# Starbucks coffee customer sentiment analysis

1/6/2025 Hyunyong Lee

# Agenda

1. Problem Statement & Objective
Value proposition

2. Data:
Preprocessing, EDA & Feature engineering

3. Prediction model

4. Recommendation

# Problem Statement & Objective

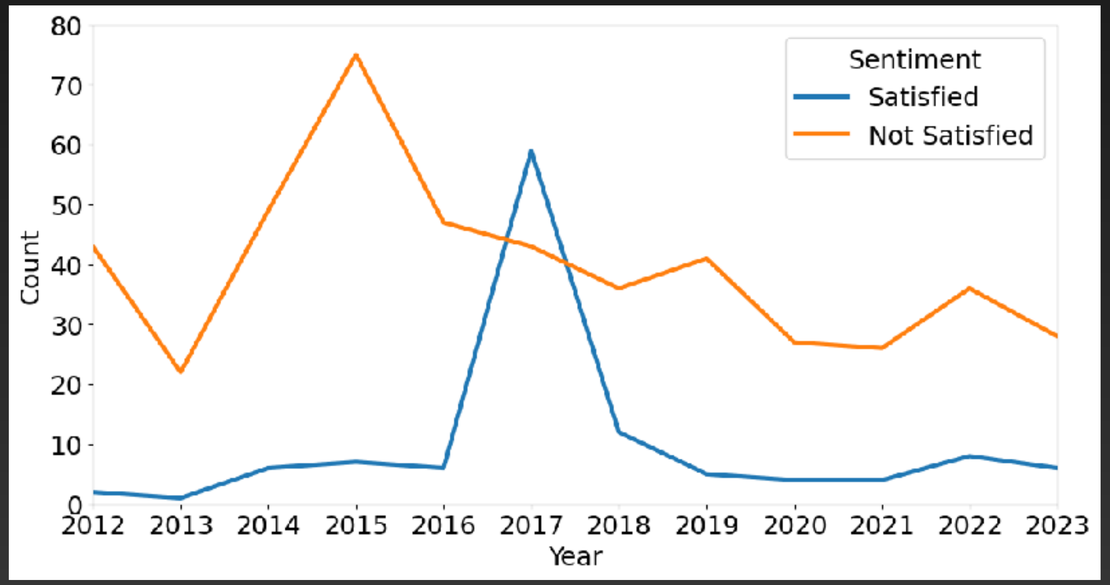
**Customers are not satisfied with our services!**

- Traditional approach: Human sorts out comments and summarize (Cannot done by real-time)

**Objective**

1) Identify opportunity areas
2) To develop a predictive model on sentiment analysis of customer review

**Stakeholder**: Operation managers



# Value proposition

- Provide priority to focus; region/store and issue to check
- Develop a real-time monitoring system of customer feedback

# Data

Data extraction
Preprocessing
EDA
Feature engineering
- Text/NLP based
- Vectorization & Topic modeling
- Word embedding with PCA

# Datasets

- Downloaded from Kaggle

https://www.kaggle.com/datasets/harshalhonde/starbucks-reviews-dataset/data

Review comment, Location (City & State), Name, Rating, Date, Images
*Extract only US data (majority) for last 10 years (2012-2023)

```
RangeIndex: 850 entries, 0 to 849
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   name         850 non-null    object
 1   location     850 non-null    object
 2   Date         850 non-null    object
 3   Rating       705 non-null    float64
 4   Review       850 non-null    object
 5   Image_Links  850 non-null    object
```

# Pre-processing & Feature Engineering

- Correct format of State code using web scraping of Wikipedia
  i.e. New York & NY > NY

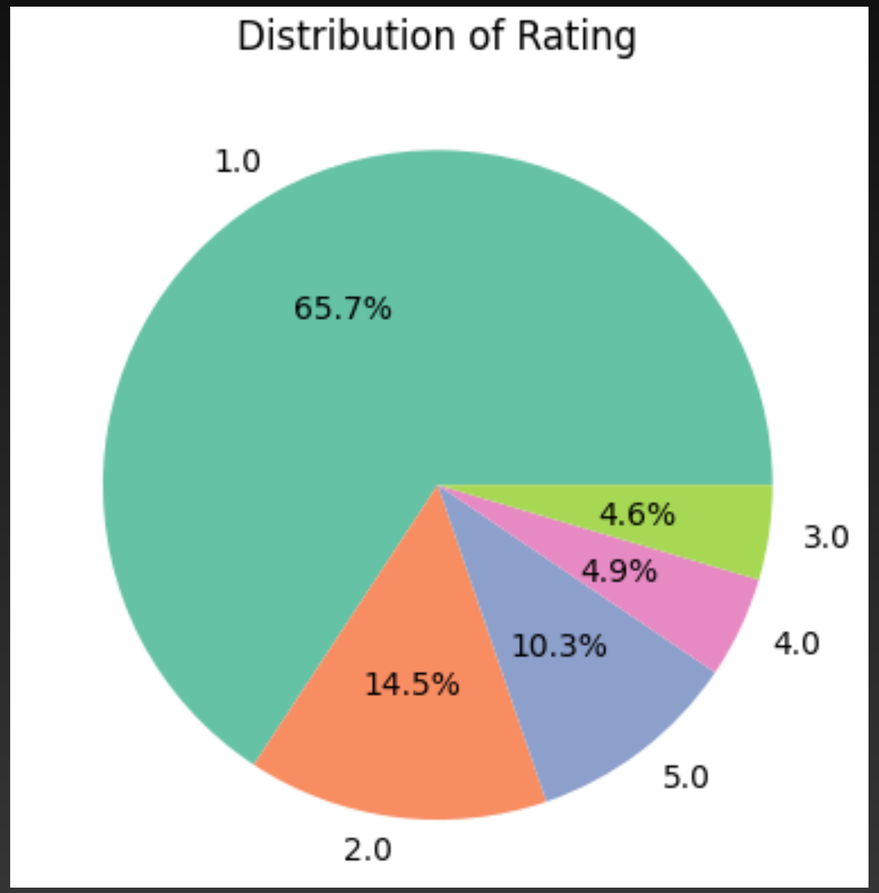
```
['TX' 'FL' 'PA' 'WA' 'OR' 'NC' 'MD' 'OTHER' 'CA' 'OH' 'HI' 'NJ' 'GA' 'DC'
 'AZ' 'MA' 'VA' 'NV' 'TN' 'IA' 'WI' 'NH' 'AR' 'MN' 'IN' 'MO' 'IL' 'MI'
 'MS' 'CO' 'OK' 'UT' 'KY' 'ME' 'KS' 'ON' 'BC' 'NY' 'NE' 'AK' 'AB' 'ID'
 None 'LA' 'UK' 'SC' 'MB' 'SK' 'CALIFORNIA' 'WYOMING' 'VIRGINIA'
 'SASKATCHEWAN' 'SOUTH CAROLINA' 'NL' 'NM' 'MINNESOTA' 'FLORIDA' 'ALBERTA'
 'ALA' 'WV' 'MAINE' 'NEW YORK' 'NS' 'ND' 'COLORADO' 'RI' 'MICHIGAN' 'WY'
 'AL' 'QC' 'MT' 'CT' 'NO OTHER LINE NEEDED']
```

- NLP with nltk
  > Applied lower case, lemmatization, removing stop words, tokenization

- Vectorization with TF-IDF word-level
  > Topic models

- Word embedding by word2vec
  > Dimensionality reduction by PCA
  ~~> Unsupervised learning: Clustering (NO)~~

# Exploratory Data Analysis



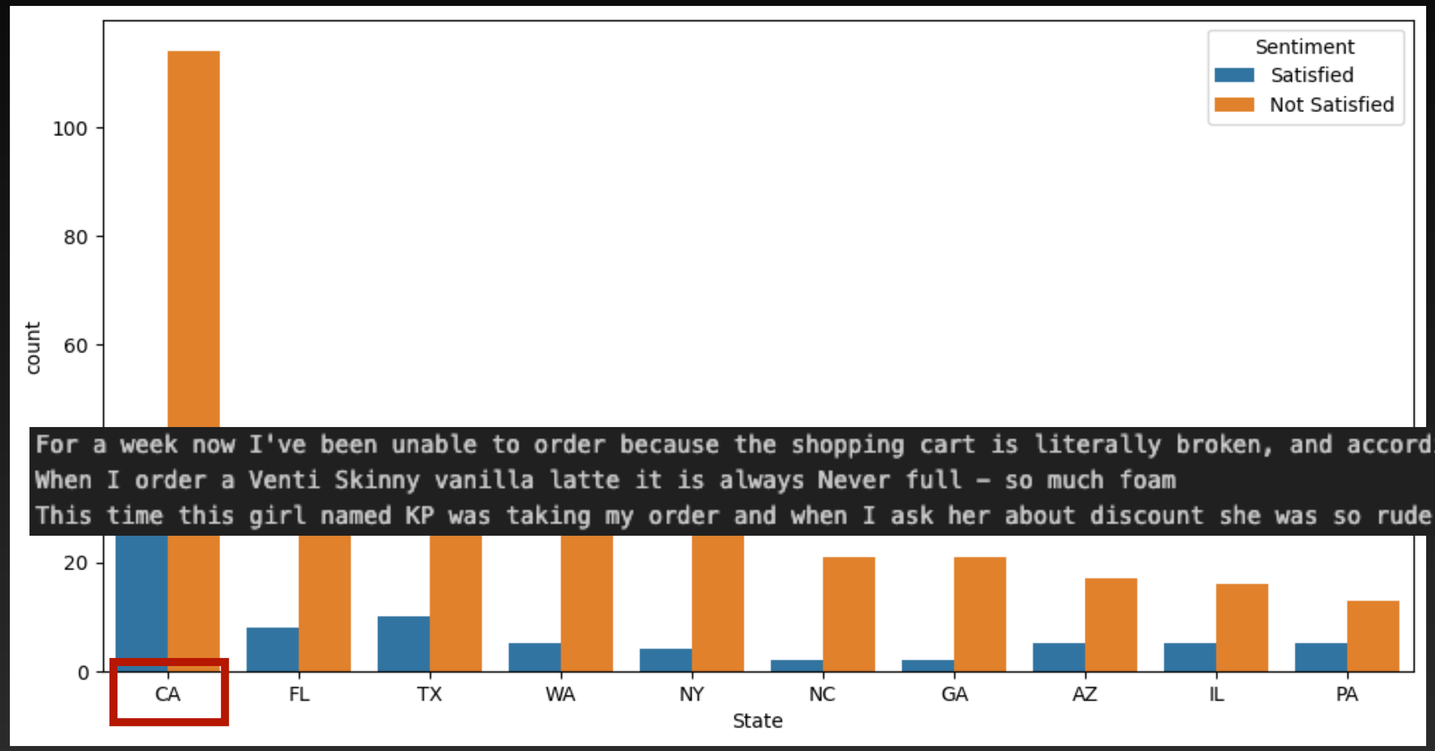

Rating converted to sentiment

1 & 2: Not satisfied (as 0)

3-5: Satisfied (as 1)

**<u>** Imbalanced label distribution (Over sampling for model training)</u>**

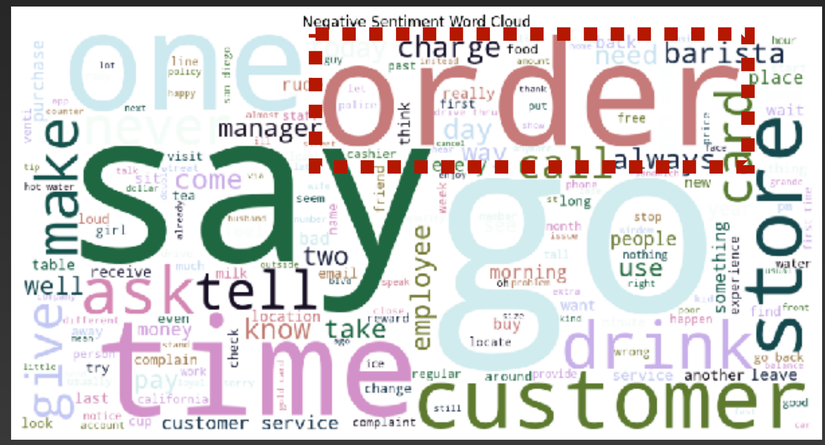
# Exploratory Data Analysis




For a week now I've been unable to order because the shopping cart is literally broken, and according to each person I've dealt with, it has something to do with PayPal
When I order a Venti Skinny vanilla latte it is always Never full - so much foam
This time this girl named KP was taking my order and when I ask her about discount she was so rude about it and she is like "There is no button on screen that I can give

## Satisfied


## Not satisfied




***Further investigation required**

Area: CA (California)

Issues: Order-related

# Prediction model

- Classification models:

Logistic regression, SVC, KNN, RF, LGBM, XGB, DNN

- Train-test data split

Hold-out (due to small data size)

- Hyper parameter tuning of ML

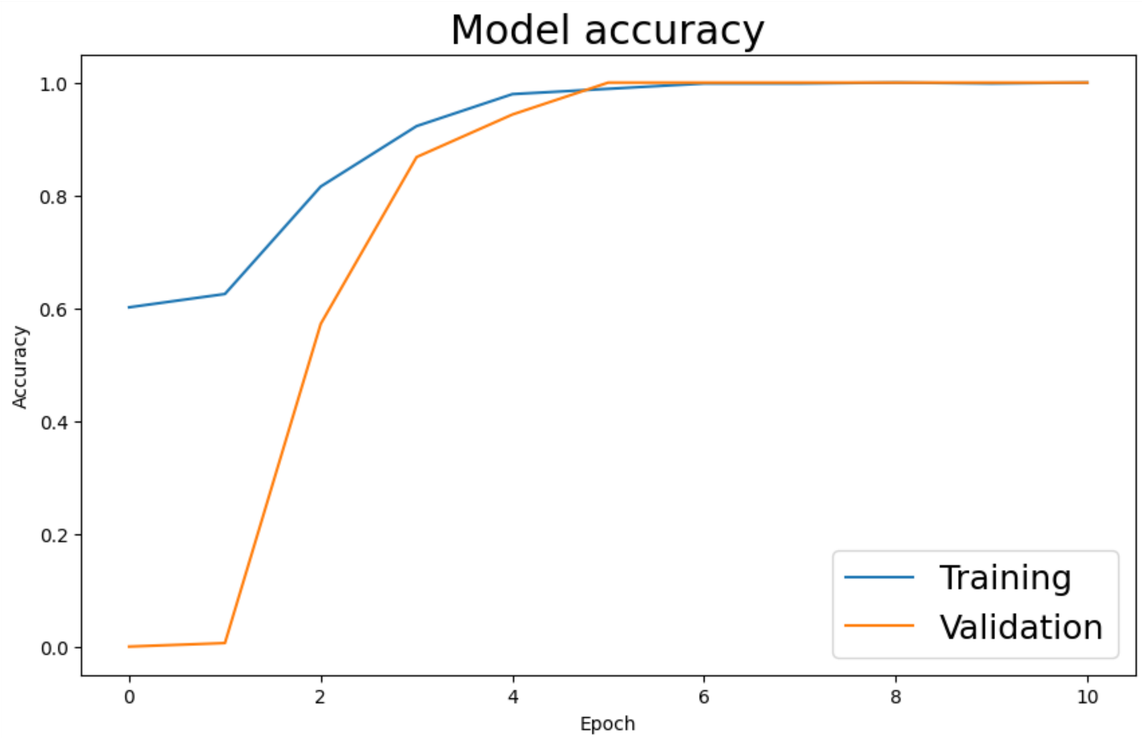
GridSearchCV with train data

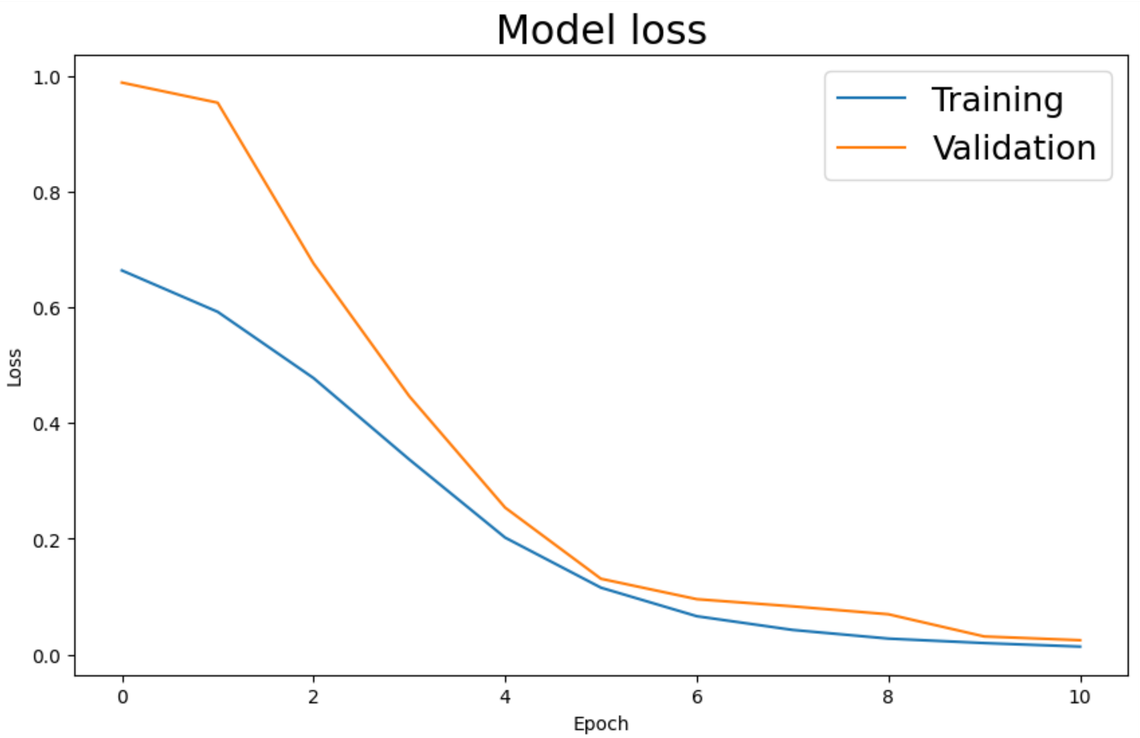
- Model selection: Acc with train/test data

# Model Selection/Evaluation

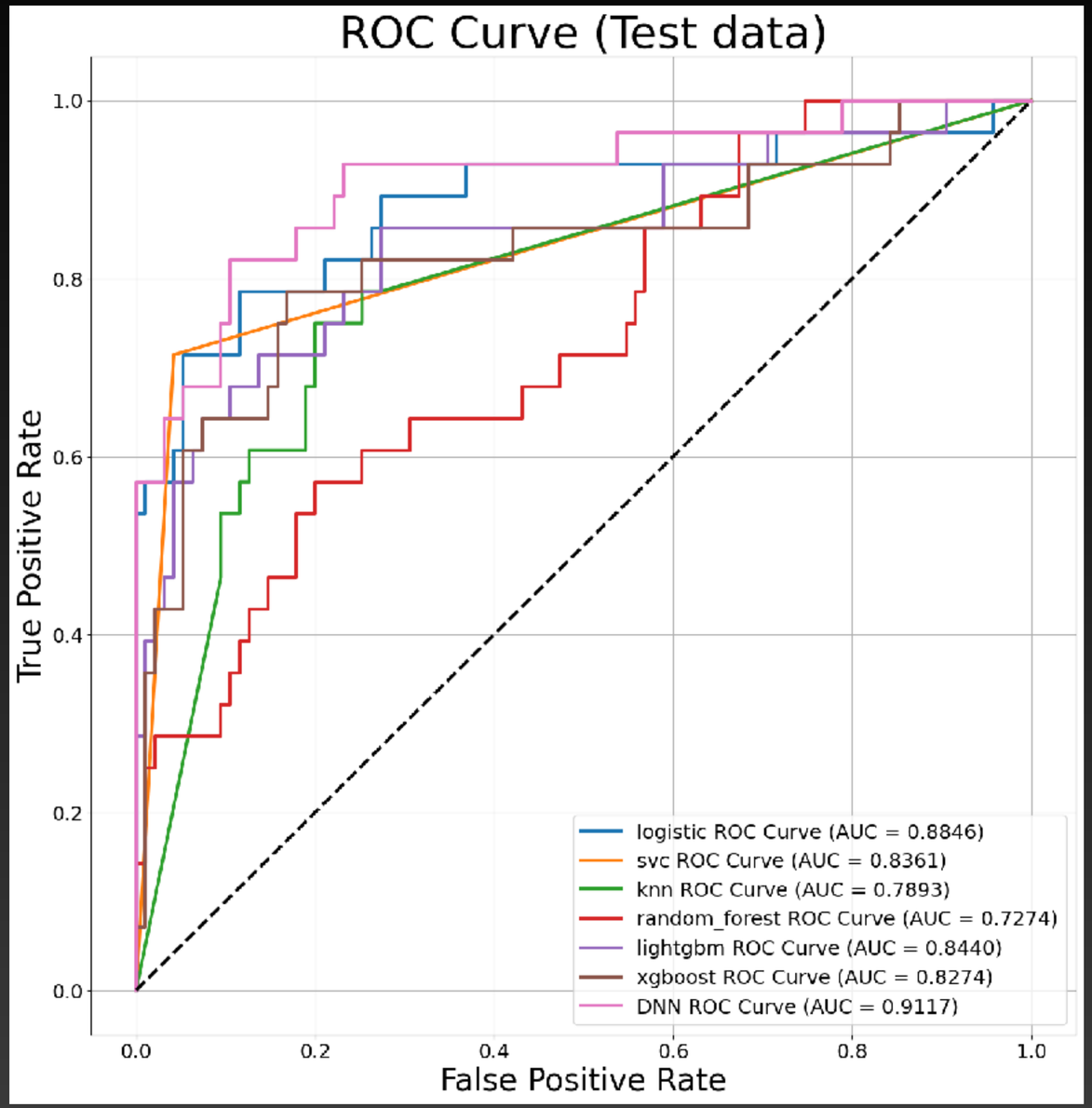
| Model | Train (Accuracy) | Test (Accuracy) | Test (AUC) | Execution time (Train, msec) | Execution time (Test, msec) |
|---|---|---|---|---|---|
| Logistic Regression | **98.5** | **84.6** | **0.885** | **149** | **5** |
| SVC | **98.6** | **90.2** | **0.836** | **2423** | **93** |
| KNN | **86.8** | **77.2** | **0.789** | **2383** | **578** |
| Random Forest | **95.5** | **77.2** | **0.727** | **1083** | **2** |
| LightGBM | **97.0** | **86.2** | **0.844** | **1504** | **2** |
| XGBoost | **97.2** | **68.2** | **0.827** | **5160** | **2** |
| DNN | **99.9** | **88.6** | **0.912** | **1300** | **43** |

# Model Selection/Evaluation



# Model Selection/Evaluation




- **Selected Model: DNN**
  Accuracy: 88.6%
  AUC: 0.912

# Recommandation

- **Develop a real-time monitoring system, such as a dashboard, for sentiment analysis of customer reviews.**

Route information/alert to region HQs or stores
Monthly review process on sentiment

# Limitation & Improvement idea

- **Limitation**

1) Small size of dataset = 800 records & biased labels
   More data might be helpful for model performance, like recall score, without oversampling.
2) It is sentiment analysis, so cannot work on non-sentiment review
   i.e. Sifat went to Starbucks yesterday.
3) Only for US

- **Improvement**

1) Model tuning; hyper parameters & design
2) Continuous training is required
   *Watch-out: Service got better > More positive > Data drift
3) Summary of feedback; Negative feedbacks

# Conclusion

- Focus on **California & Order related Issues**
- Predictive model of sentiment analysis: DNN
  **Integrate to a real-time monitoring system**