Predicting the 'Usefulness' of Yelp Reviews

Presented by Gyujin Seo | UCLA | Major in Statistics and Data Science

Table of Contents

1. Introduction
2. Pre-processing
3. Modeling
4. Results and Conclusion
5. Future Research

1. Introduction

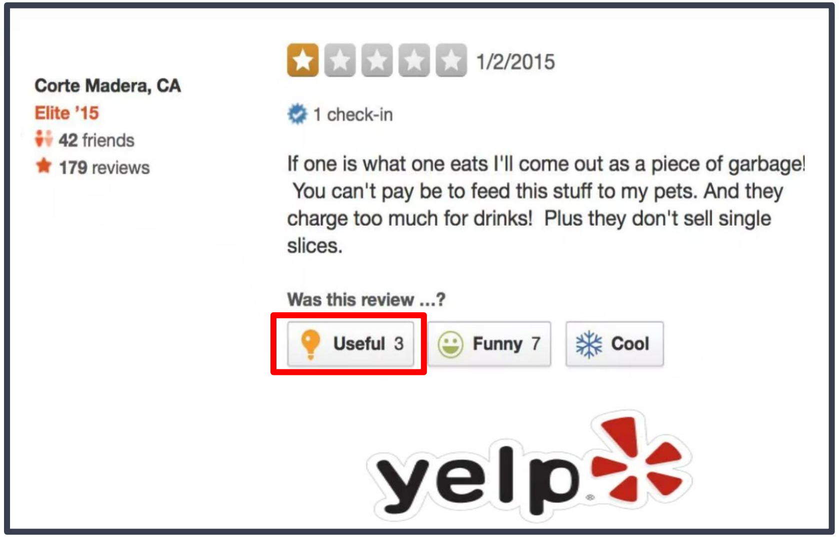
- **Background**
  - Today, we live an era where users' experiences are highly valued and expressing them is crucial.
  - Companies recognize the importance of such experiential data(review data) and utilize them as significant marketing resources.
  - Through the "yelp" website, which provides consumer review datasets, these data are collected.

- **Problem**
  - Excessive time spent on purchase decision-making due to the ambiguity of consumer review data reliability.
  - Companies spending excessive time determining the usefulness of review data.

1. Introduction(Cont'd)

- **Aim of my Project**
  - To predict the usefulness of a review prior to user interaction.

- **Yelp Overview**
  - Yelp is a consumer-driven platform providing business information, photos, and user reviews.
  - The **'useful'** button on Yelp, which plays a critical role in guiding user decisions.



1/2/2015

Corte Madera, CA
Elite '15
42 friends
179 reviews

1 check-in

If one is what one eats I'll come out as a piece of garbage! You can't pay be to feed this stuff to my pets. And they charge too much for drinks! Plus they don't sell single slices.

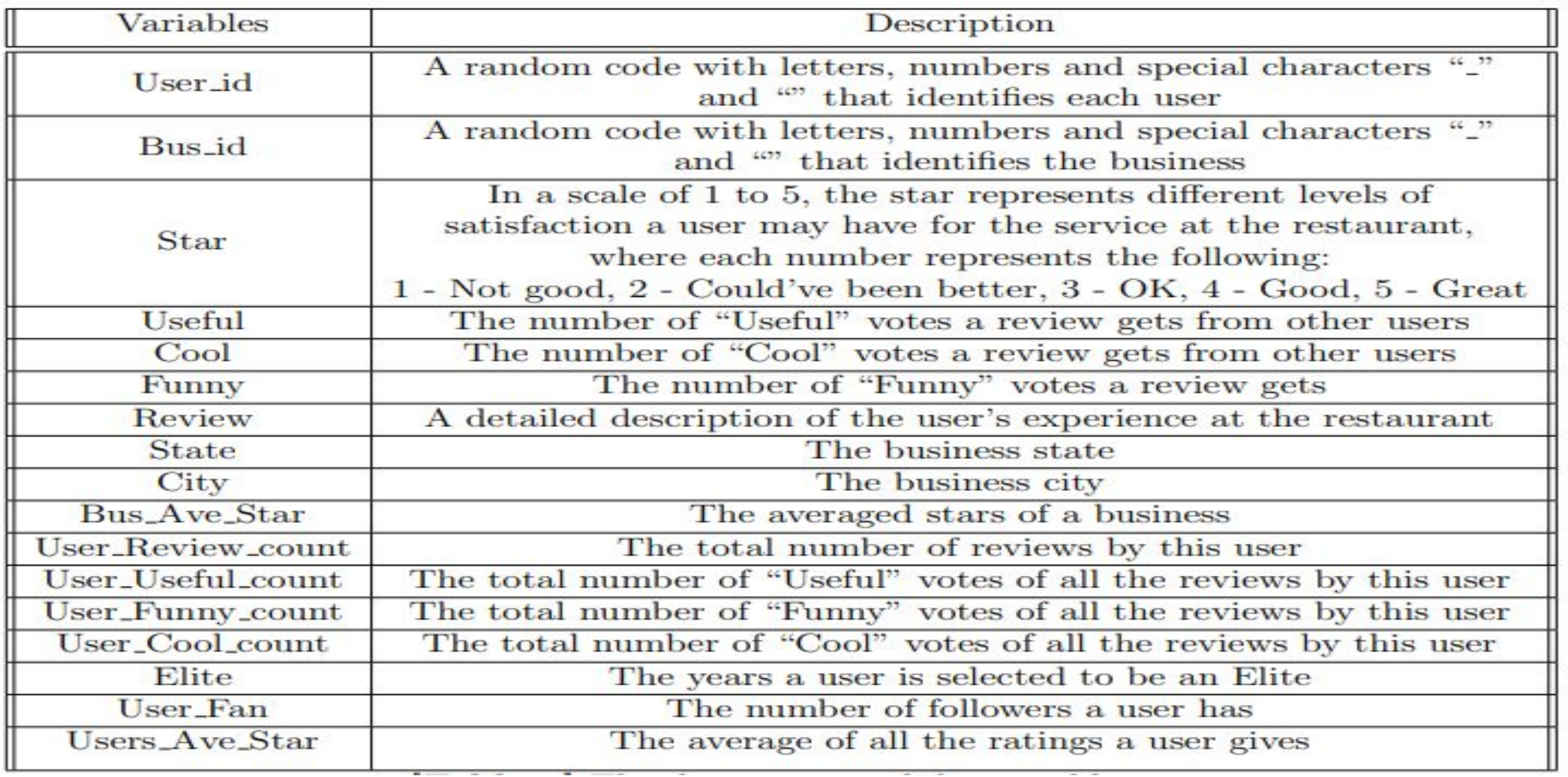
Was this review ...?

1. Introduction(Cont'd)

- **Programming language**
  - R

- **Data overview**
  - It has 53,845 unique observations, with 17 variables.
  - For the variables, I have User ID, Business ID, Star, Useful, Cool, Funny, Review, State, City, Business average count, User review count, User useful count, Userfunny count, User cool count, Elite, User fans, and User average star.

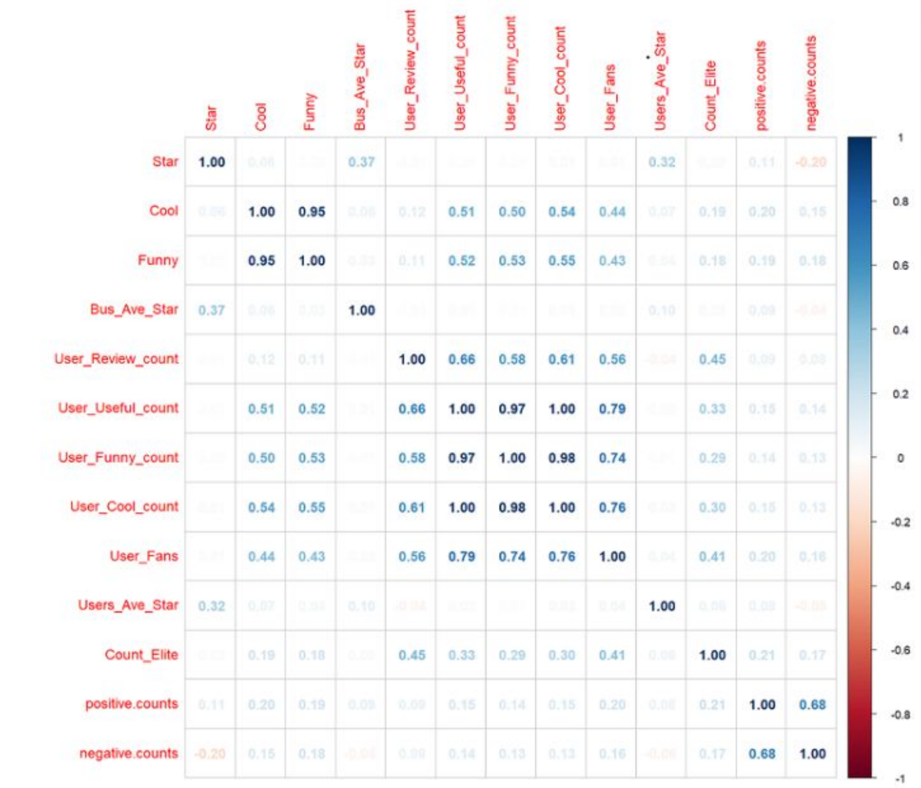
1. Introduction(Cont'd)



Variables	Description
User_id	A random code with letters, numbers and special characters “_” and “” that identifies each user
Bus_id	A random code with letters, numbers and special characters “_” and “” that identifies the business
Star	In a scale of 1 to 5, the star represents different levels of satisfaction a user may have for the service at the restaurant, where each number represents the following: 1 - Not good, 2 - Could've been better, 3 - OK, 4 - Good, 5 - Great
Useful	The number of “Useful” votes a review gets from other users
Cool	The number of “Cool” votes a review gets from other users
Funny	The number of “Funny” votes a review gets
Review	A detailed description of the user's experience at the restaurant
State	The business state
City	The business city
Bus_Ave_Star	The averaged stars of a business
User_Review_count	The total number of reviews by this user
User_Useful_count	The total number of “Useful” votes of all the reviews by this user
User_Funny_count	The total number of “Funny” votes of all the reviews by this user
User_Cool_count	The total number of “Cool” votes of all the reviews by this user
Elite	The years a user is selected to be an Elite
User_Fan	The number of followers a user has
Users_Ave_Star	The average of all the ratings a user gives

2. Pre-processing

- Transforming Useful to binary: "Not Useful" (0) if less than or equal to 1, "Useful" (1) otherwise.
- Review analysis with NLP: Counted positive and negative words in each review.
- Eliminating highly correlated variables: Dropped Funny, User Funny count, User Cool count due to high correlation (>90%).



	Star	Cool	Funny	Bus_Ave_Star	User_Review_count	User_Useful_count	User_Funny_count	User_Cool_count	User_Fans	Users_Ave_Star	Count_Elite	positive.counts	negative.counts
Star	1.00	0.06	[illegible]	0.37	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.32	[illegible]	0.11	-0.20
Cool	0.06	1.00	0.95	0.06	0.12	0.51	0.50	0.54	0.44	0.07	0.19	0.20	0.15
Funny	[illegible]	0.95	1.00	[illegible]	0.11	0.52	0.53	0.55	0.43	0.04	0.18	0.19	0.18
Bus_Ave_Star	0.37	0.06	[illegible]	1.00	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.10	[illegible]	0.09	-0.04
User_Review_count	[illegible]	0.12	0.11	[illegible]	1.00	0.66	0.58	0.61	0.56	-0.04	0.45	0.09	0.08
User_Useful_count	[illegible]	0.51	0.52	[illegible]	0.66	1.00	0.97	1.00	0.79	[illegible]	0.33	0.15	0.14
User_Funny_count	[illegible]	0.50	0.53	[illegible]	0.58	0.97	1.00	0.98	0.74	[illegible]	0.29	0.14	0.13
User_Cool_count	[illegible]	0.54	0.55	[illegible]	0.61	1.00	0.98	1.00	0.76	[illegible]	0.30	0.15	0.13
User_Fans	[illegible]	0.44	0.43	[illegible]	0.56	0.79	0.74	0.76	1.00	0.04	0.41	0.20	0.16
Users_Ave_Star	0.32	0.07	0.04	0.10	-0.04	[illegible]	[illegible]	[illegible]	0.04	1.00	0.06	0.08	-0.06
Count_Elite	[illegible]	0.19	0.18	[illegible]	0.45	0.33	0.29	0.30	0.41	0.06	1.00	0.21	0.17
positive.counts	0.11	0.20	0.19	0.09	0.09	0.15	0.14	0.15	0.20	0.08	0.21	1.00	0.68
negative.counts	-0.20	0.15	0.18	-0.04	0.08	0.14	0.13	0.13	0.16	-0.06	0.17	0.68	1.00

3. Modeling

3.1 Model Validation

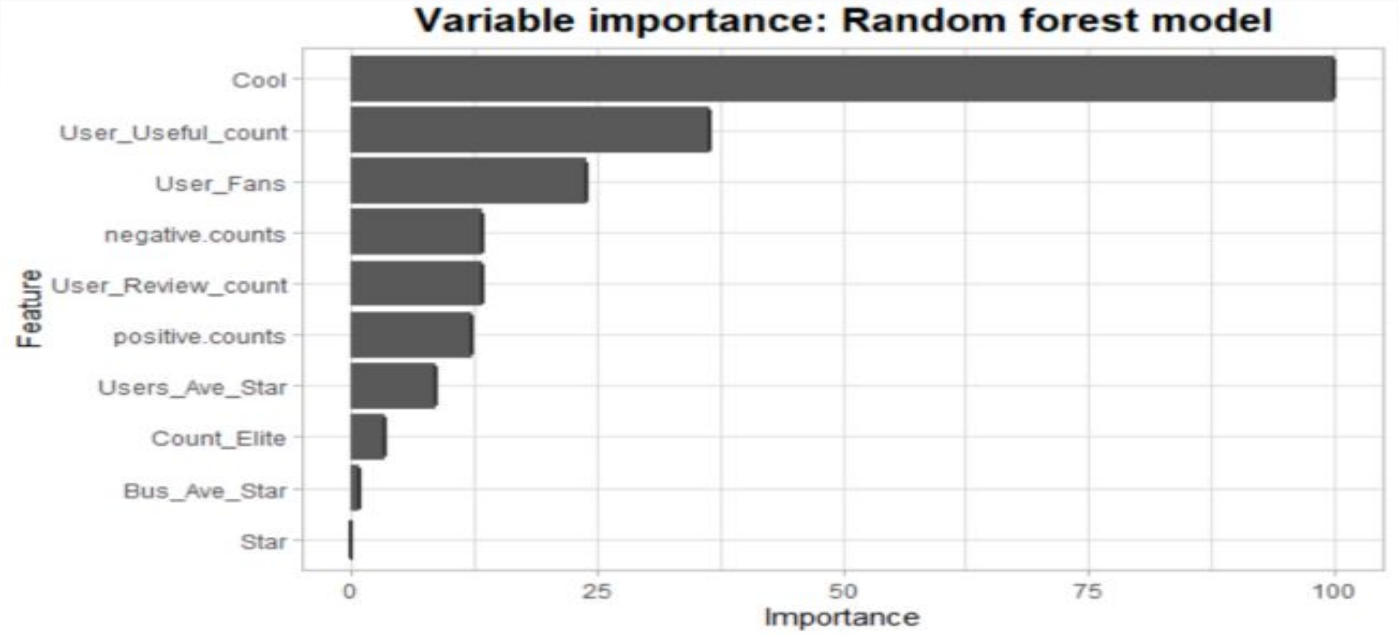
- **Data Partitioning: 70% Training Data, 30% Test Data.**
- **Cross-Validation: Applied 10-fold Cross-Validation for more accurate performance estimation.**

3.2 Methods

- **5 Machine Learning Methods Applied:**
  - **Random Forests**
  - **Logistic Regression**
  - **LASSO Regression**
  - **Decision Tree**
  - **K-NN**

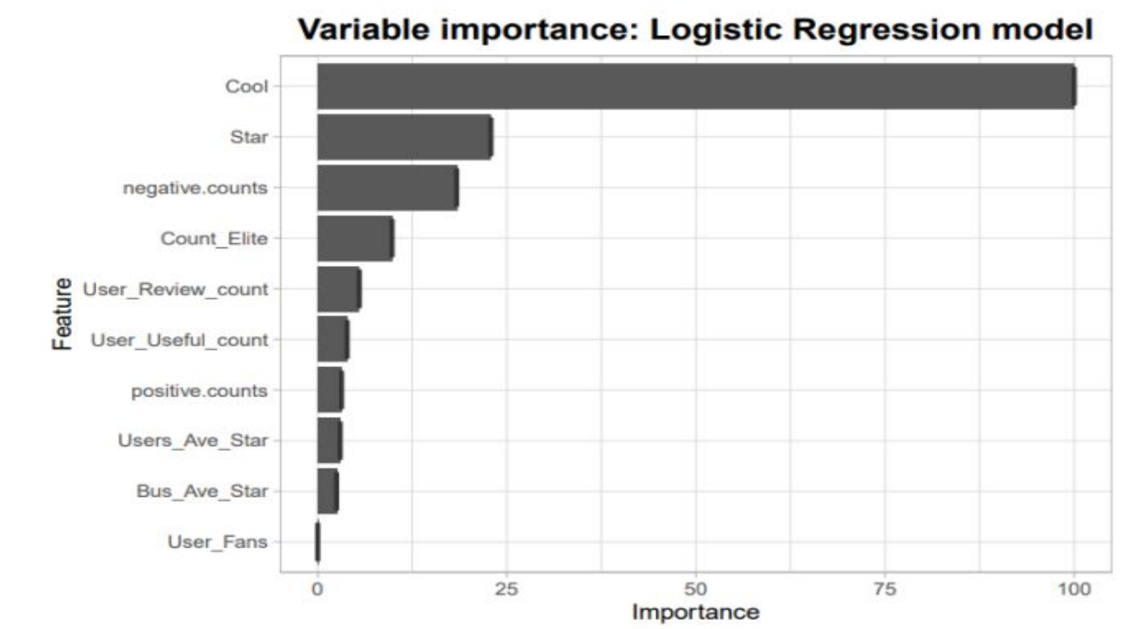
3. Modeling(Cont'd)

3.2 Methods - Random Forest



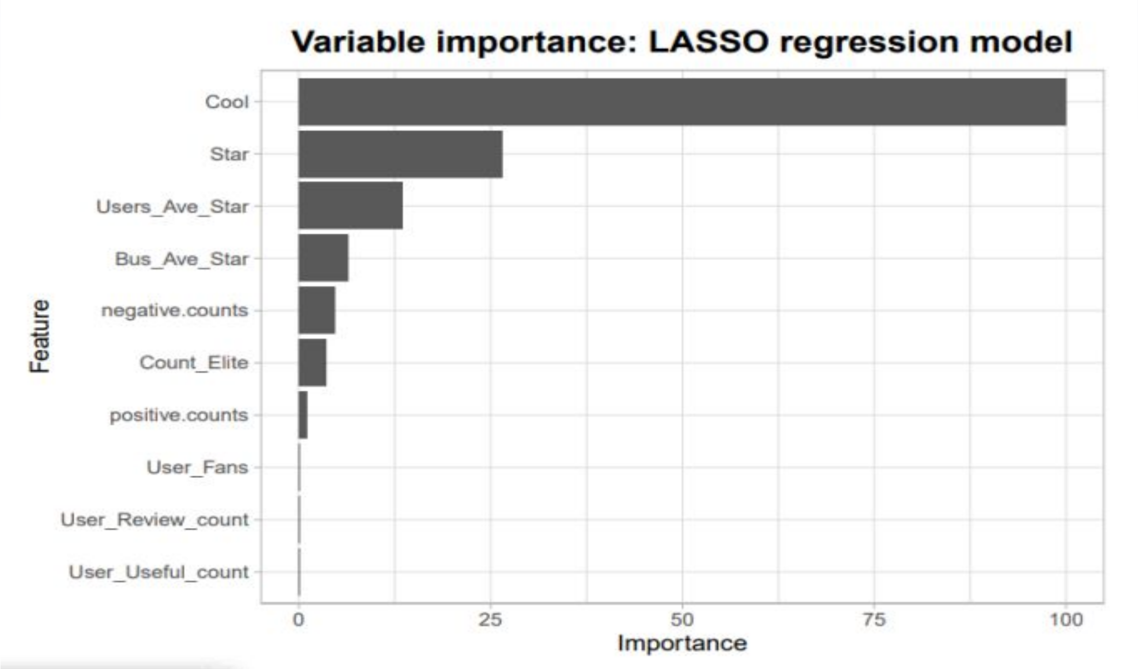
3. Modeling(Cont'd)

3.2 Methods - Logistic Regression



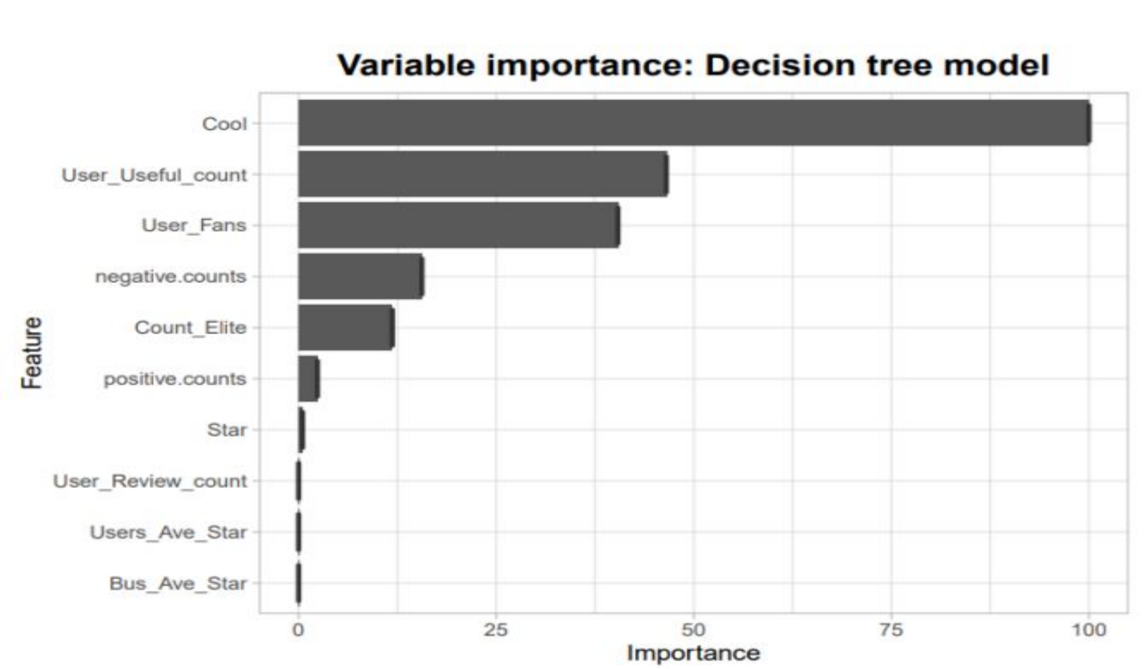
3. Modeling(Cont'd)

3.2 Methods - LASSO Regression



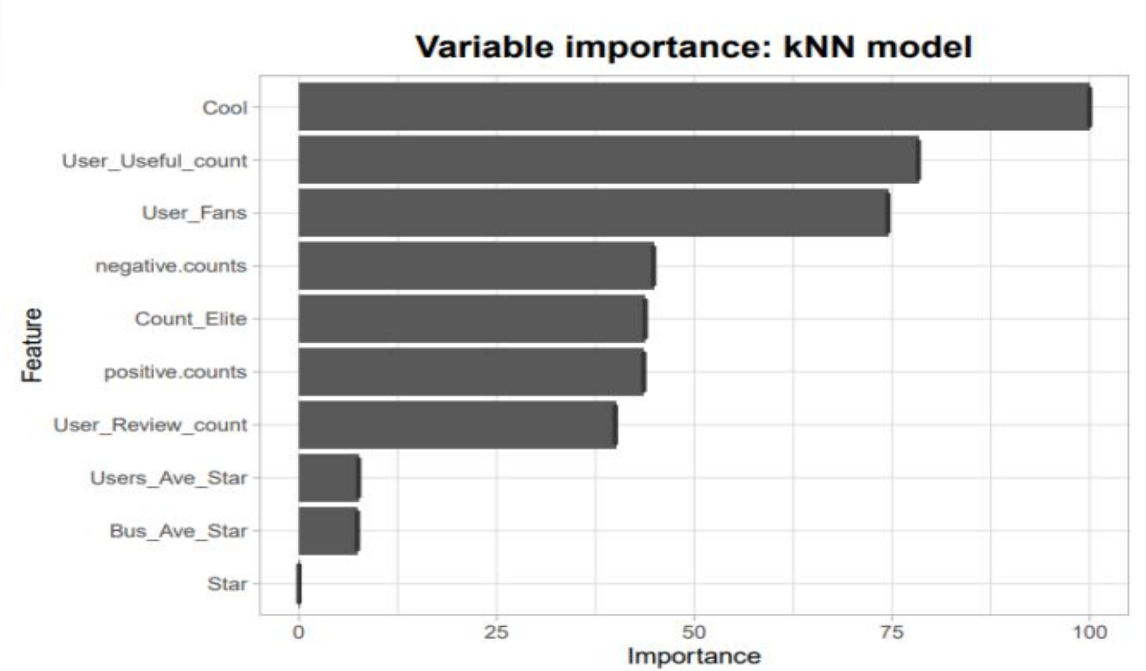
3. Modeling(Cont'd)

3.2 Methods - Decision Tree

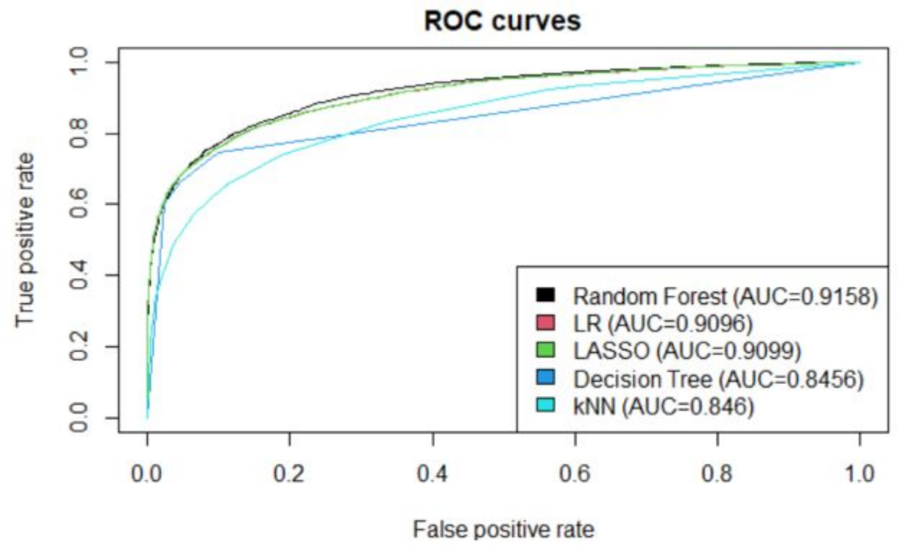


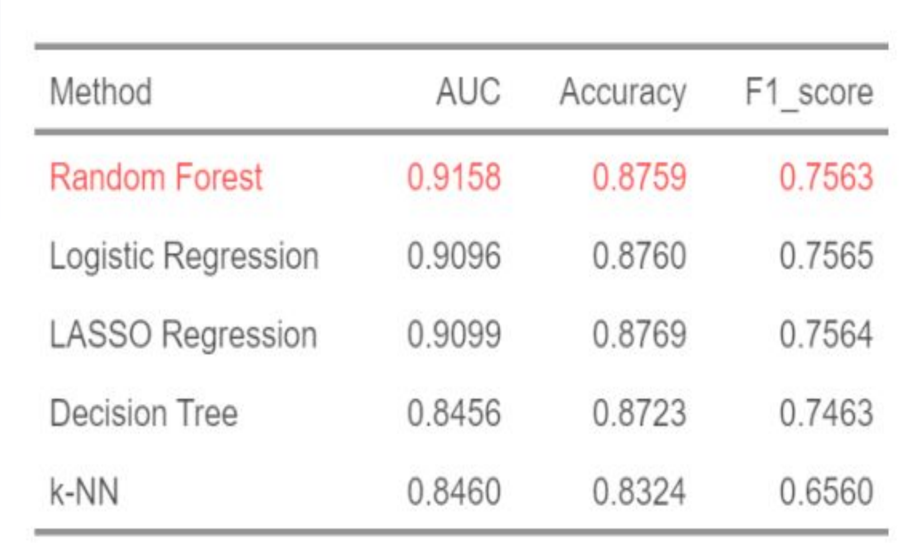
3. Modeling(Cont'd)

3.2 Methods - K-NN



4. Results and Conclusion





Method	AUC	Accuracy	F1_score
Random Forest	0.9158	0.8759	0.7563
Logistic Regression	0.9096	0.8760	0.7565
LASSO Regression	0.9099	0.8769	0.7564
Decision Tree	0.8456	0.8723	0.7463
k-NN	0.8460	0.8324	0.6560

- **‘Random Forest’ has an AUC score of 91.58% that shows the model is learning the data well enough.**

4. Results and Conclusion(cont'd)

- **Consumer Perspective**
  - By predicting the usefulness of reviews, we can secure the trust of buyers and expedite their purchase decision-making process.

- **Restaurant Perspective**
  - Predicting the 'usefulness' of reviews not only allows for swift utilization as marketing material, but can also enhance customer service satisfaction.

5. Future Research

- **Research to derive meaningful patterns and insights from the given dataset using various data mining techniques.**
  - Using algorithms like K-means clustering for customer segmentation.
- **Research on the development of predictive models using machine learning algorithms.**
  - Used for trend prediction, user behavior prediction, recommendation systems, etc.
- **Research to improve model performance through optimization theory and techniques.**
  - Solving issues such as model complexity, overfitting, and underfitting.

Thank you!