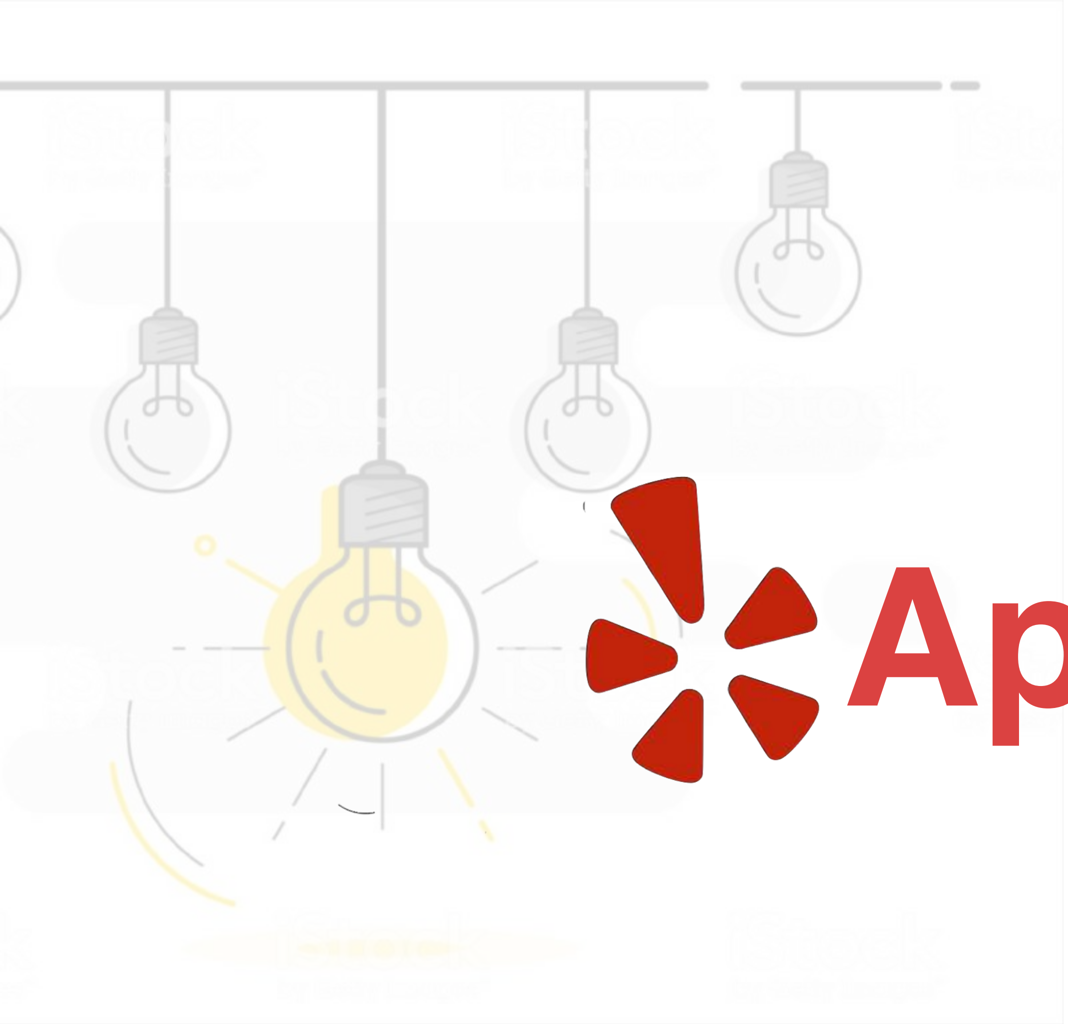
- There are certain noise level in the dataset
- The different weights of user's comments could be considered
- More features are needed to fully explain the remaining 20%

Original Model Performance: 67% ⟹ Current Model Performance: 79%





Q & A



Appendix

Appendix

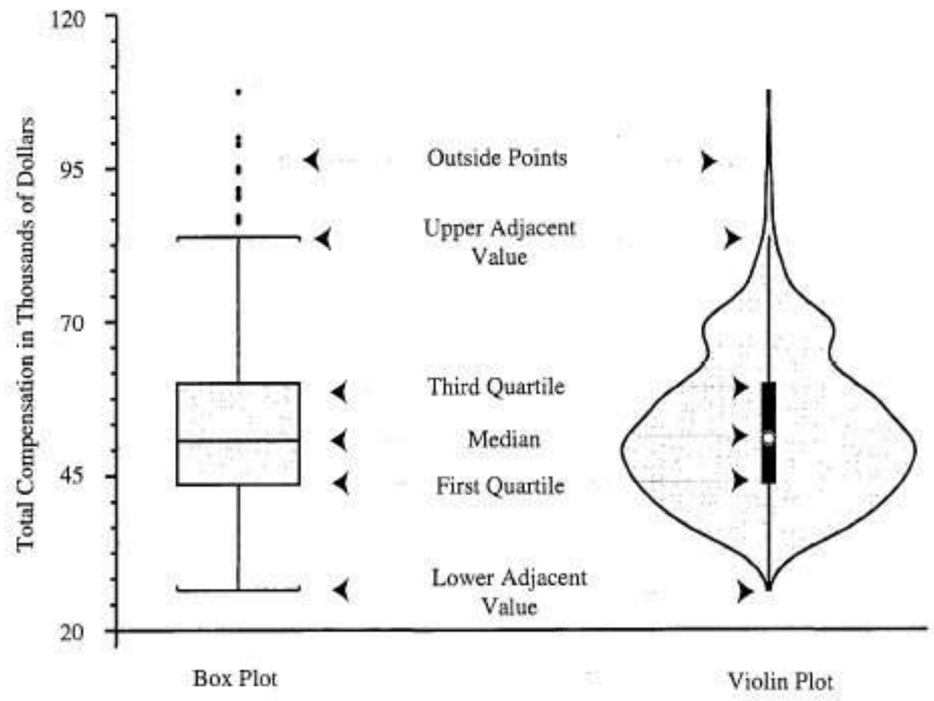
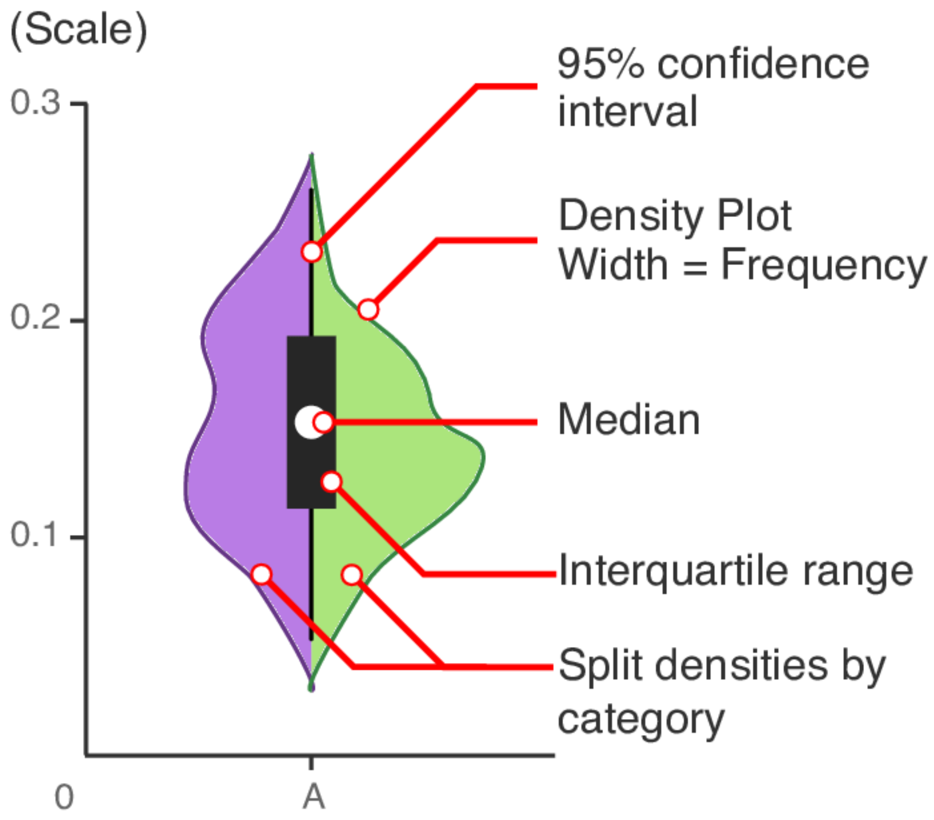
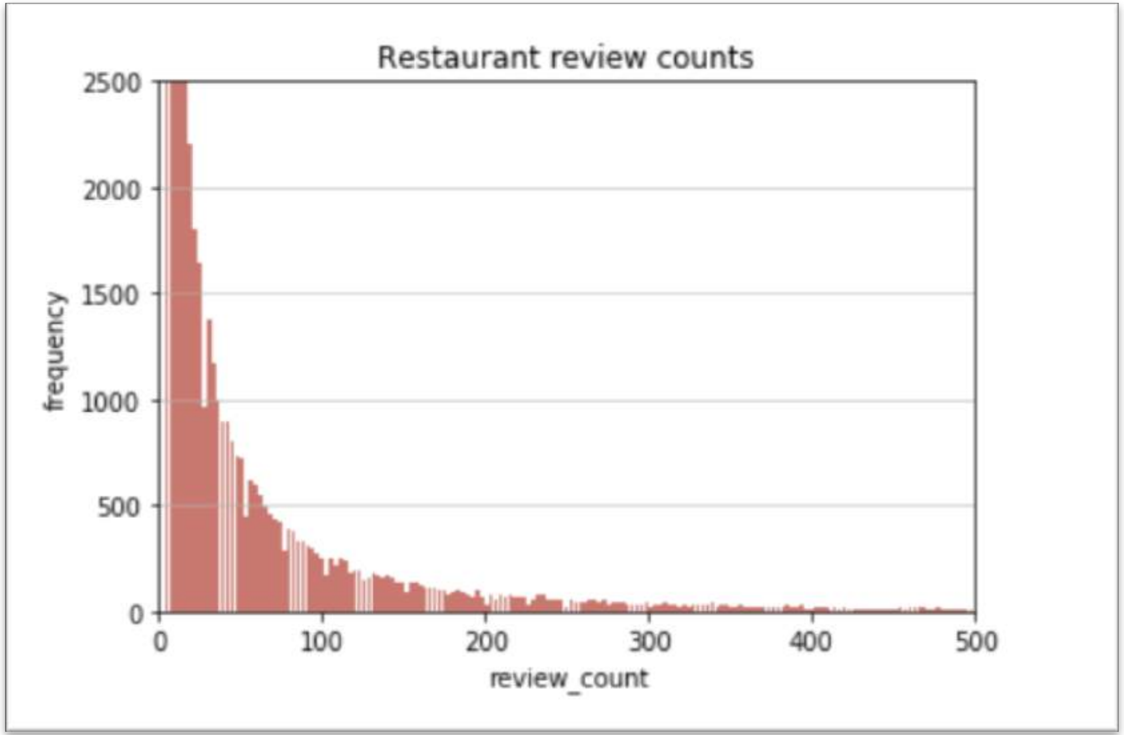
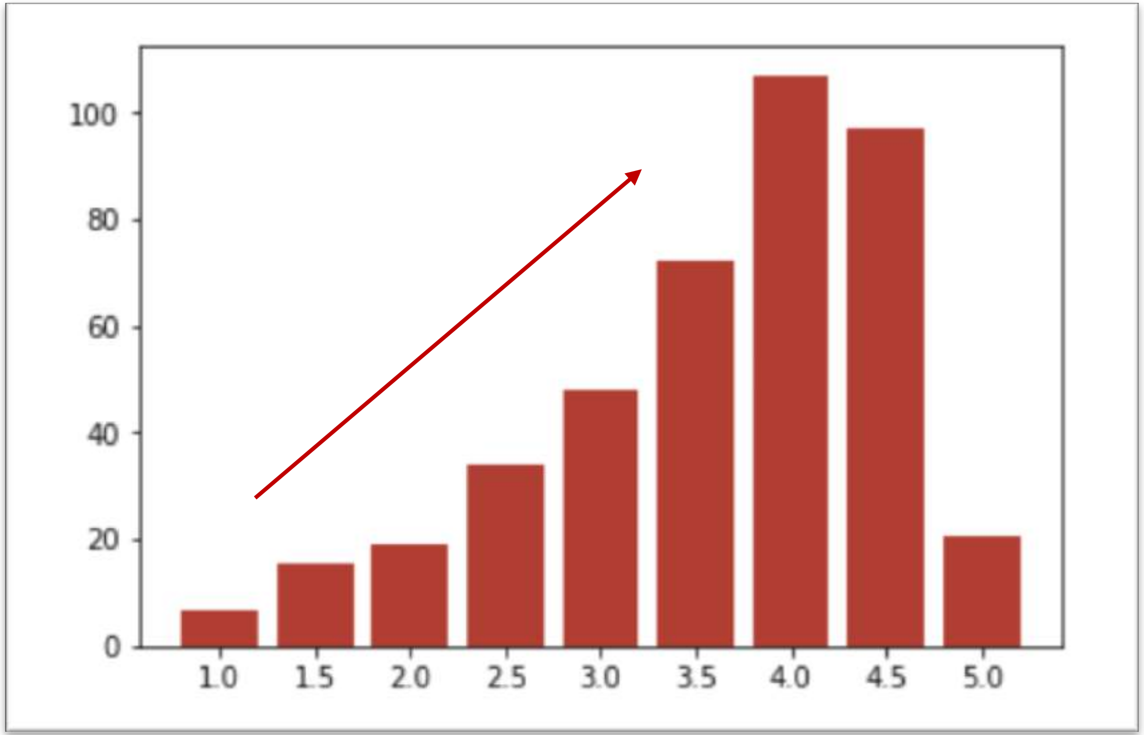


Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.



Data Visualization – Review Counts





- Based on the left figure, we found out that 1/3 of restaurants has lower than 8 reviews.
- In the right figure, we could tell that the average numbers of review increase as the rating stars increasing.

Appendix – Confusion Matrix

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Appendix – Average Star Rating by Industry