# Optimizing Business Success on Yelp: Analyzing Factors for Star Ratings

ADSP 31008 - Group 10

Steven Luo
Jiayi Deng
Xiaobing Xu
Serena Shi
Jiawei Xu

# TABLE OF CONTENTS

# Problem Statement & Objectives

## Problem Statement

Yelp's challenge is to refine business onboarding to better reflect the impact of customer interactions and preferences on star ratings.

## Purpose and Objectives

1. Predict star ratings and understand driving factors
2. Equip businesses with actionable strategy to elevate their ratings
3. Boost user engagement through data models for Yelp

## Solution Value

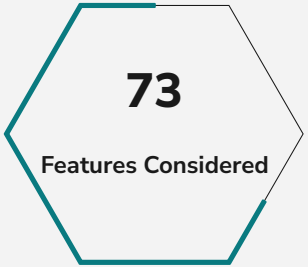
Analyze the Yelp dataset to predict star ratings, provide businesses with strategic insights to enhance user experience and foster a robust Yelp community.

## Target Variables

- **Primary Target Variable**
  - **Star Ratings (star_x)**: The average star rating a business receives.

- **Target City: Santa Barbara**
  - Tourist Attraction
    - Rich data for Yelp
  - High Economic Activities
    - Nightlife +, bars +
  - Cultural Diversity
  - Seasonal Business Pattern

## Definition & Scope of Features





**Main Categories for features:**

| | |
|---|---|
| **Business:**  <br>**Services options** (Takeout, caters), **dining experience** (ambience, outdoor seating), **presence of amenities** (e.g., TV, Wi-Fi, parking) | **Review:**  <br>Sentiment scores, frequency of positive/negative words, length of review. |
| **User Engagement:**  <br>Number of reviews received, average stars given by user, number of friends, elite status duration. | **Temporal Data:**  <br>Check-in times, frequency of reviews over time, temporal patterns in user activity, tips information. |

# Data Exploration

**5 Data Sets** (Business, User, Review, Checkin, Tips)
**150,346 businesses** (location, ratings, attributes)
**1,987,897 users**
**6,990,280 reviews** (reviews with rating and text)
**908,915 tips** (user tips with text and compliment)
**131,930 check-ins** (user check-ins at business)

**Primary Unit**

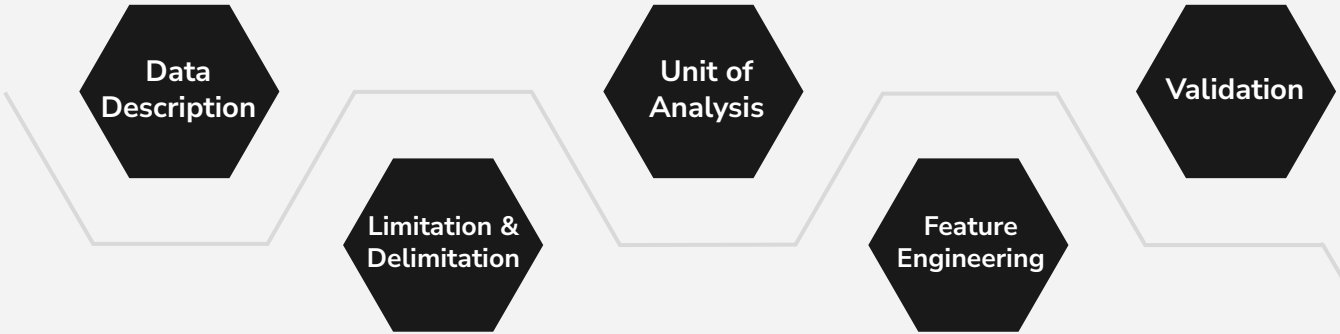
Businesses: Evaluated through ratings, categories, and attributes.

Users: Analyzed via reviews, tips, and engagement metrics.

**Secondary Unit**

Reviews: Insights on customer satisfaction and feedback.

**Cross Validation:** Employ cross-validation techniques to ensure model reliability across different subsets of data.




**Subjectivity: Check-in and tips,** being subjective, might introduce bias in understanding business quality so we didn't incorporate in feature engineering.

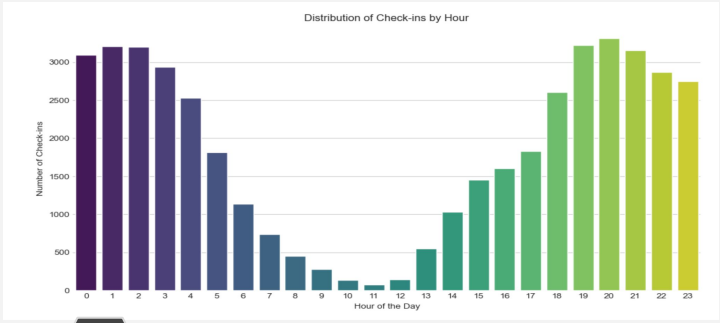
**Focus Area:** Analysis centers on **user interactions, review and business-related metrics.**

**Text Features:** Identify significant words or phrases (TF-IDF).

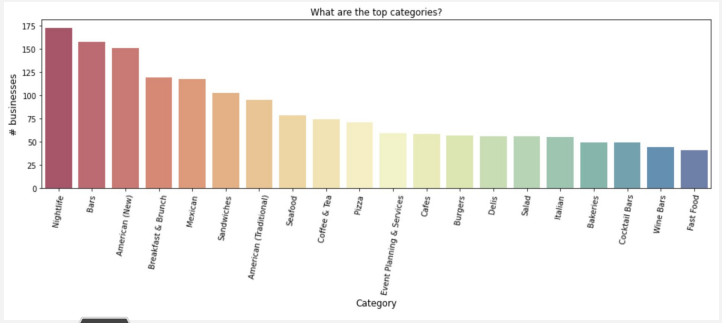
**Truncated SVD:** Avoid sparse matrix issue since user compliment/feedback/fans have a lot zeros.

**One-Hot Encoding:** Each category value is converted into a new column and assigned a 1 or 0 (notation for true/false) value to the column.

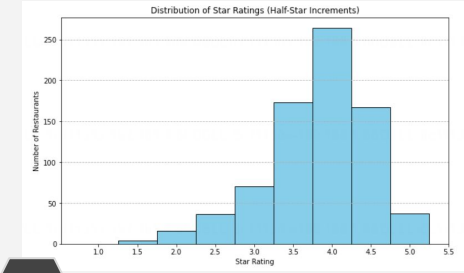
# Exploratory Data Analysis



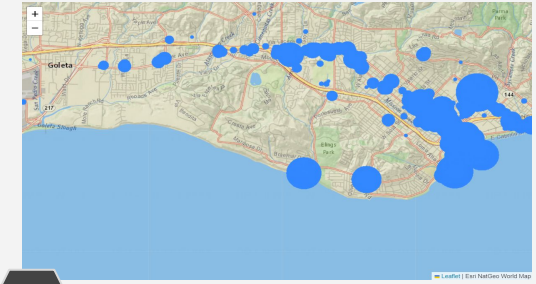

3 The distribution of check-in times is concentrated around **midnight**.




4 Nightlife is the top category

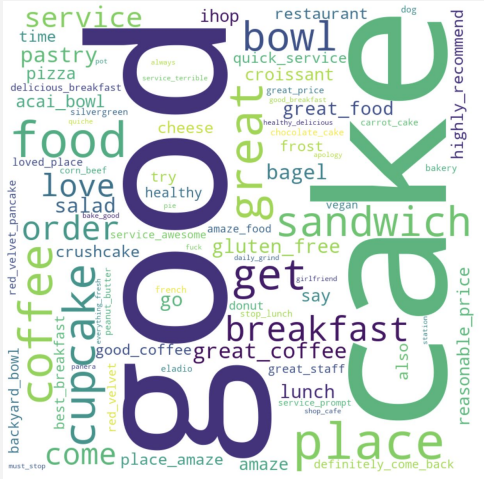



1 Star Ratings Distribution



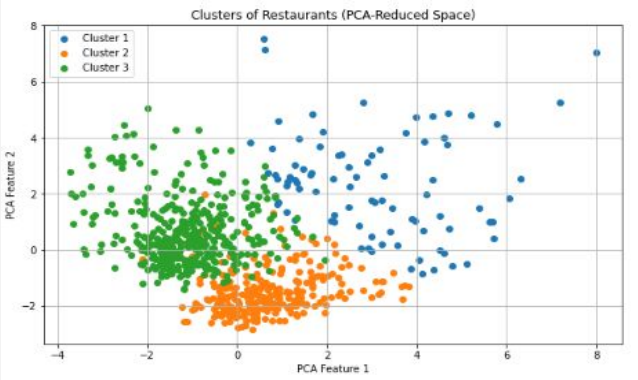

2 Geographic Business Density


5 Nightlife wordcloud (apply topic modeling)



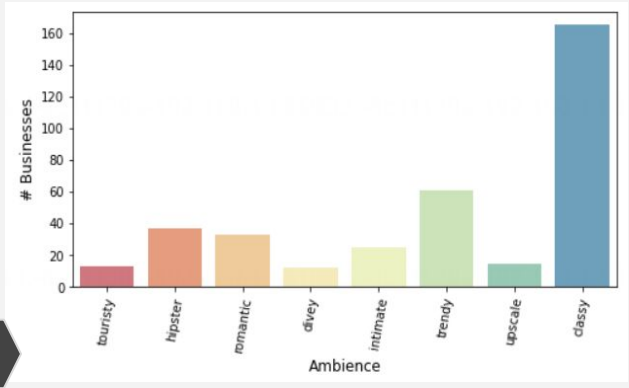

6 Breakfast & Brunch wordcloud

# Exploratory Data Analysis



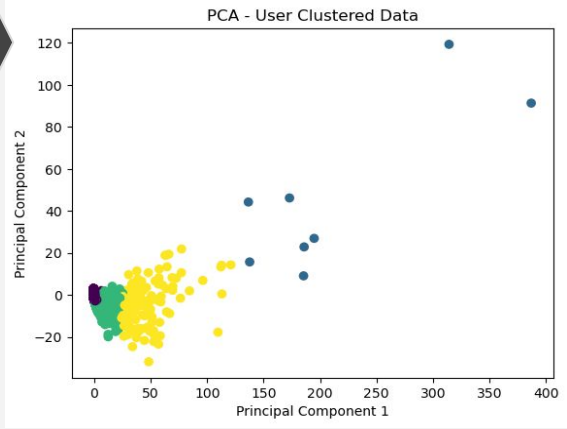

- Generate **loadings** of features
- Find clusters of restaurants

3



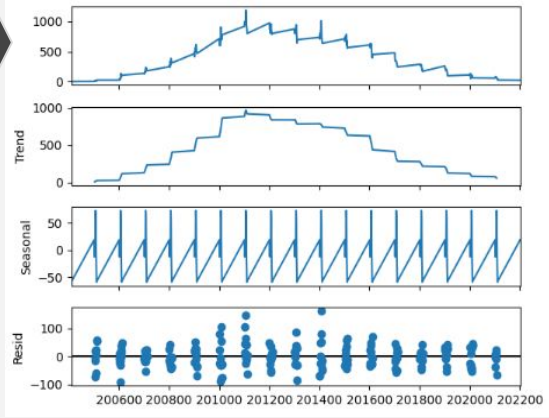

- Distribute of Ambience - (one of high loadings features)

4




- Used **T-SNE** to reduce sparse matrix
- Find KOL (more fans, high feedback(cool, funny))

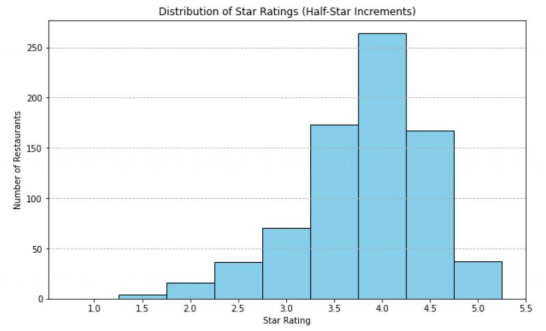
5



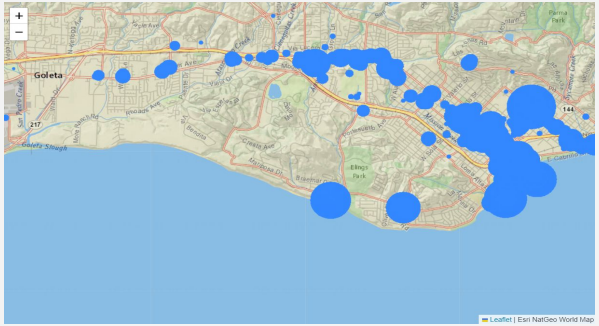

- A time series into **trend**, **seasonal**, and **residual** components for **users**
- Growth phase of user activity on Yelp, peaking around 2014 and then decrease

# Exploratory Data Analysis

## General Info Yelp Reviews in Santa Barbara



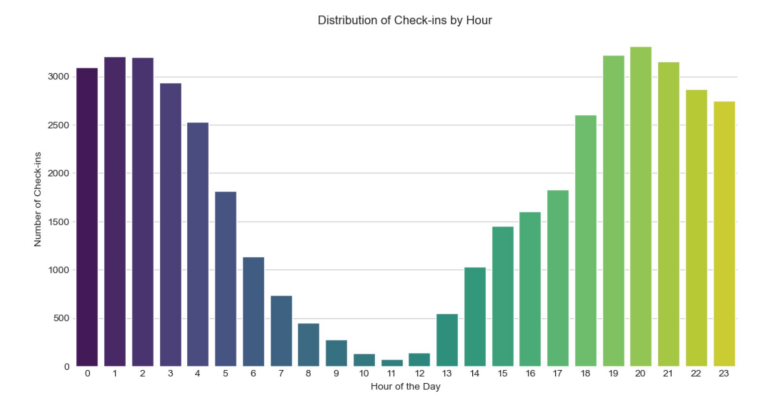

Star Ratings Distribution



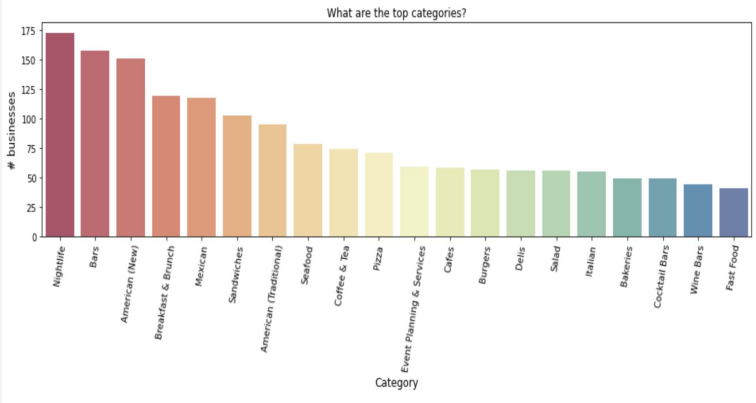

Geographic Business Density

# Exploratory Data Analysis

## Deep Dive into Consumer Behavior and Preferences




Check-in time concentrated around midnight

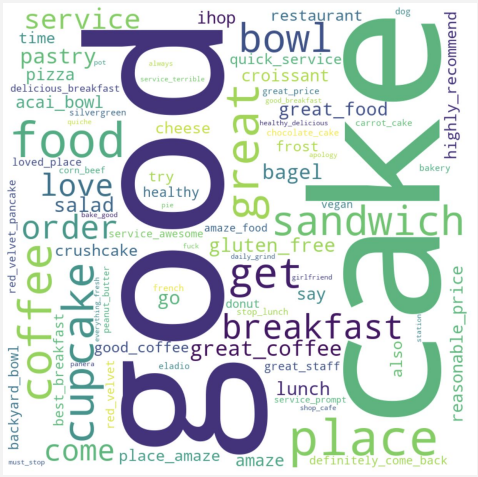



Nightlife and bars are the top two categories

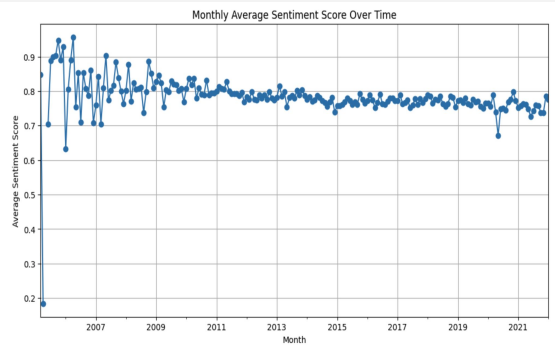
# Exploratory Data Analysis

## Topic Modeling & Sentiment Analysis towards reviews


Nightlife Wordcloud




Breakfast & Brunch Wordcloud



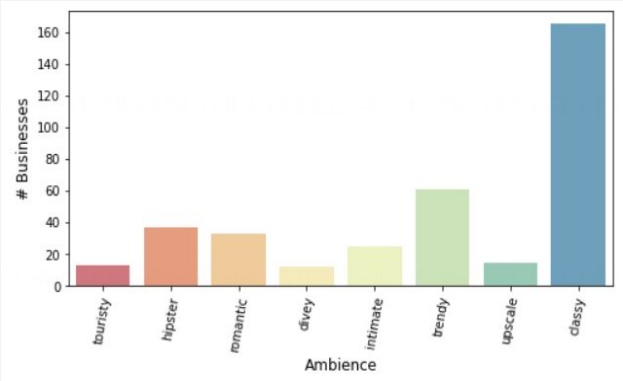

Time series analysis of **sentiment scores**

- Converge to mean score of 0.75

# Exploratory Data Analysis

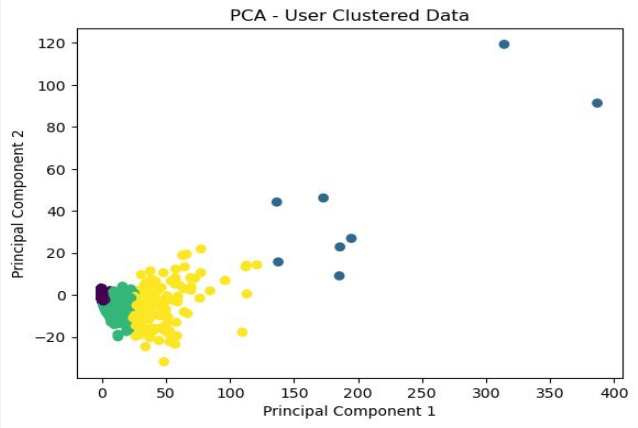
## Business Features Analysis


- Generate **loadings** of features
- Find clusters of restaurants

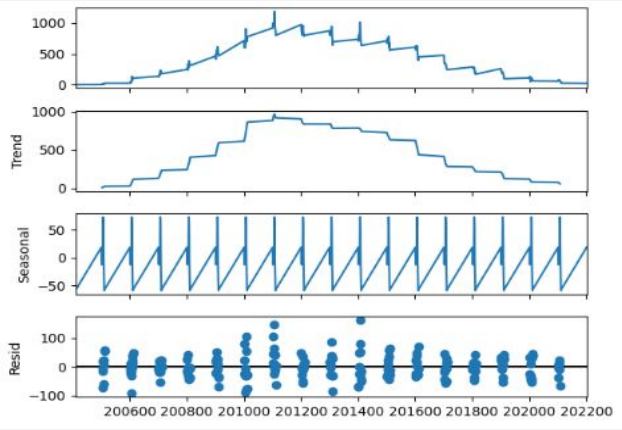



- Distribute of Ambience - (one of high loadings features)

## User Data Analysis



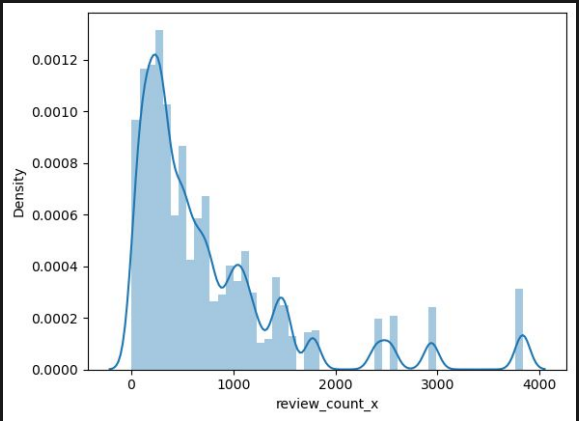

- Used **T-SNE** to reduce sparse matrix
- Find KOL (more fans, high feedback(cool, funny))




- A time series into **trend**, **seasonal**, and **residual** components for **users**
- Growth phase of user activity on Yelp, peaking around 2014 and then decrease

# Methodology: Feature Engineering

## 1. Numerical




## 2. Categorical (HC)

- Casual Dining & Drinks
- Diverse Cuisine & Entertainment
- Health & Lifestyle
- Nightlife Essentials
- Sophisticated Eats
- Traditional Comfort Foods

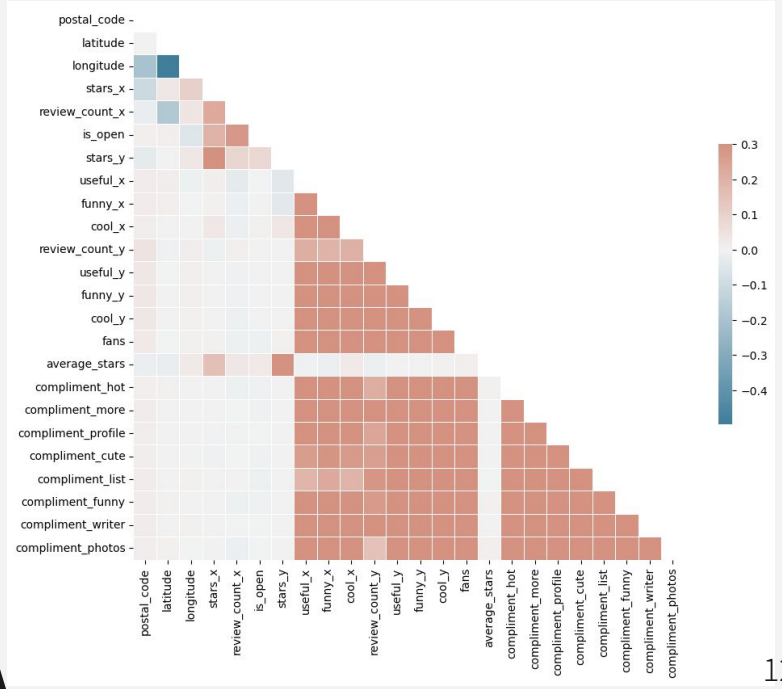
## 3. Date

Number of days using Yelp

## 4. Text(TF-IDF Vectorizer)

- Preprocessed the text data
- Removed stopwords
- Lemmatized the text
- Vectorized with TF-IDF

## 5. Sparse-Matrix (Truncated SVD)



# Methodology
# Modeling Framework

## Model Descriptions

Baseline:
– Linear Regression

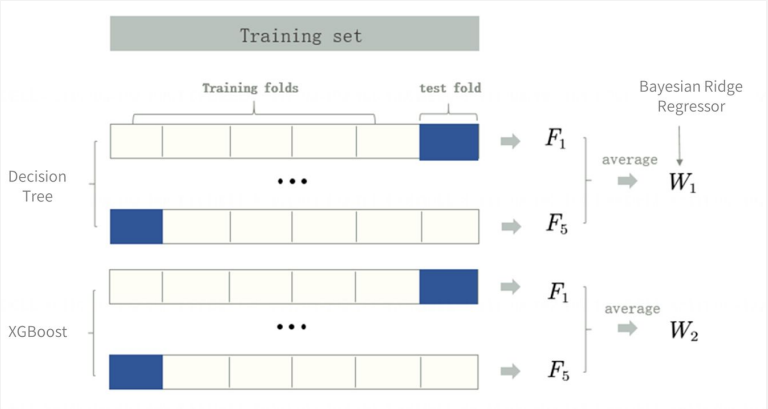
Non-Linear:
– SVM (linear & non-linear)
– Decision Tree
– Random Forest
– AdaBoost
– XGBoost

– Customized Ensemble Model: Stacking the base models above using **Bayesian Ridge Regression**

## Model Selection and Validation

**Model Performance Evaluation:**
CV error > training set error: overfitting
CV error ≈ training set error >> test set error: underfitting

**Randomized Search Cross-validation (5-fold)** was employed to assess model generalizability, with **RMSE (Root Mean Squared Error)** as the performance metric to compare different models' predictive accuracy.

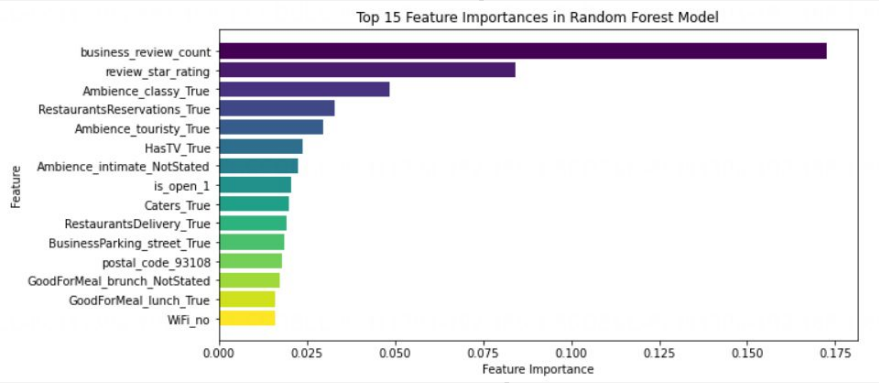



Ensemble Model -> Stacking

# Modeling Results

| | Default Parameters | | Cross-Validation (5-fold) | | | Stacking |
|---|---|---|---|---|---|---|
| **Model** | **Training RMSE** | **Test RMSE** | **CV RMSE** | **Training RMSE** | **Test RMSE** | |
| **Linear Regression** | 0.36 | 0.36 | 0.36 | 0.36 | 0.36 | |
| **SVM** | 0.47 (linear) | 0.47 (linear) | 0.078 (rbf) | 0.057 (rbf) | 0.077 (rbf) | |
| **Decision Tree** | 0.0 | 0.04 | 0.051 | 0.012 | 0.042 | |
| **Random Forest** | 0.014 | 0.039 | 0.045 | 0.018 | 0.039 | **Bayesian Ridge (Decision Tree, XGBoost)** |
| **AdaBoost** | 0.41 | 0.41 | 0.41 | 0.40 | 0.41 | Training RMSE: 0.022<br>Test RMSE: 0.035 |
| **XGBoost** | 0.064 | 0.074 | 0.050 | 0.025 | 0.047 | |

# Findings & Recommendations




**Impact of Review Volume and Quality**

Higher review counts and quality positively impact Yelp star ratings, indicating popularity and customer satisfaction.

***Encouraging Customer Reviews:*** Implement a feedback initiative, incentivize reviews, and train staff to prompt satisfied customers for feedback.

**Role of Key Opinion Leaders (KOLs)**

KOLs have a significant impact on star ratings due to their following and credibility.

***KOL Engagement:*** Invite KOLs for unique dining experiences, develop ongoing collaborations, and utilize their feedback for improvement and credibility-building.

**Significance of Ambience and Service:**

Ambience, service, and staff attitude strongly influence positive reviews.

***Focus on Experience and Service:*** Regularly update ambience and service based on feedback, consider customer preferences in strategic planning.

# Future Steps

## Improve Model

**Overfitting Mitigation:** Enhance overfitting prevention with improved regularization, better feature selection, and advanced validation techniques.

**Deep Learning Techniques:** Evaluate the use of deep learning models, like RNNs, to capture complex patterns in review sequences.

**Enhanced Sentiment Analysis:** Implement state-of-the-art NLP models (BERT, GPT) for nuanced understanding of customer sentiment in reviews.

## Improve Business

**Customer Experience Personalization:**

- Use review insights to customize experiences.
- Target marketing to diverse customer segments.

**Operational Excellence:**

- Utilize analytics for inventory, staffing, and menu optimization.
- Implement real-time quality control from Yelp feedback.

**Reputation Management and Response:**

- Establish a team to engage with reviews.
- Turn negative reviews into public service opportunities.

**KOL and Influencer Collaborations:**

- Partner with local influencers for events and promotions.
- Track influencer impact on customer perceptions.

## Improve Yelp Platform

**Enhanced Analytics Dashboard for Businesses:** Develop advanced dashboards with predictive analytics to provide businesses insights, enabling anticipation of trends and customer needs.

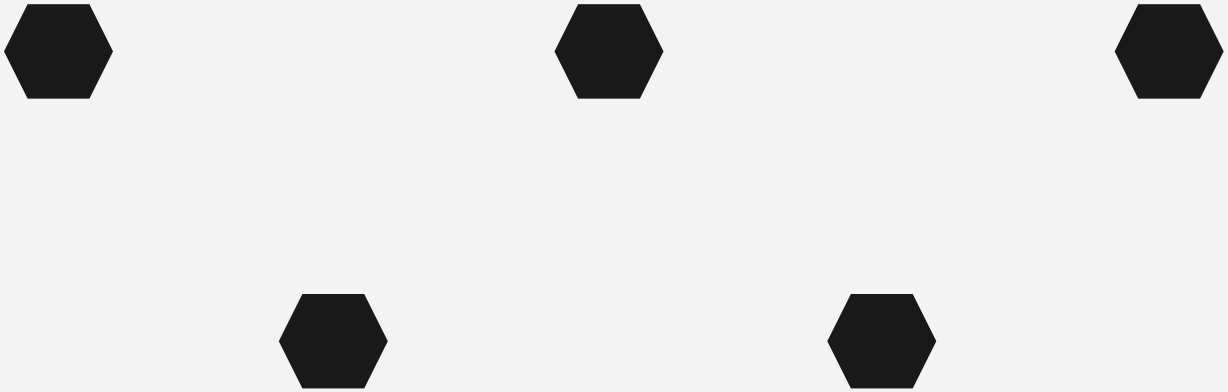
**Business Engagement Tools:** Create tools within Yelp that make it easier for businesses to engage with customers, respond to reviews, and manage their online reputation.

**Personalized Business Recommendations:** Use Yelp's data to offer personalized recommendations to businesses for improvement based on similar successful businesses in their area.


# THANKS

Open to questions!

# Results & Recommendations



# Reference

This is where you give credit to the ones who are part of this project.

- Presentation template by Slidesgo
- Icons by Flaticon
- Infographics by Freepik
- Images created by Freepik - Freepik
- Author introduction slide photo created by Freepik
- Text & Image slide photo created by Freepik.com

# RESOURCES

Did you like the resources on this template? Get them for free at our other websites.

## VECTORS

- Pack human resources 11
- Pack social media 24

## PHOTOS

- Reflection of rocky mountains and sky on beautiful lake
- Breathtaking view of sunnylvsfjorden fjord and famous seven sisters waterfall; norway
- Side view of man thinking in office while looking at sticky notes
- Medium shot women working together
- Red boat moored on the idyllic lake near the rocky mountains
- Leopard
- Rear view of a man with black backpack standing on bridge
- Medium shot smiley women making business plan
- Medium shot bored people working
- People assisting at business meeting
- Silhouette of a person's hand holding paper airplane against dramatic sky
- Railway train tracks with platforms

## Methodology

# Fonts & colors used

This presentation has been made using the following fonts:

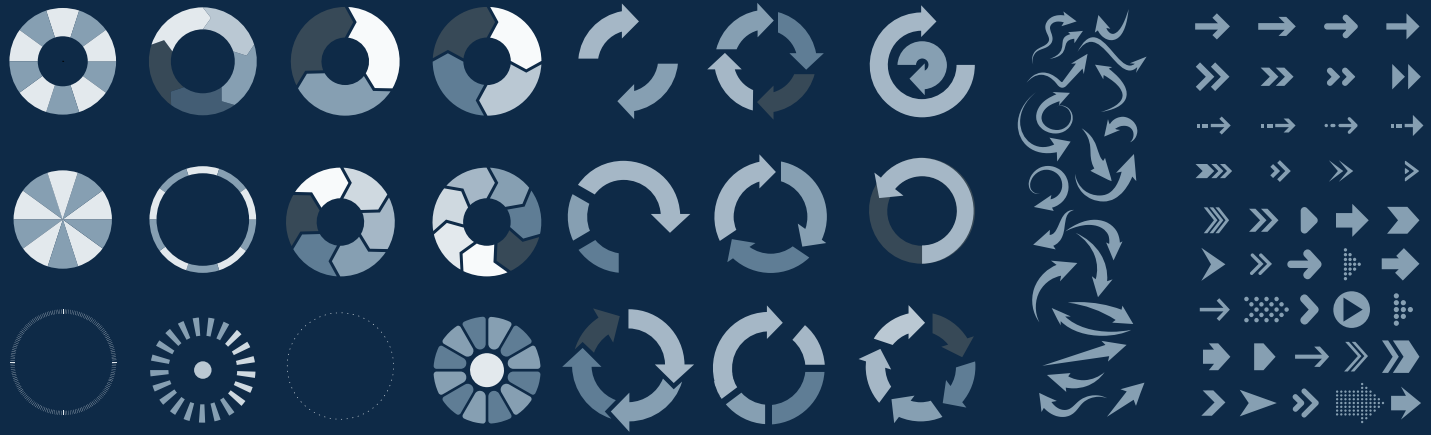
**Nunito Sans**
(https://fonts.google.com/specimen/Nunito+Sans)

**Assistant**
(https://fonts.google.com/specimen/Assistant)

#191919 #d9d9d9 #f3f3f3 #097a80

# Use our editable graphic resources...

You can easily resize these resources, keeping the quality. To change the color, just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want. Don't forget to group the resource again when you're done.



# ...and our set of editable icons

You can resize these icons, keeping the quality.
You can change the stroke and fill color; just select the icon and click on the paint-bucket/pen.
In Google Slides, you can also use Flaticon's extension, allowing you to customize and add even more icons.

# Business Icons

# Avatar Icons

# Creative Process Icons

# Educational Process Icons

# Help & Support Icons

# Medical Icons

# Nature Icons

# Performing Arts Icons

# SEO & Marketing Icons

# Teamwork Icons


slidesgo