SENTIMENT ANALYSIS AND RATING

IN YELP RESTAURANTS USER REVIEWS

BY:

Steven Too Heng Kwee

ID: 304449

Business Problem

Data Preparation

Model Development

Model Evaluation

Solution Deployment

ABSTRACT

We perform sentiment analysis based on Yelp user reviews. Reviews are extracted, segmented and cleaned. Text vectorization is performed and different algorithms are used, namely perceptron learning algorithm, Naive Bayes, Random Forest Classifier, Decision Tree and SVM to predict an actual rating from 1 to 5.

Business Problem → Data Preparation → Model Development → Model Evaluation → Solution Deployment

INTRODUCTION

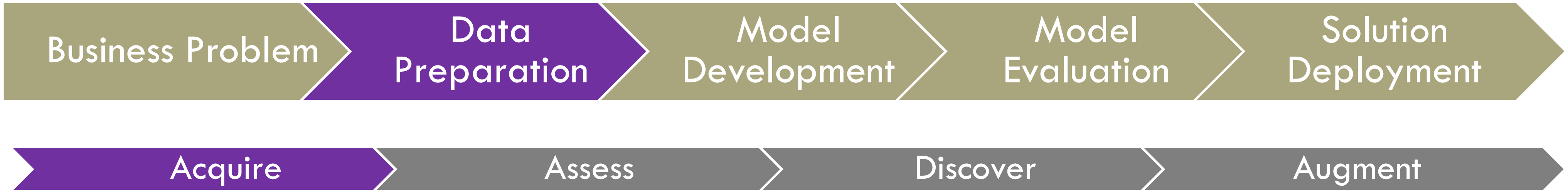
- Yelp, Inc. is a company that enables users to rate and review all kinds of businesses. In the area of restaurants, such reviews and ratings essentially function as crowd-sourced food (and drink) criticism. Yelp is the largest such user-supplied review web service, and as such has very large amounts of review data. The two main parts of a review are the text of the actual review and a star rating from one to five.

- Empirical data research demonstrated that an average one-star increase led to 59% increase in revenue of independent restaurants (Lucas, 2011)

- However with so many reviews per restaurant, users tend to rely heavily on the ratings distribution which is very subjective and biased. The same expressed opinion can be rated differently by users.

Research Question:

> ***"Analyzing the reviews and star rating set by reviewers. Can we provide an unbiased star rating ? "***


Business Problem

Data Preparation

Model Development

Model Evaluation

Solution Deployment

Acquire

Assess

Discover

Augment

The Dataset

6,685,900 reviews

192,609 businesses

200,000 pictures

10 metropolitan areas

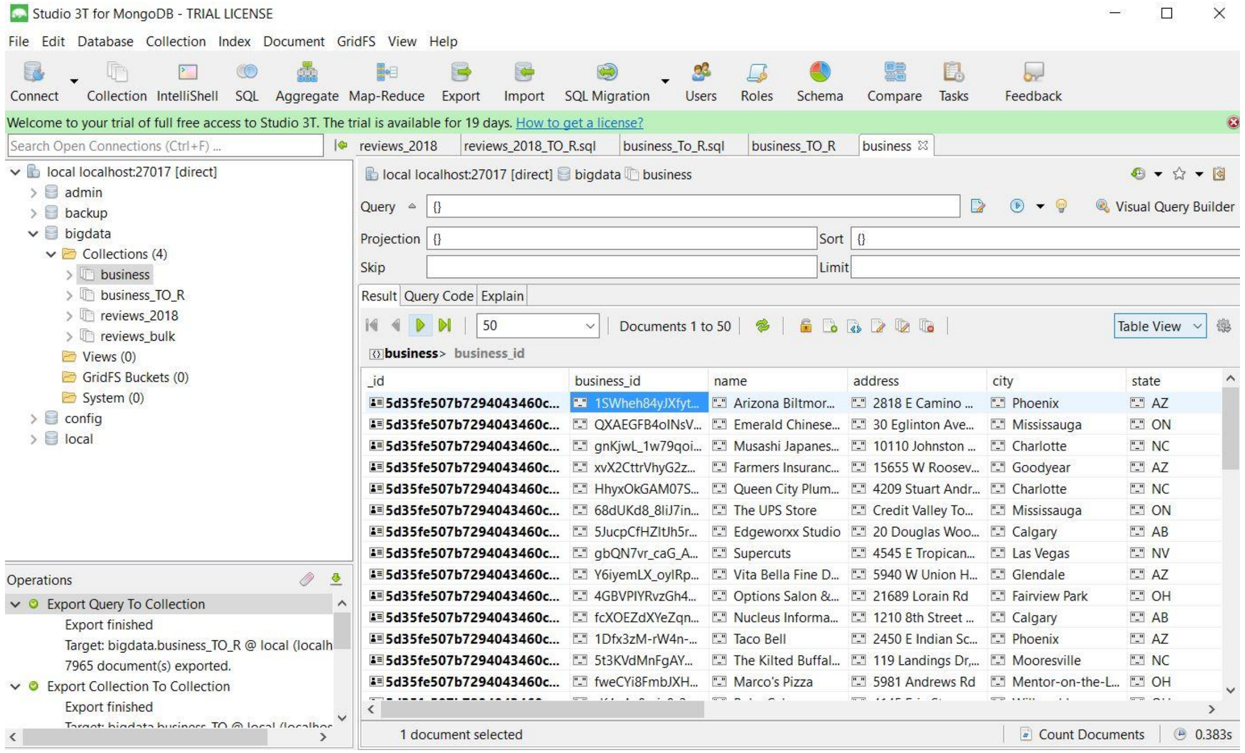
1,223,094 tips by 1,637,138 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 192,609 businesses

- The data comes from Yelp Dataset Challenge available at https://www.yelp.ca/dataset
- Files extracted in this project:

Business.json: 131 Mb. Contains business data including location data, attributes, categories and average star rating

Review.json: 4.97 Gb. Contains full review text data including the user_id that wrote the review and the business_id the review is written for.

- Review.json file cannot be loaded directly into Python dataframe
- MongoDB community edition used as repository and segmention tool
- PyMongo used as connection client for data retrieval
- Data segmented to 2018 reviews for restaurants in Toronto



Welcome to your trial of full free access to Studio 3T. The trial is available for 19 days. How to get a license?

- local localhost:27017 [direct]
  - admin
  - backup
  - bigdata
    - Collections (4)
      - business
      - business_TO_R
      - reviews_2018
      - reviews_bulk
    - Views (0)
    - GridFS Buckets (0)
    - System (0)
  - config
  - local

local localhost:27017 [direct] bigdata business

Query {}

Projection {} Sort {}

Skip Limit

50 Documents 1 to 50

business> business_id

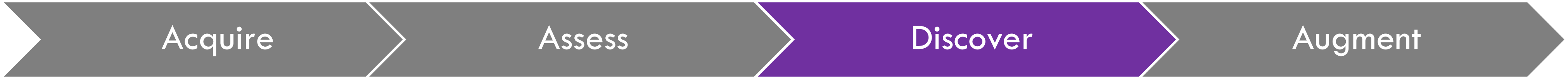
_id	business_id	name	address	city	state
5d35fe507b7294043460c...	1SWheh84yJXfyt...	Arizona Biltmor...	2818 E Camino ...	Phoenix	AZ
5d35fe507b7294043460c...	QXAEGFB4oINsV...	Emerald Chinese...	30 Eglinton Ave...	Mississauga	ON
5d35fe507b7294043460c...	gnKjwL_1w79qoi...	Musashi Japanes...	10110 Johnston ...	Charlotte	NC
5d35fe507b7294043460c...	xvX2CttrVhyG2z...	Farmers Insuranc...	15655 W Roosev...	Goodyear	AZ
5d35fe507b7294043460c...	HhyxOkGAM07S...	Queen City Plum...	4209 Stuart Andr...	Charlotte	NC
5d35fe507b7294043460c...	68dUKd8_8liJ7in...	The UPS Store	Credit Valley To...	Mississauga	ON
5d35fe507b7294043460c...	5JucpCfHZItJh5r...	Edgeworxx Studio	20 Douglas Woo...	Calgary	AB
5d35fe507b7294043460c...	gbQN7vr_caG_A...	Supercuts	4545 E Tropican...	Las Vegas	NV
5d35fe507b7294043460c...	Y6iyemLX_oylRp...	Vita Bella Fine D...	5940 W Union H...	Glendale	AZ
5d35fe507b7294043460c...	4GBVPIYRvzGh4...	Options Salon &...	21689 Lorain Rd	Fairview Park	OH
5d35fe507b7294043460c...	fcXOEZdXYeZqn...	Nucleus Informa...	1210 8th Street ...	Calgary	AB
5d35fe507b7294043460c...	1Dfx3zM-rW4n-...	Taco Bell	2450 E Indian Sc...	Phoenix	AZ
5d35fe507b7294043460c...	5t3KVdMnFgAY...	The Kilted Buffal...	119 Landings Dr,...	Mooresville	NC
5d35fe507b7294043460c...	fweCYi8FmbJXH...	Marco's Pizza	5981 Andrews Rd	Mentor-on-the-L...	OH

1 document selected

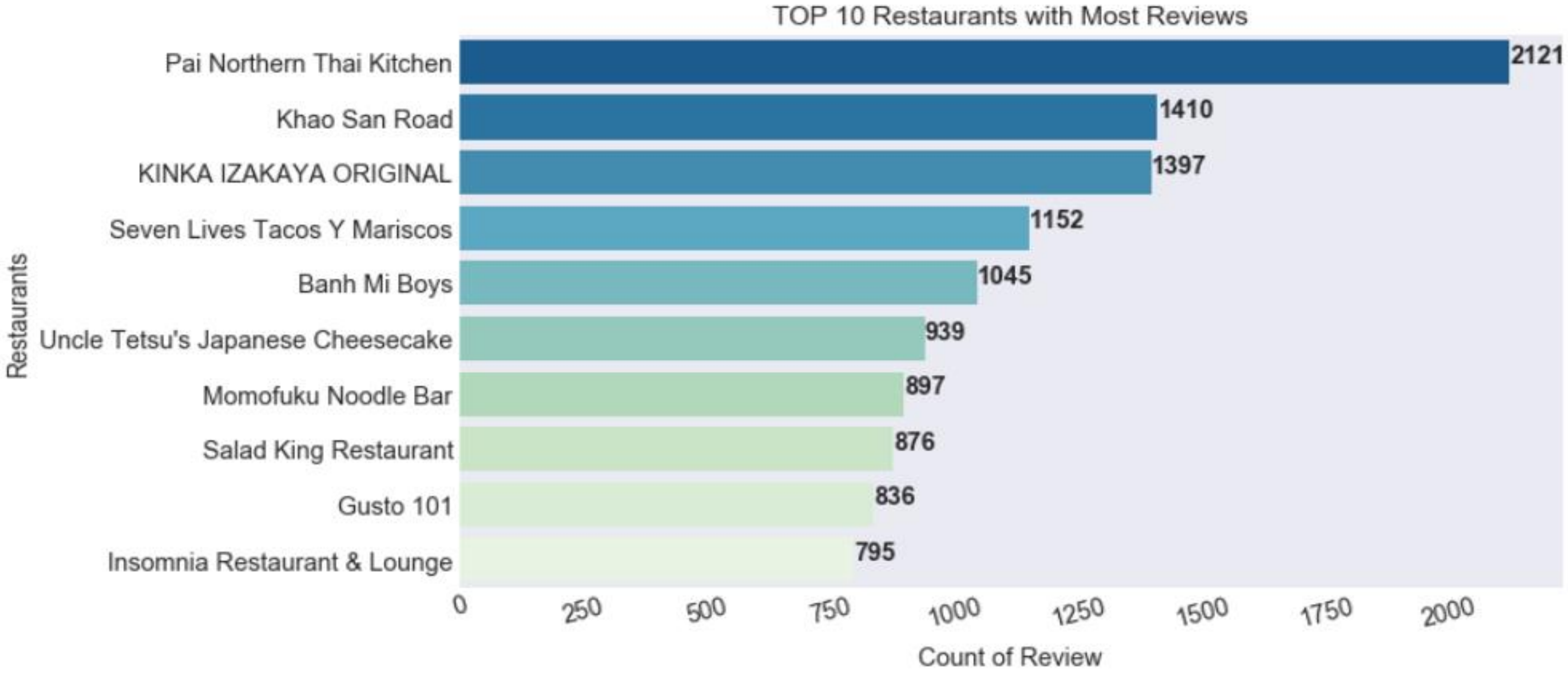
0.383s

Operations

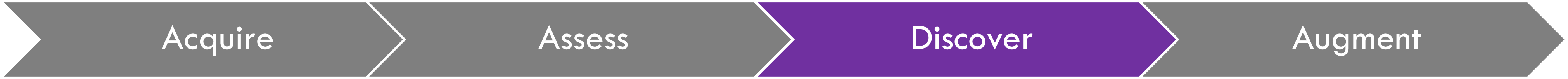
- Export Query To Collection
  - Export finished
  - Target: bigdata.business_TO_R @ local (localh
  - 7965 document(s) exported.
- Export Collection To Collection
  - Export finished



- Number of reviews extracted: 57,047
- Number of businesses: 7,965



- Top 10 restaurants with most reviews. Grouped by name.

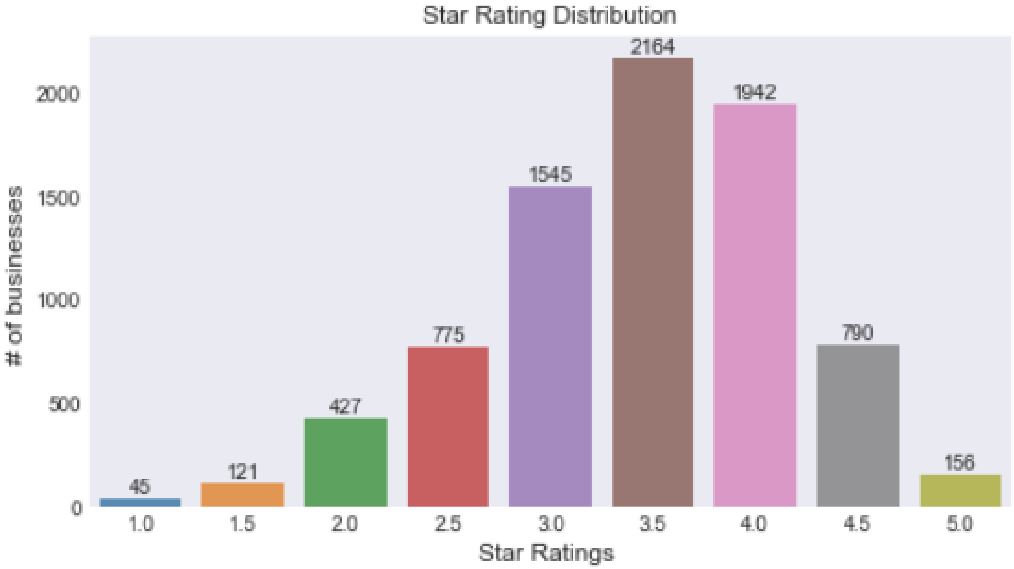


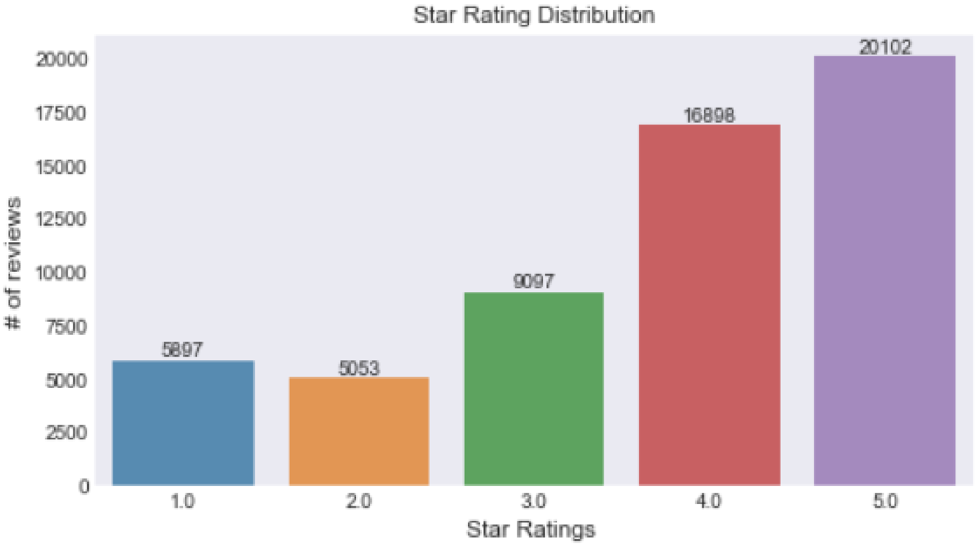
Acquire

Assess

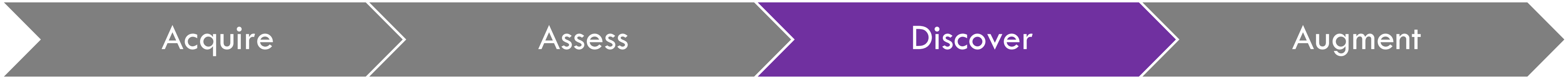
Discover

Augment





The business ratings set by Yelp follows a normal distribution. However, the users review ratings are mostly 4.0, 5.0 and notice the higher proportion of 1 star. Based on the reviews alone we can see the tendency by reviewers to over and under rate.

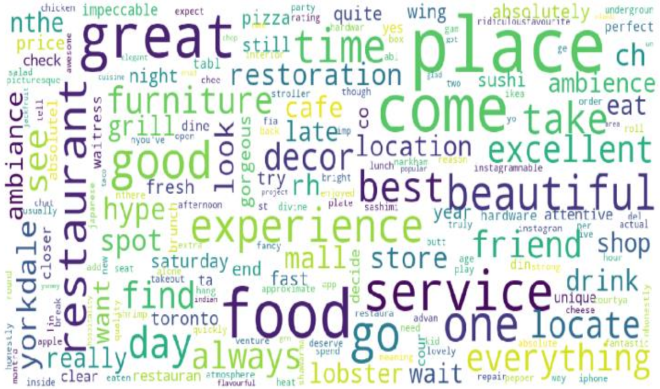


Acquire

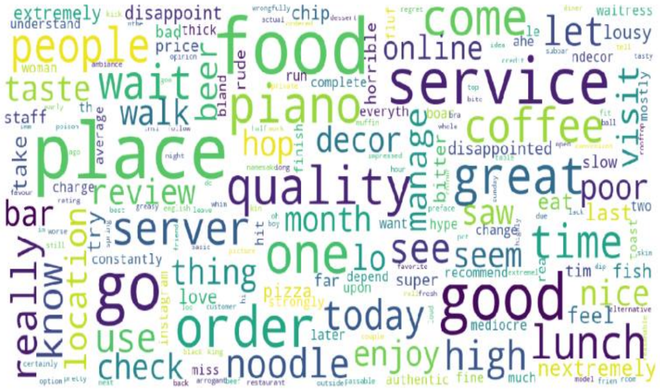
Assess

Discover

Augment



Positive user reviews: place, food, service, time, décor, furniture



Negative user reviews: food, service, place, order, wait, server, lunch

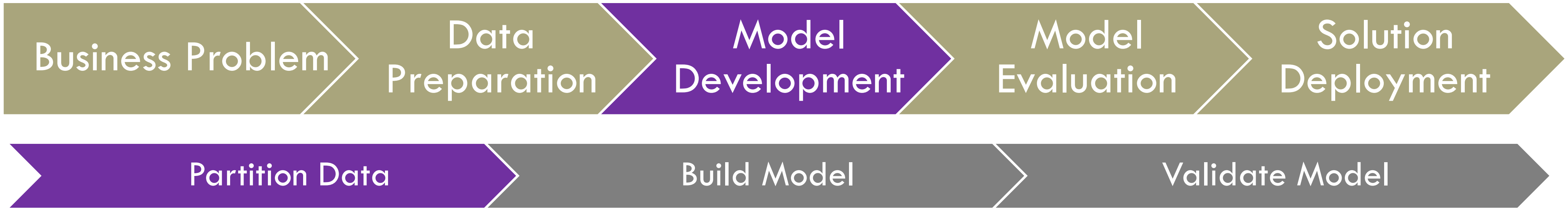
Observations: Most words are common on both positive and negative reviews. This denotes the need for context to determine the actual sentiment

Cleaning the text reviews

Create clean_text function for EDA and text_process function for modelling

- lower the text
- tokenize the text (split the text into words) and remove the punctuation
- remove useless words that contain numbers
- remove useless stop words like ‘the’, ‘a’ ,’this’ etc.
- Part-Of-Speech (POS) tagging: assign a tag to every word to define if it corresponds to a noun, a verb etc. using the WordNet lexical database
- lemmatize the text: transform every word into their root form (e.g. restaurants -> restaurant, ate -> eat)

Sample: All stars go to the decor and atmosphere of th...
Cleaned text: star go decor atmosphere cafe make feel like p...



Business Problem

Data Preparation

Model Development

Model Evaluation

Solution Deployment

Partition Data

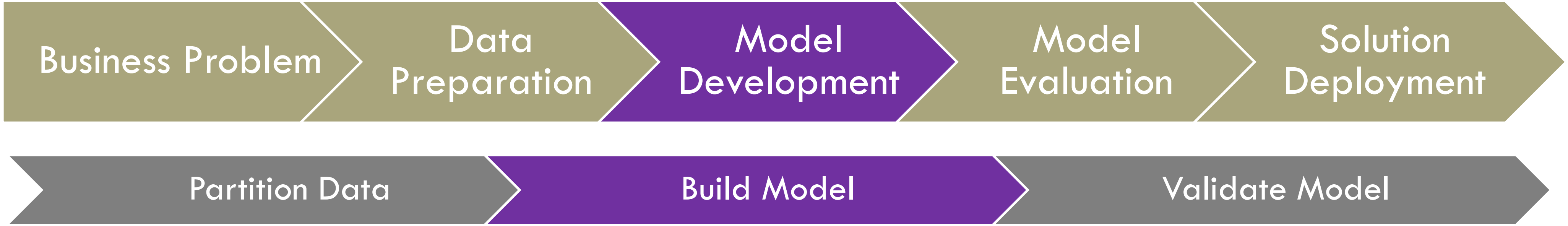
Build Model

Validate Model

- The dataset was categorized into 1, 3 and 5 stars thereby reducing the reviews to 35,096. This should reduce the processing time and increase accuracy.

- Attempt was done with all 5 categories but proven less accurate. See chart below.

- The train and test dataset was split into 80/20 sets.



Business Problem

Data Preparation

Model Development

Model Evaluation

Solution Deployment

Partition Data

Build Model

Validate Model

- Stopwords and punctuation removed from the classified dataset

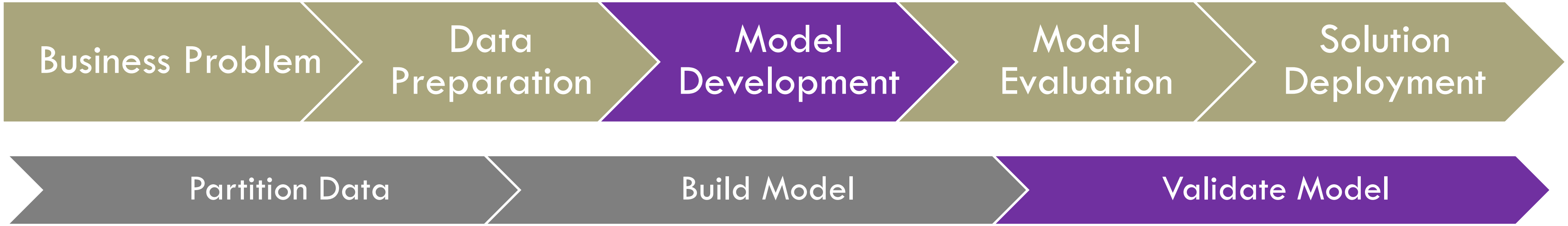
```
def text_process(text):
    nopunc = [char for char in text if char not in string.punctuation]
    nopunc = ''.join(nopunc)
    return [word for word in nopunc.split() if word.lower() not in stopwords.words('english')]
```

- Using CountVectorizer function to Tokenize and perform word count

```
CountVectorizer(analyzer=text_process).fit(x)
```

- *Splitting the dataset into Training Set and Testiong Set on a 80/20 partition*

```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2,random_state=101)
```



Business Problem

Data Preparation

Model Development

Model Evaluation

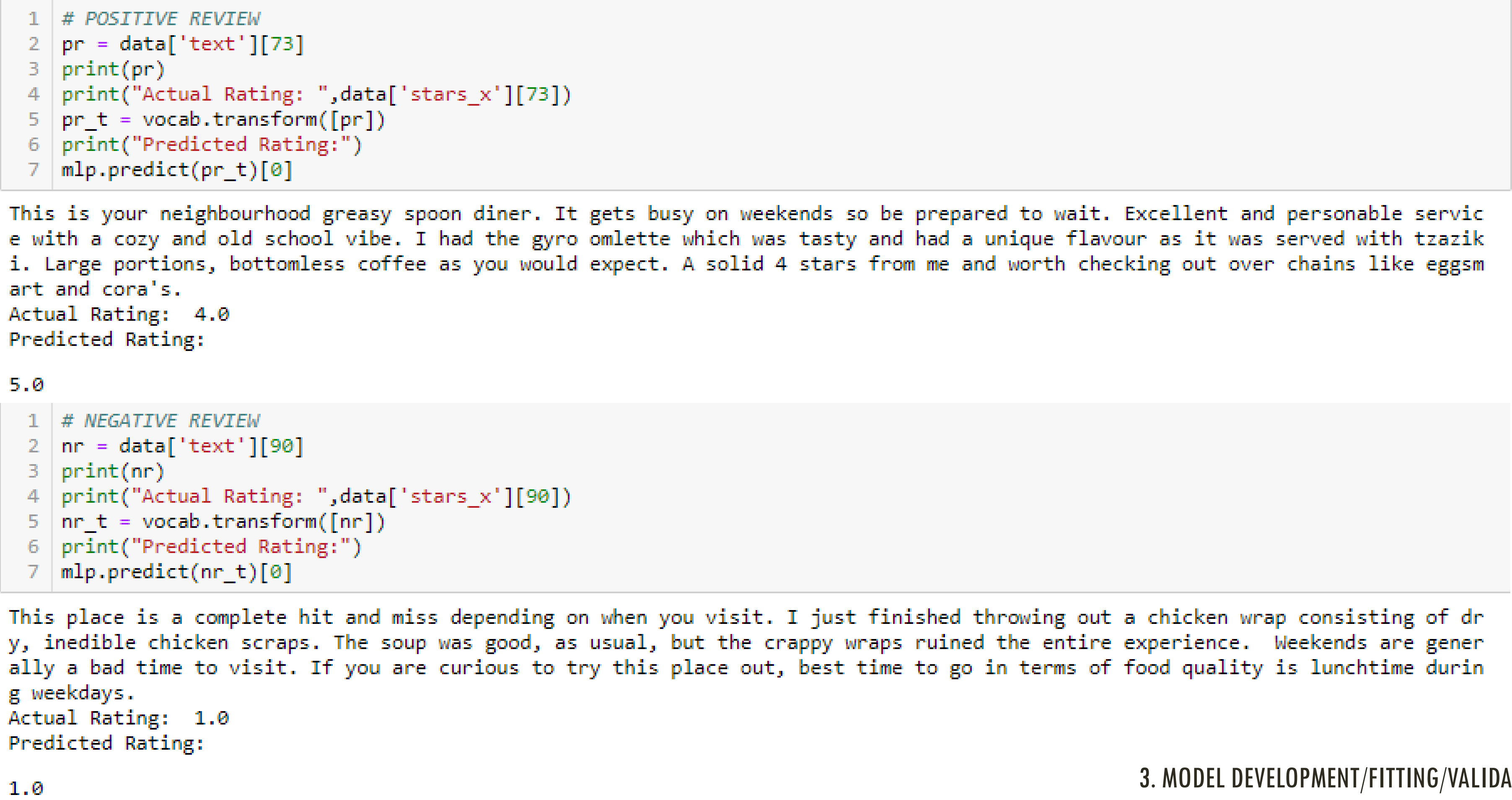
Solution Deployment

Partition Data

Build Model

Validate Model

Sample test with Multiplayer Perceptron Classifier



```
# POSITIVE REVIEW
pr = data['text'][73]
print(pr)
print("Actual Rating: ",data['stars_x'][73])
pr_t = vocab.transform([pr])
print("Predicted Rating:")
mlp.predict(pr_t)[0]
```

```
This is your neighbourhood greasy spoon diner. It gets busy on weekends so be prepared to wait. Excellent and personable servic
e with a cozy and old school vibe. I had the gyro omlette which was tasty and had a unique flavour as it was served with tzazik
i. Large portions, bottomless coffee as you would expect. A solid 4 stars from me and worth checking out over chains like eggsm
art and cora's.
Actual Rating:  4.0
Predicted Rating:

5.0
```

```
# NEGATIVE REVIEW
nr = data['text'][90]
print(nr)
print("Actual Rating: ",data['stars_x'][90])
nr_t = vocab.transform([nr])
print("Predicted Rating:")
mlp.predict(nr_t)[0]
```

```
This place is a complete hit and miss depending on when you visit. I just finished throwing out a chicken wrap consisting of dr
y, inedible chicken scraps. The soup was good, as usual, but the crappy wraps ruined the entire experience.  Weekends are gener
ally a bad time to visit. If you are curious to try this place out, best time to go in terms of food quality is lunchtime durin
g weekdays.
Actual Rating:  1.0
Predicted Rating:

1.0
```

Business Problem

Data Preparation

Model Development

Model Evaluation

Solution Deployment

Model Evaluation

Models	Precision % - 1,3,5 stars	Precision % - all stars
Multilayer Perceptron	85.1	54.72
Multinomial Naive Bayes	85.01	56.91
Random Forest Classifier	73.7	47.25
Decision Tree	70.07	43.3
Support Vector Machine	57.48	35.24

- Multilayer Perceptron Classifier produced the best accuracy score on a 1,3,5 stars classifier. Let us use it to predict a random positive review and a random negative review!

Business Problem

Data Preparation

Model Development

Model Evaluation

Solution Deployment

Discussion and Conclusion

- This above model has many applications not limited to reviews. I can be uses to any text that required some sort of scoring or detect unfair or erroneous ratings

- We are able to accurately the stars rating according to the reviews. However when we look at the negative sample, it appears that it should have received a higher rating.

Improvement Prospects

- Review the vectorization process. Processing time too long.

- Review the whole modelling concept. Star rating prediction trained current ratings. Research to undertake for rating based on just reviews text and sentiment alone. Also to consider normalizing the data and adding some weight element to the words.