How to open a coffee & tea shop successfully?

Team 8: Jiawen Zhang, Kaiwen Hu, Yijun Zhou

TABLE OF CONTENTS

1. Project Summary - Background

1. 64% of American adults currently consume coffee every day.

(Reuters, NCA)

According to a study conducted by the NCA, this is the highest rate since 2012. In our infographic about bizarre sleep habits, you can see that some famous writers were also regular coffee drinkers.

1. Project Summary

This project aims to develop a system to evaluate important factors that affect the **rating(response variable)** of coffee & tea shops on popular social media review systems–Yelp and Google Map. In the end, we will examine how **potential features(price, location, open hours, photos, reviews, search/popularity trend, etc)** impact the rating of the coffee & tea shop.

2. Data Description - overview

In the data set, we have **6,243** coffee & tea shops and **220+** distinct features.

2. Data Description - features

Static features (groupby **shop**):

- Shops' **basic information** like: address, whether provide takeout…
- **Number of shops** nearby
- **Yolo result** from detecting review pictures
- **Demographic Information** in state level

	business_id text	latitude double precision	longitude double precision	review_count double precision	price_level double precision	HasTV bigint	less_50km bigint	class_0 bigint	class_1 bigint
1	00rY5F9ltW-IWf2Ev96kOg	39.7791327	-86.1645255	277	2	0	466	22	0
2	00sOoojttpdZIjH8VgOU0A	27.9963217	-82.3726623	17	1	2	1090	0	0
3	018SgjlLDCKLR7gFSEkbGQ	35.9087666	-86.8843908	9	1.39	2	460	[null]	[null]
4	02nb6CI8w-2EoSEkQdk2Wg	39.9361704128	-75.1469422175	101	1	2	2152	4	0
5	02nUjwVmJGTgGyiIi-hklg	39.9525122617	-75.1717417315	11	1.39	2	2184	0	0

2. Data Description - features

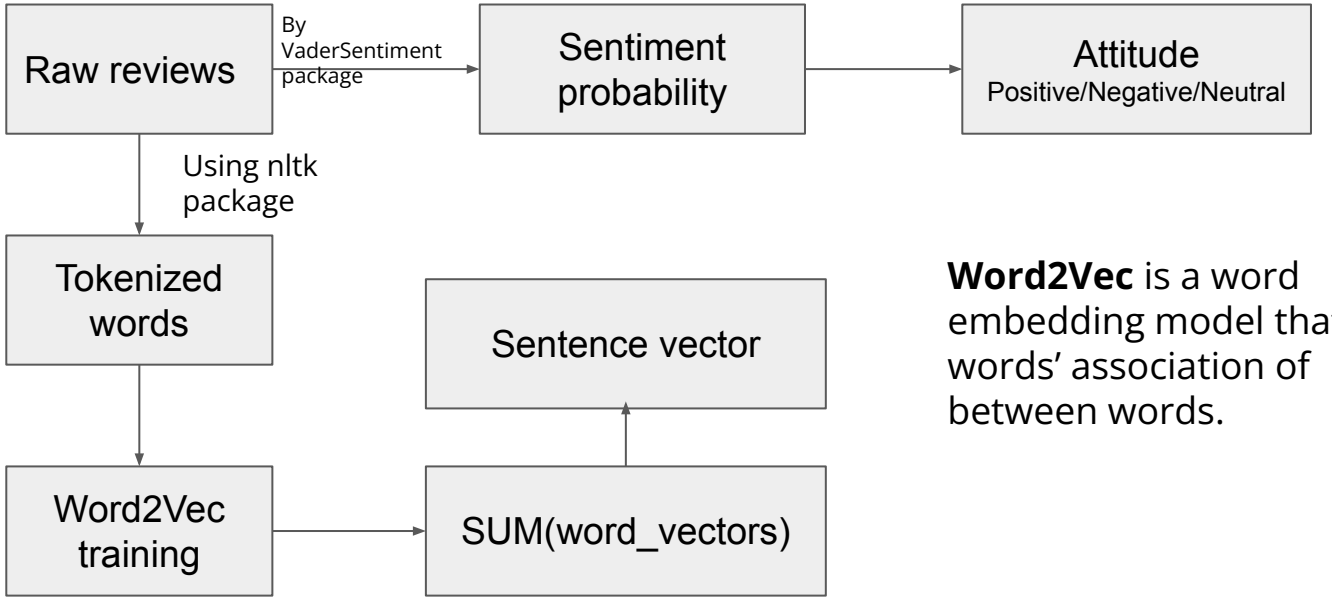
Dynamic features (averaged by **month**):

- Review **vectors** <- Word2Vec
- Review **attribute** like useful, funny
- **Previous rating** (or rating from last month)
- **Searching Index** of Coffee & Tea for this month
- Simple Sentiment result (positive,negative count) <- Vadar

Target variable:

	business_id text	year-month text	useful_review double precision	funny_review double precision	changing_rating double precision	v_0 double precision	v_1 double precision
1	-0epFLgYq2C1Jo_W4FOBKw	2012-07	1	0	5	-0.08564641478257778	0.10195921861377114
2	-0epFLgYq2C1Jo_W4FOBKw	2012-08	2	0	4.5	-0.06783923277808301	-0.4727796292889801
3	-0epFLgYq2C1Jo_W4FOBKw	2013-07	2	0	4.571428571428571	0.2837716763483032	-0.029454608160235426

2. Data Description - features

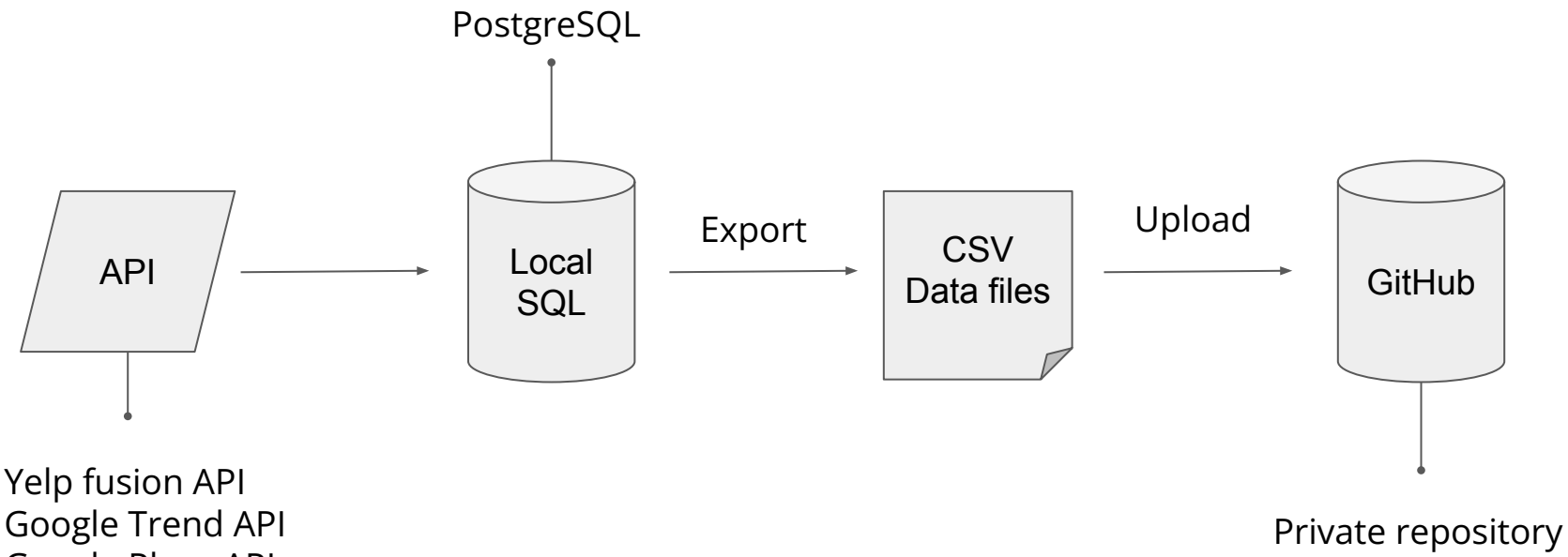


Word2Vec is a word embedding model that learn words' association of between words.

Sentence Vector example:

v_0	v_1	v_2	v_3	v_4	v_5	v_6	...	v_90	v_91	v_92	v_93	v_94
0.190541	-0.534683	0.424322	-0.731165	1.229790	0.464220	0.591767	...	-0.499821	-0.139621	1.353730	0.509824	-0.335431
0.821401	-0.667513	0.503169	-0.728497	1.548672	0.393993	-0.216269	...	-1.045970	-0.148073	0.789526	0.318352	0.284340

2. Data Management - collecting



- Yelp fusion API
- Google Trend API
- Google Place API
- Census Bureau API

2. Data Management - distributing

Direct data packages:

yelp_academic_dataset_business.json	116,078 KB
yelp_academic_dataset_checkin.json	280,234 KB
yelp_academic_dataset_review.json	5,216,669 KB
yelp_academic_dataset_tip.json	176,372 KB
yelp_academic_dataset_user.json	3,284,501 KB

import

Download & Execute Directly To Google Cloud

GitHub

Setup files:

- yelp_data_setup.sql
- yelp_import.ipynb

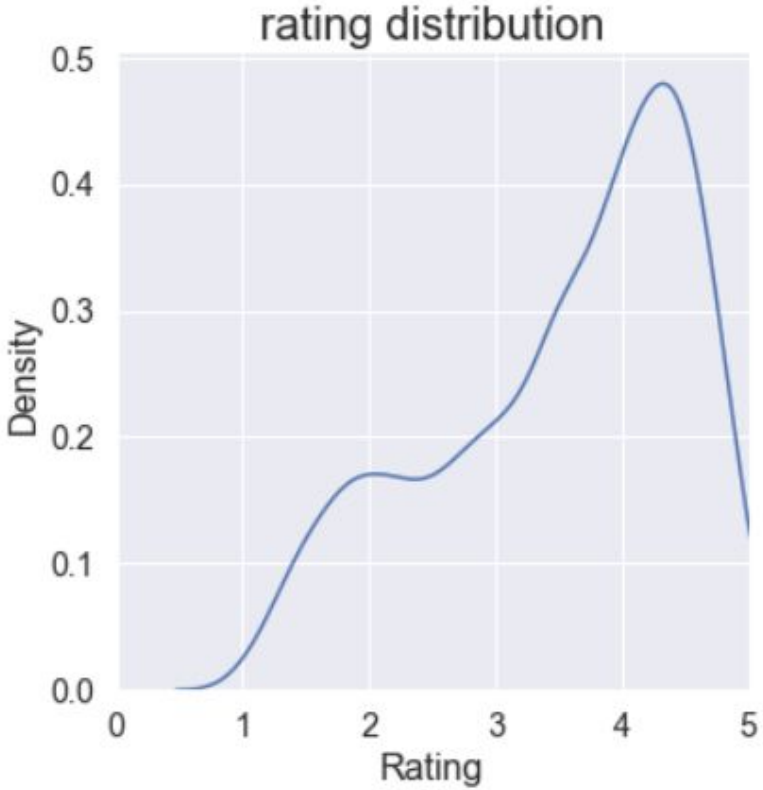
Google Cloud PostgreSQL

- census_Bureau_final
- census_bureau
- census_bureau_clean
- google_place_distance
- google_place_info
- google_trend
- state_code
- yelp_checkin
- yelp_coffee_info
- yelp_coffee_info_processed

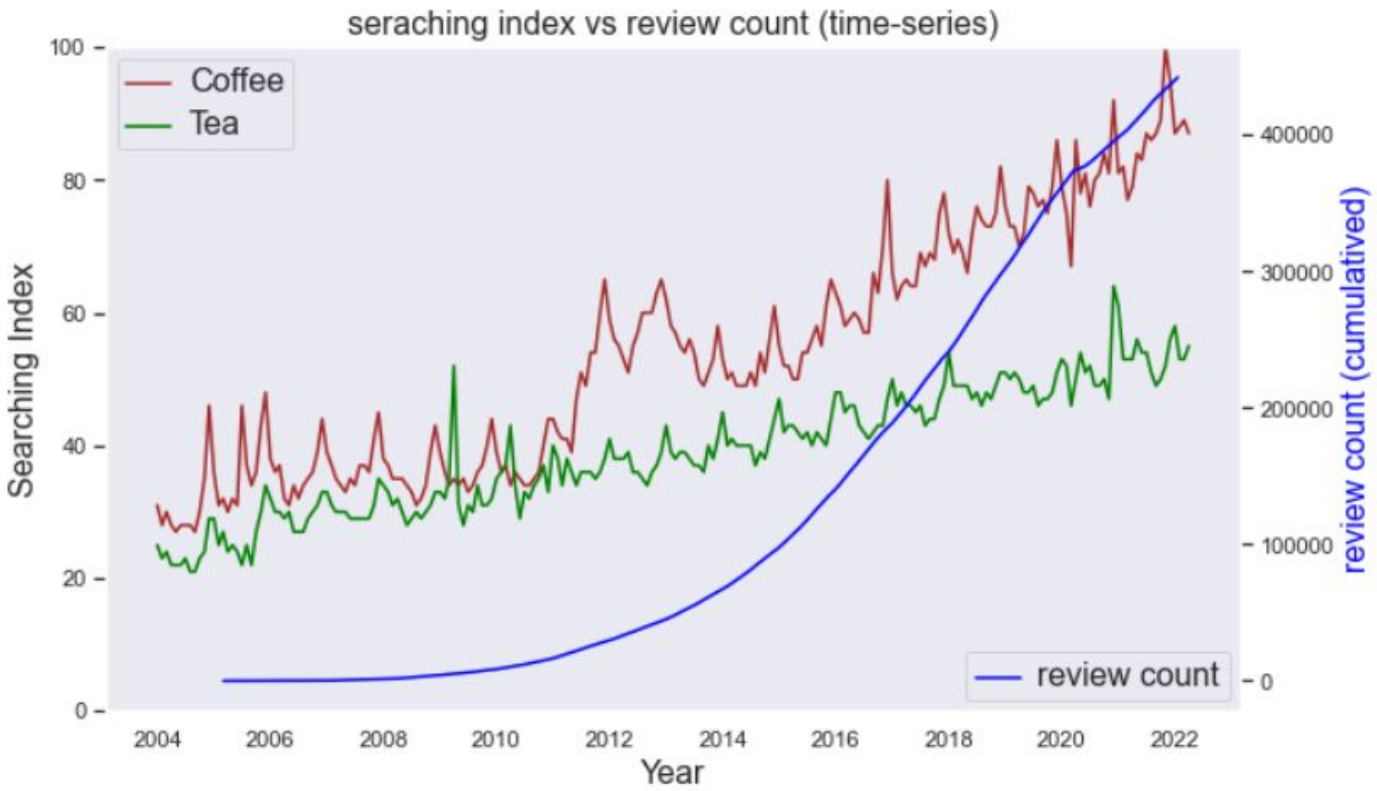
Data files:

- coffee_info.csv
- google_place.csv

3. Data Visualization - rating



3. Data Visualization



4. Model - Overview

- **Multiple Linear Regression**
- **Penalized Regression**
  - **Lasso**
  - **Ridge**
  - **Elastic Net**
- **Autoregressive**

4. Model - Multiple Linear Regression

Multiple Linear Regression - good fit

- Uses two or more independent variables to predict the outcome of a dependent variable
- $$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1,i} + \hat{b}_2 X_{2,i} + \ldots + \hat{b}_p X_{p,i}$$
- Assumption check
  - Linearity: The relationship between X and the mean of Y is linear.
  - Homoscedasticity: The variance of residual is the same for any value of X.
  - Independence: Observations are independent of each other.
  - Normality: For any fixed value of X, Y is normally distributed.

4. Multiple Linear Regression Preliminary results

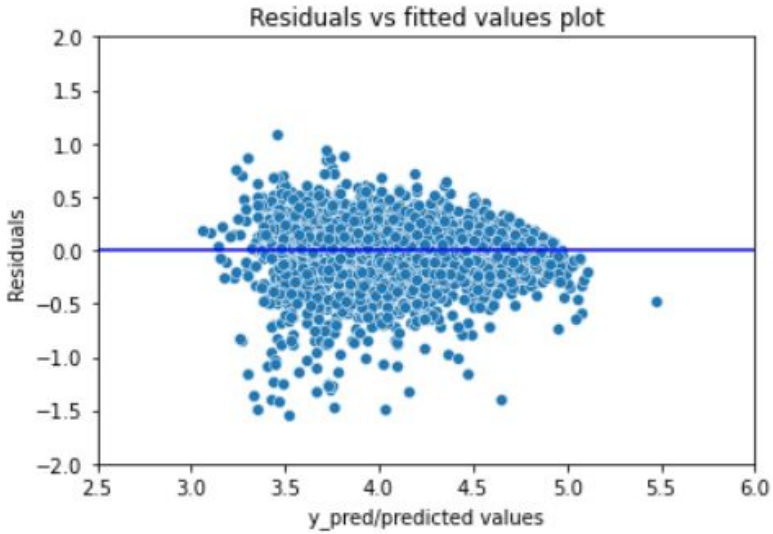
Model performance:

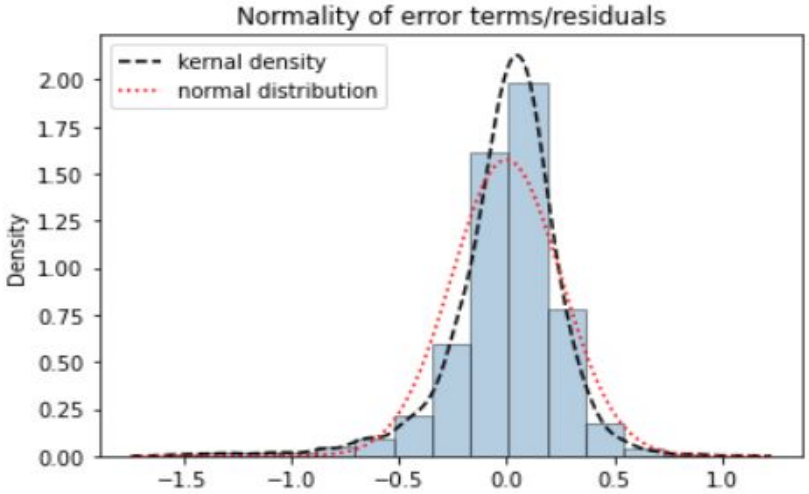
R-squared:	0.690
Adj. R-squared:	0.677
F-statistic:	52.51
Prob (F-statistic):	0.00
Log-Likelihood:	-183.29
AIC:	692.6
BIC:	1718.

Sample model summary:

	coef	std err	t	P>|t|	[0.025	0.975]
avg_review_funny	-0.0578	0.022	-2.655	0.008	-0.101	-0.015
avg_review_cool	0.0863	0.015	5.754	0.000	0.057	0.116
avg_stars	0.1588	0.023	6.823	0.000	0.113	0.204
avg_review_count	0.0002	7.73e-05	2.522	0.012	4.34e-05	0.000
avg_useful	9.901e-05	5.73e-05	1.728	0.084	-1.33e-05	0.000
avg_funny	0.0001	3.78e-05	2.814	0.005	3.22e-05	0.000
avg_cool	-0.0002	7.16e-05	-2.363	0.018	-0.000	-2.88e-05
avg_fans	-0.0017	0.000	-3.448	0.001	-0.003	-0.001
avg_compliment_hot	-0.0005	0.000	-1.266	0.206	-0.001	0.000
avg_compliment_more	-0.0023	0.004	-0.606	0.544	-0.010	0.005
avg_compliment_profile	-0.0008	0.003	-0.248	0.804	-0.007	0.006
avg_compliment_cute	0.0003	0.004	0.075	0.941	-0.008	0.008
avg_compliment_list	0.0060	0.006	1.084	0.278	-0.005	0.017
avg_compliment_note	-0.0001	0.000	-0.505	0.614	-0.001	0.000
avg_compliment_plain	-4.009e-05	0.000	-0.172	0.863	-0.000	0.000
avg_compliment_cool	-0.0003	0.000	-1.044	0.296	-0.001	0.000
avg_compliment_funny	-0.0003	0.000	-1.044	0.296	-0.001	0.000
avg_compliment_writer	-0.0010	0.001	-1.822	0.069	-0.002	7.68e-05
avg_compliment_photos	-0.0006	0.000	-1.892	0.059	-0.001	2.03e-05
yelp_years	-0.0303	0.005	-6.608	0.000	-0.039	-0.021
avg_total_compliment	0.0003	0.000	1.462	0.144	-0.000	0.001

4. Multiple Linear Regression - Assumption Check



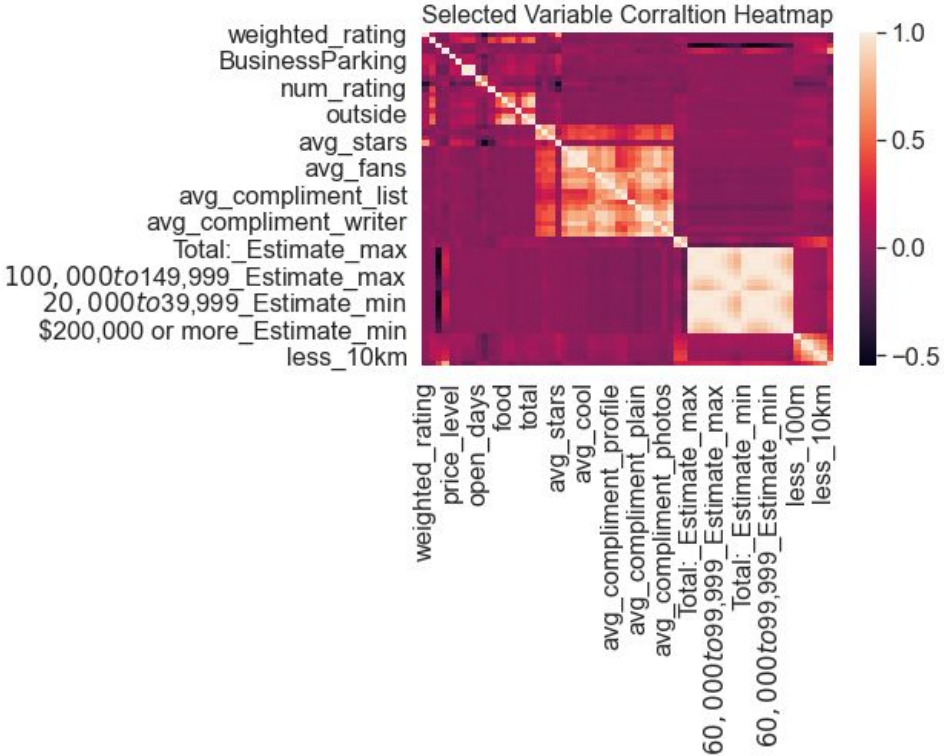


The points are basically symmetrically distributed around horizontal line in the plot, with a roughly constant variance. And the point does not follow a linear or quadratic shaped pattern.

Residuals are pretty much normally distributed.

Linearity	✓	Homoscedasticity	✓
Normality	✓	Independence	✓

4.Model - Correlation plot for some variable

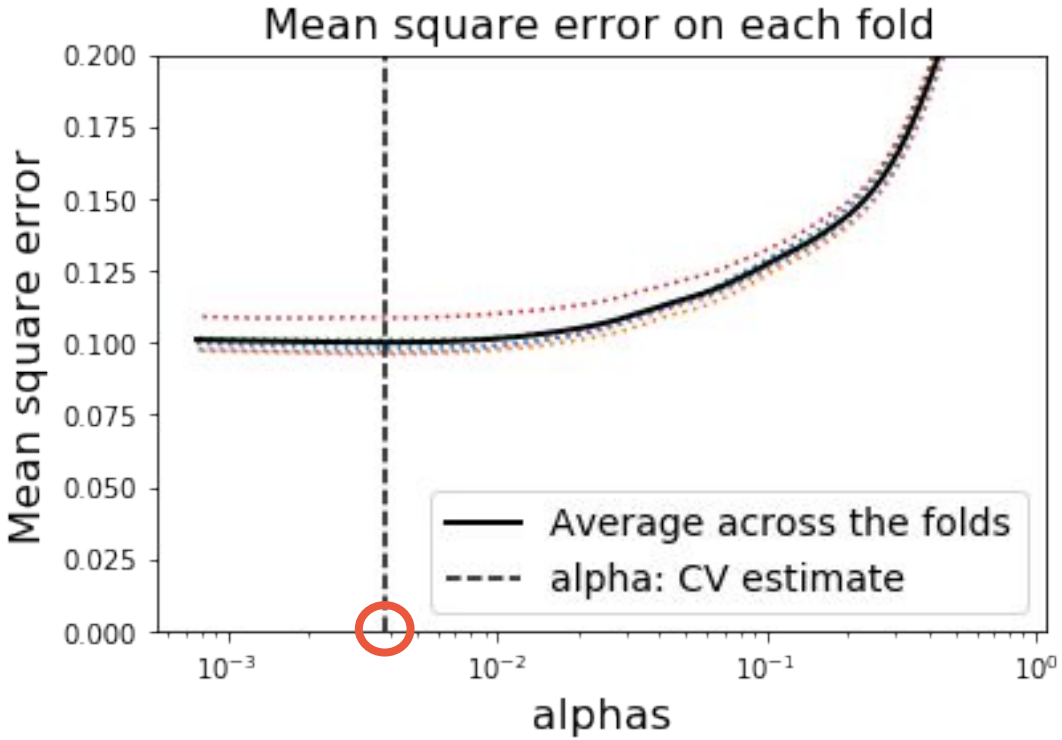


The left shows the correlation heatmap of some variables.

4. Model - Penalized Regression

- **Lasso Regression**
  - Lasso Regression shrinks the regression coefficients towards zero by penalizing the regression model with a penalty term called **L1-norm**, which is the sum of the absolute coefficients.
- **Ridge Regression**
  - Ridge regression shrinks the regression coefficients with minor contribution to the outcome to zero by penalizing the regression model with a penalty term called **L2-norm**.
- **Elastic Net regression**
  - Elastic Net produces a regression model that is penalized with both the **L1-norm** and **L2-norm**. The consequence of this is to effectively shrink coefficients (like in ridge regression) and to set some coefficients to zero (as in LASSO).

4. Model - Penalized Regression Hyperparameter Tuning



Use K-Fold to find optimized alpha:

```
Config: {'alpha': 0.0030150753768844224}
```

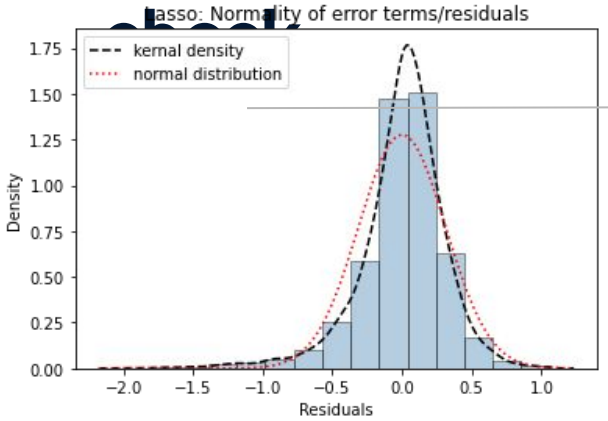
These two results match.

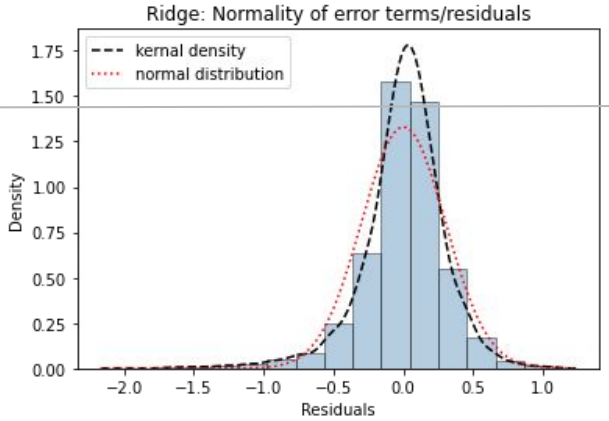
Note: This method is used for all these three penalized regressions to find the best model.

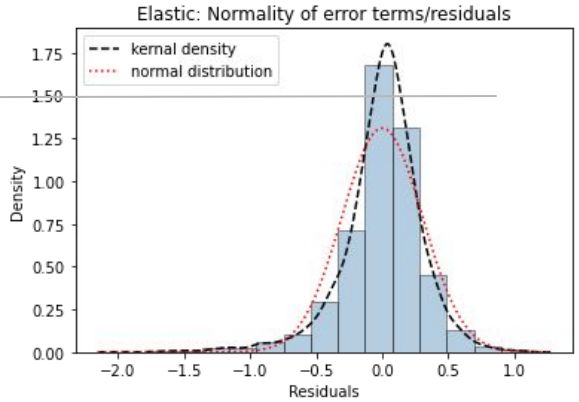
4. Model - Penalized Regression - result

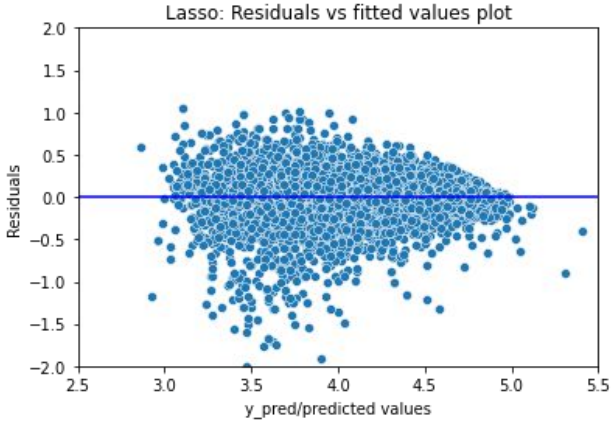
LASSO Regression	Elastic Net Regression	Ridge Regression
R^2: 0.6747 Test MAE: 0.236	R^2: 0.6916 Test MAE: 0.243	R^2: 0.7005 Test MAE: 0.262

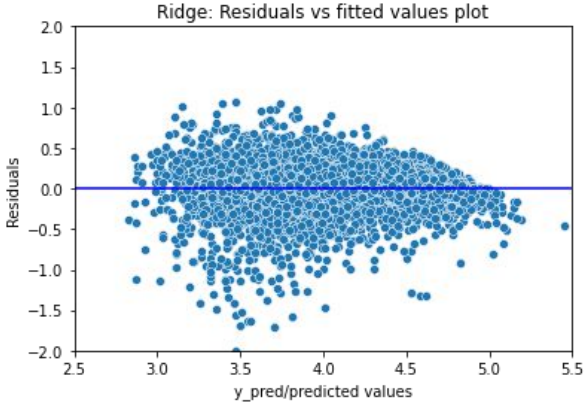
4.Model - Penalized Regression - assumption

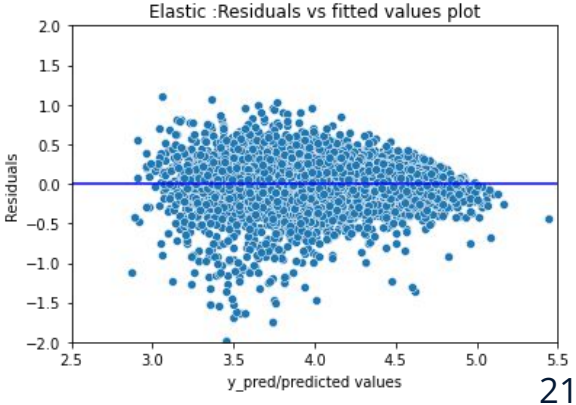












4. Model - Autoregressive - description

Autoregressive Model

- Model factors with autocorrelation (correlation between observations at adjacent time), and current Y response variable depend on previous Y variable.

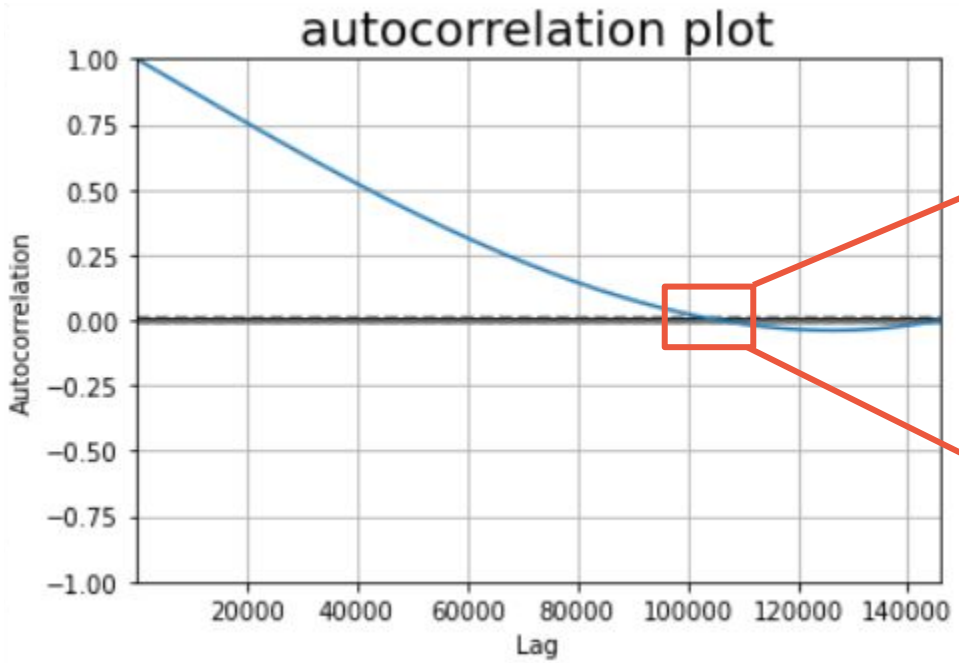
- $$Yt = \beta_0 + \beta_1 * Y_{t-1} + \sum \alpha_i * X_{t,i} + error_t$$

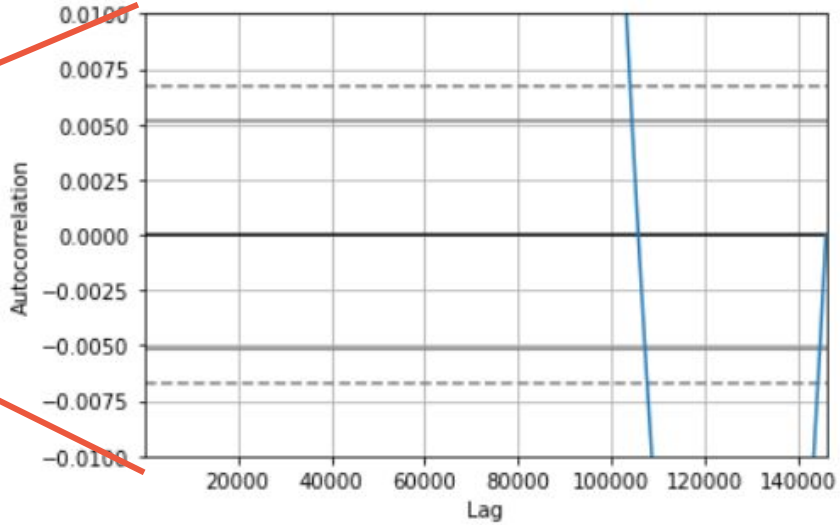
4. Model - Autoregressive - Result

OLS Regression Results

Dep. Variable:	y	R-squared:	0.771
Model:	OLS	Adj. R-squared:	0.771
Method:	Least Squares	F-statistic:	1620.
Date:	Sun, 22 May 2022	Prob (F-statistic):	0.00
Time:	00:02:45	Log-Likelihood:	-48581.
No. Observations:	113189	AIC:	9.763e+04
Df Residuals:	112953	BIC:	9.991e+04
Df Model:	235		
Covariance Type:	nonrobust		

4. Model - AutoRegressive - Assumption check





Enlarged

4. Model - Result Comparison

Multiple Linear Regression	LASSO Regression	Elastic Net Regression	Ridge Regression	AutoRegressive
R^2: 0.67	R^2: 0.6747 Test MAE: 0.236	R^2: 0.6916 Test MAE: 0.243	R^2: 0.7005 Test MAE: 0.262	R^2: 0.771 Test MAE: 0.256

Getting better!

Note: All R^2 are adjusted R^2.

5. Project conclusion

Significant factors: (Factors with p value <0.05)
[This table is only part of the features]

Feature	Coefficient	Feature meaning
User_avg_stars	1.2106	The average rating reviewers gave to any shop.
Pre_rating	0.2304	Rating from last month.
avg_review_cool	0.2232	The average cool of reviews for certain shop.
v_17,39…	0.1576; 0.1486; ...	Important review vectors (hard to interpret).
class_4,8…	0.1597; 0.1813; …	Count of class_4(airplane), class_8(boat) in review pictures.
RestaurantsTakeOut	0.0321	Whether the shop can take out or not.
BusinessParking	0.0288	Whether the shop provide business parking lot.
BikeParking	0.0274	Whether the shop provide Bike parking lot.
Outdoorseating	0.0194	Whether the shop provide outdoor seating.

6. Milestones

Winter Week 8-10
Exploring the resources and start collecting data

Spring Week 1-2
Designing data schema and finish data collecting

Week 3-4
Data cleaning and data analysis

Week 5-6
Selecting model and start constructing models

Week 7-8
Improving our model and Evaluate the model

Week 9-10
Finish the final report and presentation

THANKS!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.