HOUSE PRICE PREDICTION

PRESENTED BY: SHRAVANI KSHEERASAGAR

DATA INTRODUCTION

PROBLEM STATEMENT:

To forecast housing prices based on factors like area, number of bedrooms, furnishing status, and proximity to main roads.

DATA DESCRIPTION

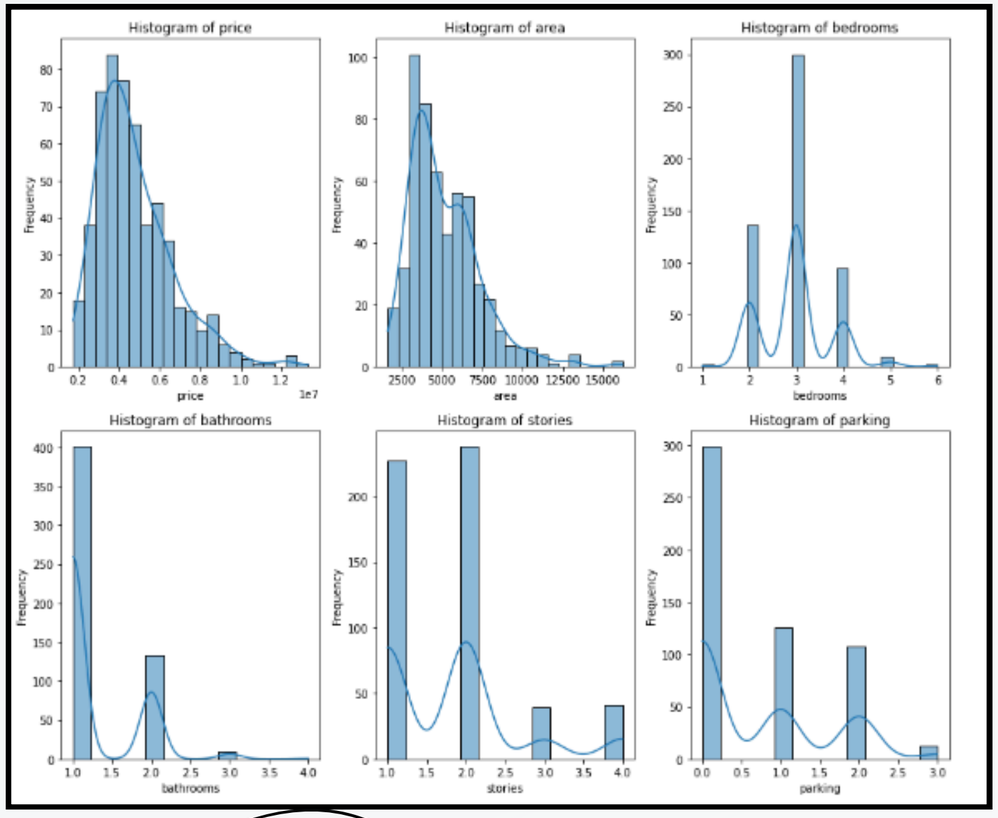
OBJECTIVE: PREDICT HOUSING PRICES BASED ON VARIOUS FEATURES.
DATASET: CONTAINS 545 RECORDS WITH 13 FEATURES.
FEATURES INCLUDE:PRICE, AREA, BEDROOMS, BATHROOMS, STORIES,MAINROAD, GUESTROOM, BASEMENT, HOTWATER,HEATING, AIRCONDITIONING,PARKING,PREFAREA, FURNISHING STATUS.
DATA TYPES:

- NUMERICAL: PRICE, AREA, BEDROOMS, BATHROOMS, STORIES, PARKING
- CATEGORICAL: MAINROAD, GUESTROOM, BASEMENT, HOTWATER HEATING, AIRCONDITIONING, PREFAREA, FURNISHING STATUS

NO MISSING VALUES.
FEATURES ENCODED AS NEEDED.

LITERATURE REVIEW

Univariate Analysis



Interpretation:

Price: Most house prices fall within the range of four million dollars.

Area: The majority of houses range from 2500 sq feet to 5000 sq feet in size.

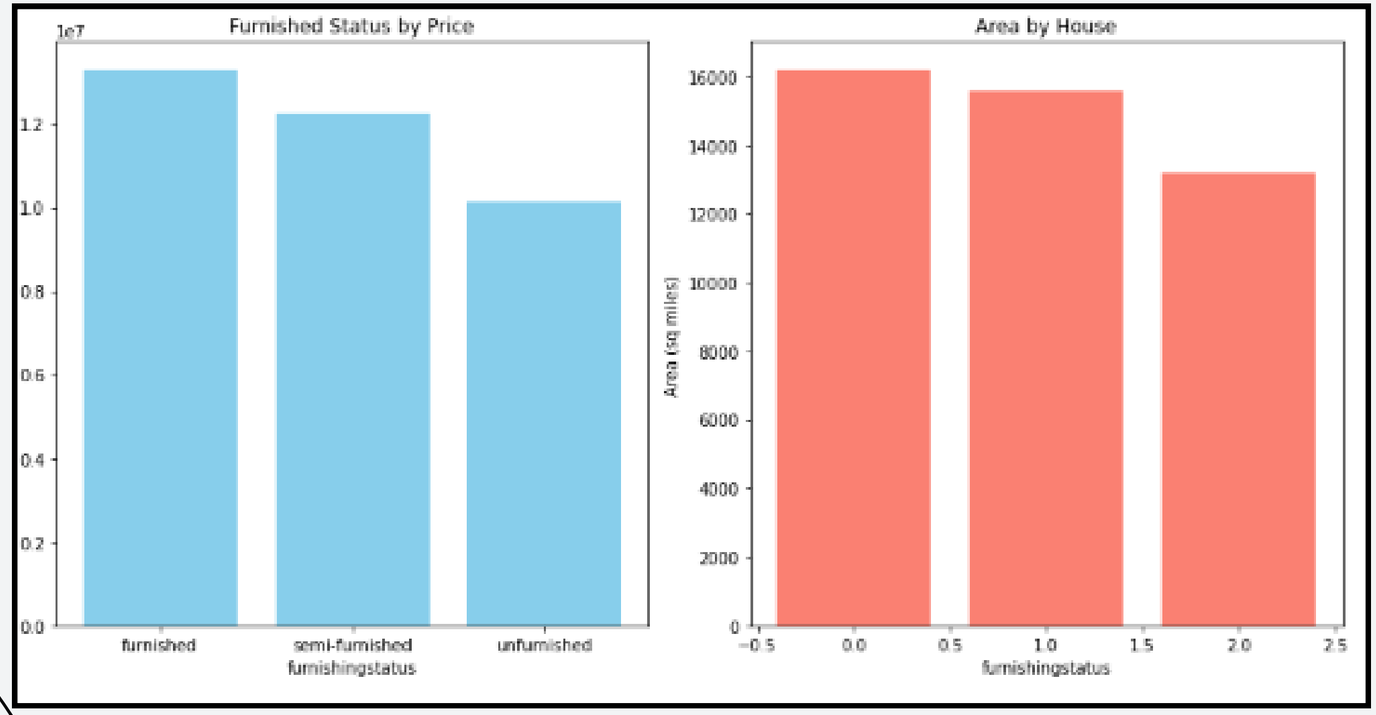
Bedrooms: Most houses have three bedrooms.

Bathrooms: The majority of houses have one bathroom.

Stories: Most houses are one to two stories tall.

Parking: Most houses do not have a designated parking area.

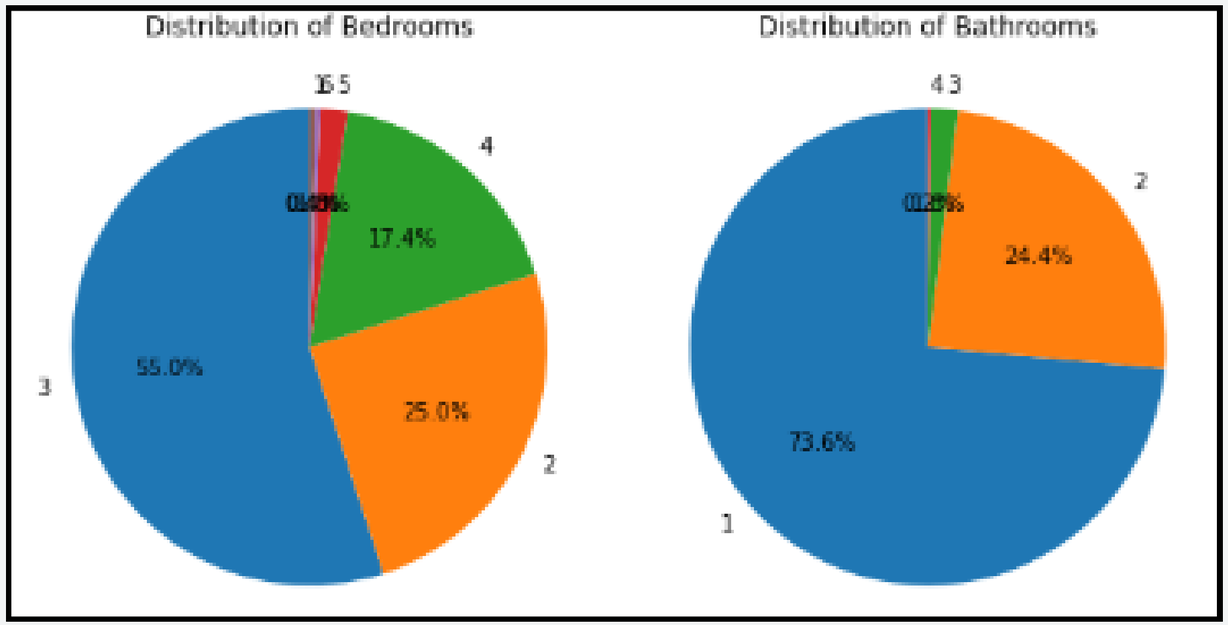
Bivariate Analysis



Interpretation:

Higher-priced houses are furnished.

Furnished options are available for larger houses.



METHODOLOGY

Initial model to understand baseline performance.
MSE: 1,466,037,860,959.45
R2 Score: 0.6591
MAPE: 17.4%

LINEAR REGRESSION

Adding regularization to handle multicollinearity.
MSE: 1,466,577,353,548.05
R2 Score: 0.6590
MAPE: 17.4%

RIDGE REGRESSION

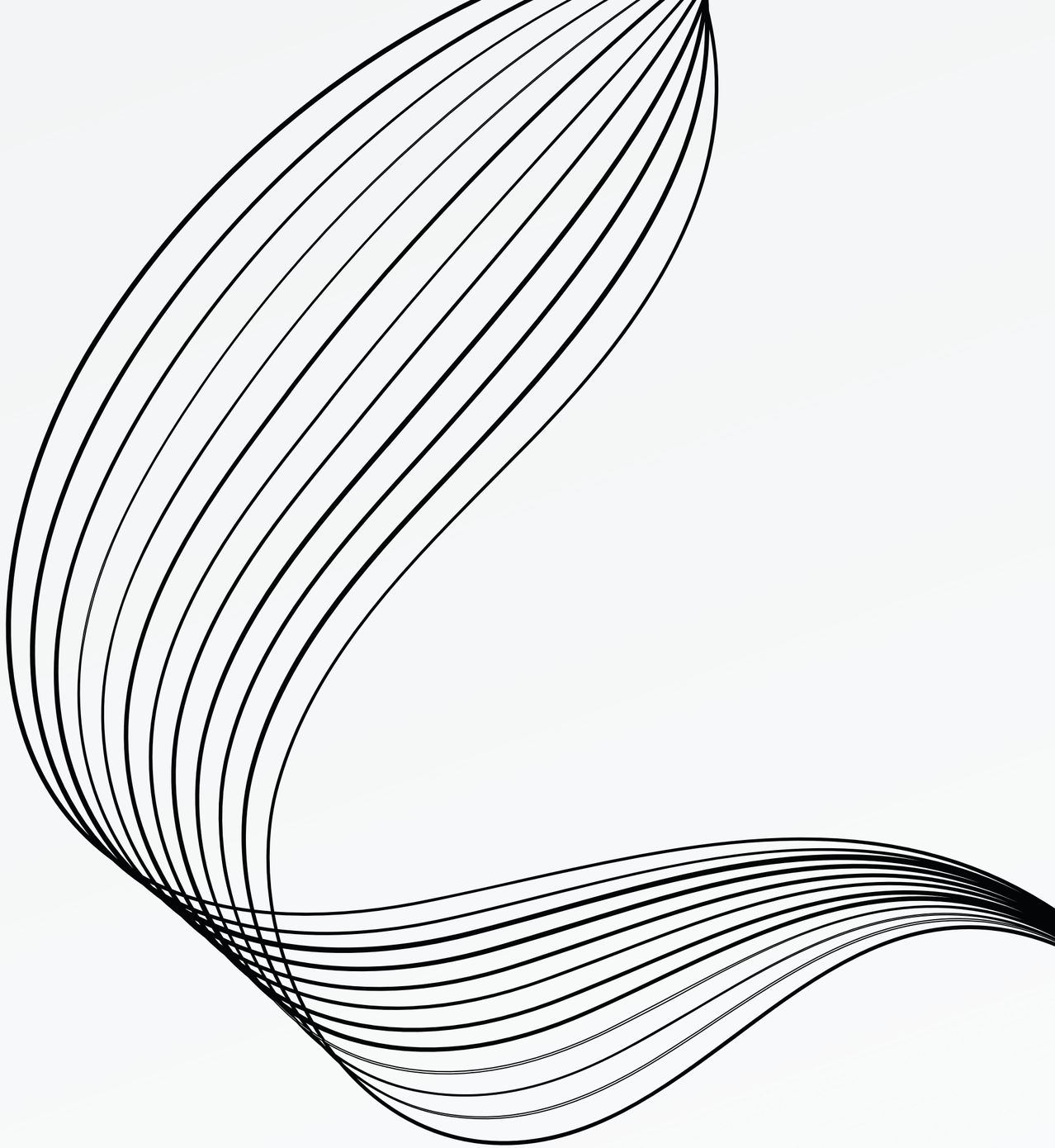
Combining L1 and L2 regularization
MSE: 1,655,855,305,781.80
R2 Score: 0.6150
MAPE: 18.1%

ELASTIC NET REGRESSION

CONCLUSION

Linear and Ridge Regression models performed well, explaining about 65.9% of the variance with a MAPE of 17.4%, indicating reasonable accuracy.

ElasticNet Regression showed lower accuracy (61.5%) and a higher MAPE (18.1%).



PREDICTING COLLEGE ADMISSION OUTCOMES USING MACHINE LEARNING

DATA INTRODUCTION

PROBLEM STATEMENT:

Predicting the chance of a student's admission to a university using various academic and test scores.

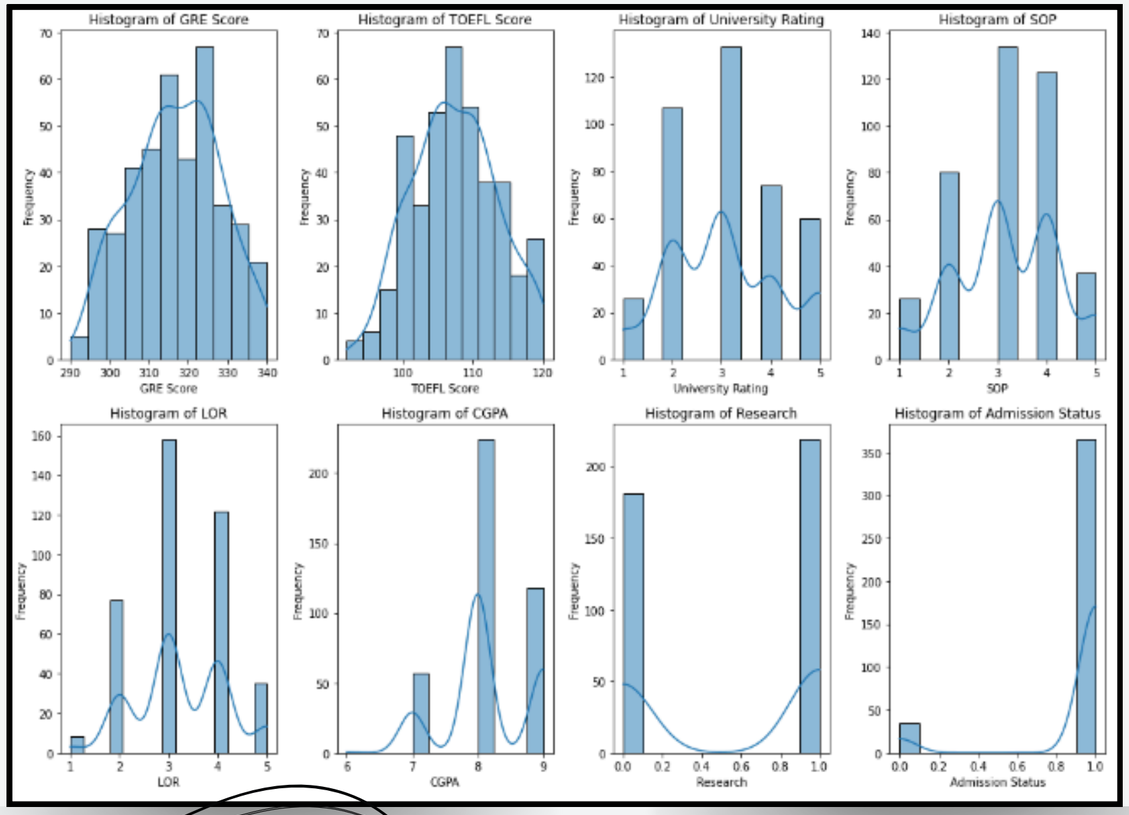
DATA DESCRIPTION

SOURCE: 'ADM_DATA.CSV'
TOTAL RECORDS: 400
FEATURES: 9 ATTRIBUTES INCLUDING GRE SCORE, TOEFL SCORE, UNIVERSITY RATING, SOP, LOR, CGPA, RESEARCH, AND CHANCE OF ADMIT.
ALL FEATURES HAVE APPROPRIATE DATA TYPES (INT64 AND FLOAT64).
NO MISSING VALUES DETECTED.
KEY STATISTICAL INSIGHTS:

- GRE SCORE: AVG - 316.8, STD - 11.5
- TOEFL SCORE: AVG - 107.4, STD - 6.1
- CGPA: AVG - 8.15, STD - 0.65
- RESEARCH EXPERIENCE: 54.75% HAVE RESEARCH EXPERIENCE.

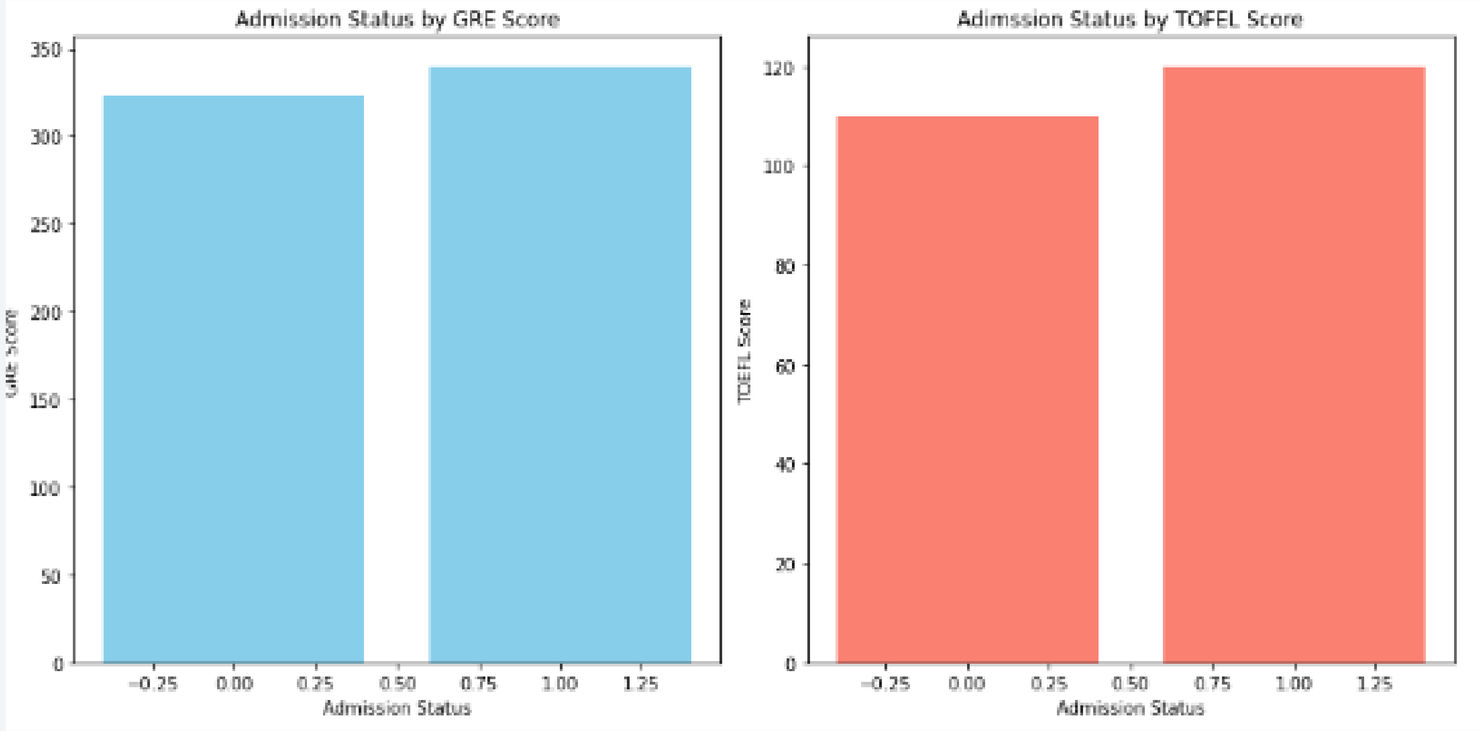
LITERATURE REVIEW

Univariate Analysis



GRE: Over 60 students scored 320 or higher on the GRE.
TOEFL: Over 60 students scored 110 or higher on the TOEFL.
University Rating: The majority of university ratings fall between 3 and 3.5.
SOP: Over 60 students received SOP scores of 4 to 5.
LOR: Over 60 students received LOR scores of 3.
CGPA: Most students have CGPA scores ranging from 8.0 to 9.0.

Bivariate Analysis



The students who has 300 to 350 GRE score are having 50% to 100% chance of getting admission.
The students who has 100 to 120 TOEFL score are having 50% to 100% chance of getting admission.

METHODOLOGY

- Accuracy: 91.25%
- F1-Scores: Class 0 - 36%, Class 1 - 95%
- Confusion Matrix:
- TP: 2, FP: 3, FN: 4, TN: 71

LOGISTIC REGRESSION

- Accuracy: 89%
- F1-Scores: Class 0 - 0%, Class 1 - 94%
- Confusion Matrix:
- TP: 0, FP: 5, FN: 4, TN: 71

RANDOM FOREST

- Accuracy: 89%
- F1-Scores: Class 0 - 31%, Class 1 - 94%
- Confusion Matrix:
- TP: 2, FP: 3, FN: 6, TN: 69

GRADIENT BOOSTING

CONCLUSION

Logistic Regression:

- Best performance with 91.25% accuracy.
- Balanced F1-scores indicating reasonable predictions.

Random Forest:

- Zero F1-score for class 0, indicating poor performance for predicting no admission.

Gradient Boosting:

- Lower performance for class 0 compared to Logistic Regression.

.

Class Imbalance:

- Models struggled with minority class (no admission).

Feature Importance:

- Logistic Regression handled feature importance and class imbalance better.

TWITTER SENTIMENT ANALYSIS

DATA INTRODUCTION

PROBLEM STATEMENT:

Understanding the sentiment behind the tweets

DATA DESCRIPTION

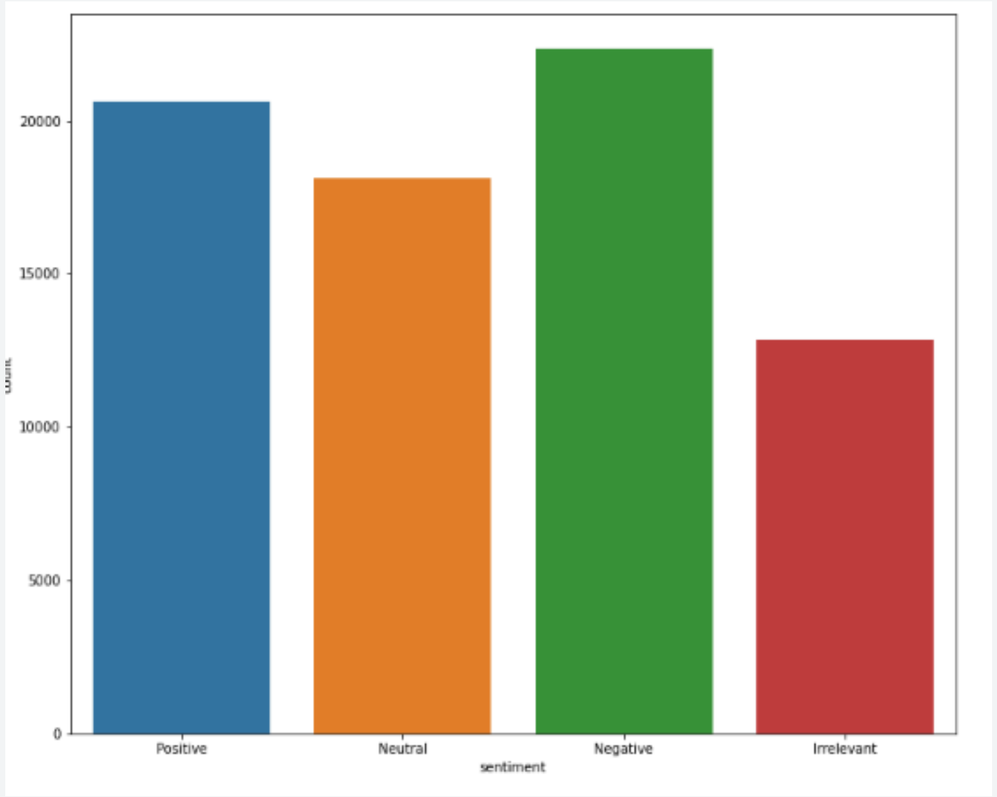
DATASET: TWITTER TRAINING DATA
COLUMNS: ID, TOPIC, SENTIMENT, TEXT
SAMPLE DATA:

- TOPIC: BORDERLANDS
- SENTIMENT: POSITIVE
- TEXT: "I'M GETTING ON BORDERLANDS AND I WILL MURDER YOU..."

DATA INFO:

- TOTAL ROWS: 74,682
- COLUMNS: 3 (AFTER DROPPING 'ID')
- SENTIMENT DISTRIBUTION:
  - NEGATIVE: 22,542
  - POSITIVE: 20,832
  - NEUTRAL: 18,318
  - IRRELEVANT: 12,990

LITERATURE REVIEW



Sentiment Distribution:

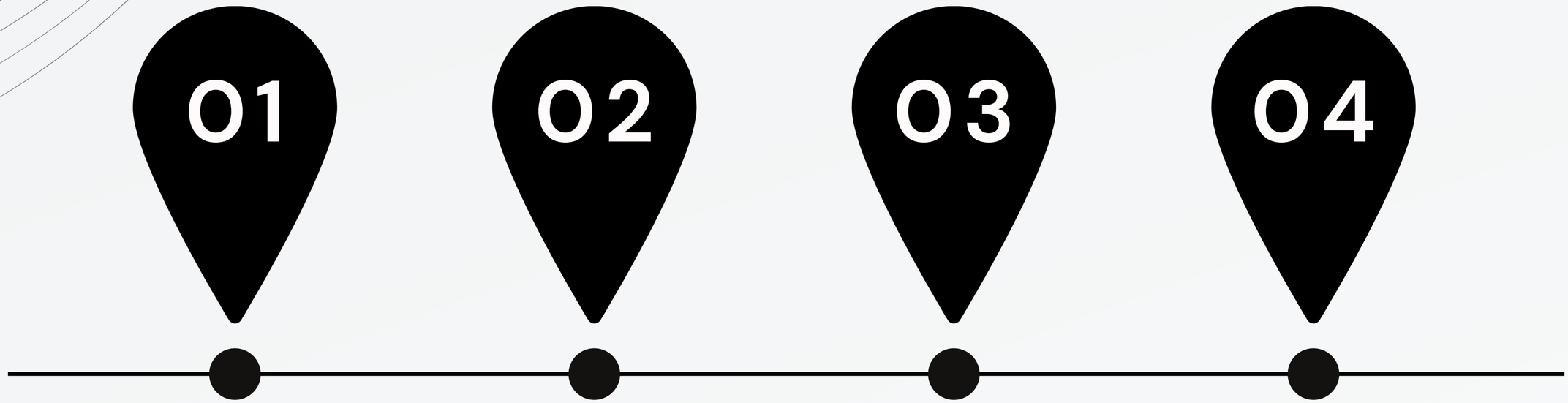
Negative: 22,542

Positive: 20,832

Neutral: 18,318

Irrelevant: 12,990

TEXT PREPROCESSING



METHODOLOGY

Accuracy: 76.31%
F1-Scores:

- Irrelevant: 70%
- Negative: 80%
- Neutral: 74%
- Positive: 77%

LOGISTIC REGRESSION

Accuracy: 90.11%
F1-Scores:

- Irrelevant: 89%
- Negative: 92%
- Neutral: 90%
- Positive: 89%

RANDOM FOREST

Accuracy: 45.16%
F1-Scores:

- Irrelevant: 18%
- Negative: 52%
- Neutral: 39%
- Positive: 42%

ADABOOST CLASSIFIER

CONCLUSION

- **Random Forest:** Achieved the highest accuracy and F1-scores across all sentiment classes
- **Text Preprocessing:** Effective in cleaning and standardizing tweet texts for better model performance
- **AdaBoost:** Performed poorly, particularly in classifying Irrelevant and Neutral sentiments

Why:

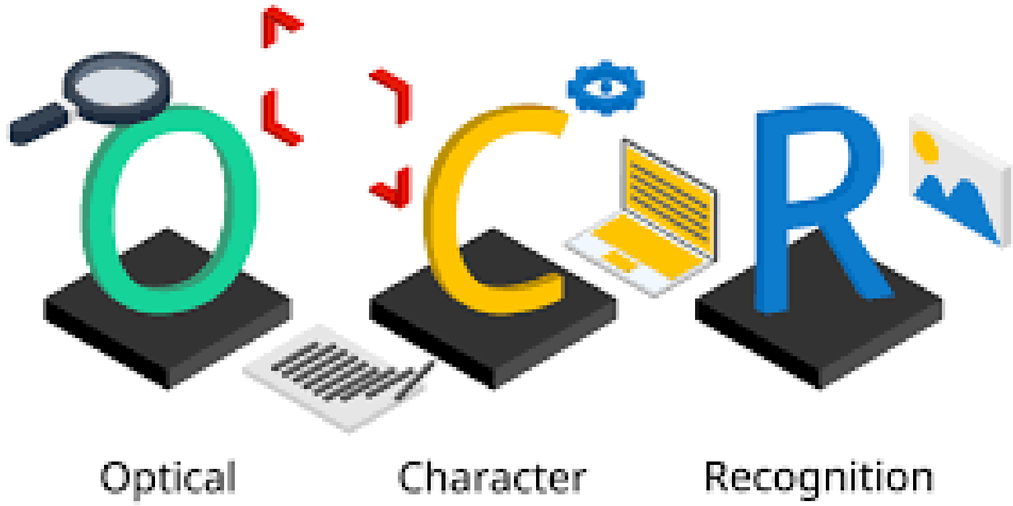
- **Model Suitability:** Random Forest handles large feature spaces well, while AdaBoost struggled due to complexity in text data
- **Data Preprocessing:** Standardized text preprocessing improved model performance, but more advanced techniques could be explored for further improvement

OPTICAL CHARACTER RECOGNITION (OCR): VARIOUS FONTS AND STYLES

DATA INTRODUCTION

PROBLEM STATEMENT:

The goal of this project is to develop an OCR model capable of recognizing characters from a dataset containing various fonts and styles. The model should be able to accurately classify characters in a diverse set of images, providing reliable text recognition for real-world applications.



DATA DESCRIPTION

OVERVIEW

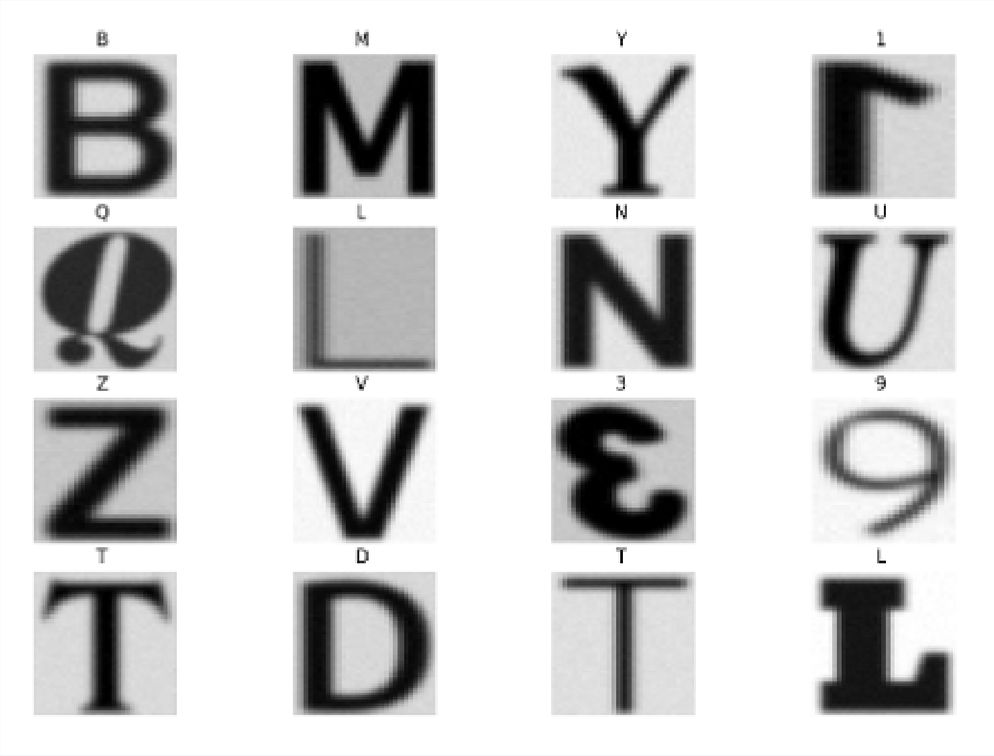
OPTICAL CHARACTER RECOGNITION (OCR) IS A TECHNOLOGY USED TO CONVERT DIFFERENT TYPES OF DOCUMENTS, SUCH AS SCANNED PAPER DOCUMENTS, PDFS, OR IMAGES TAKEN BY A DIGITAL CAMERA, INTO EDITABLE AND SEARCHABLE DATA. THE DATASET USED IN THIS PROJECT CONTAINS IMAGES OF CHARACTERS IN VARIOUS FONTS AND STYLES.

GENERATING TRAIN AND VALIDATION DATA

WE USE IMAGEDATAGENERATOR FROM TENSORFLOW KERAS TO GENERATE TRAINING AND VALIDATION DATA WITH REAL-TIME DATA AUGMENTATION

LITERATURE REVIEW

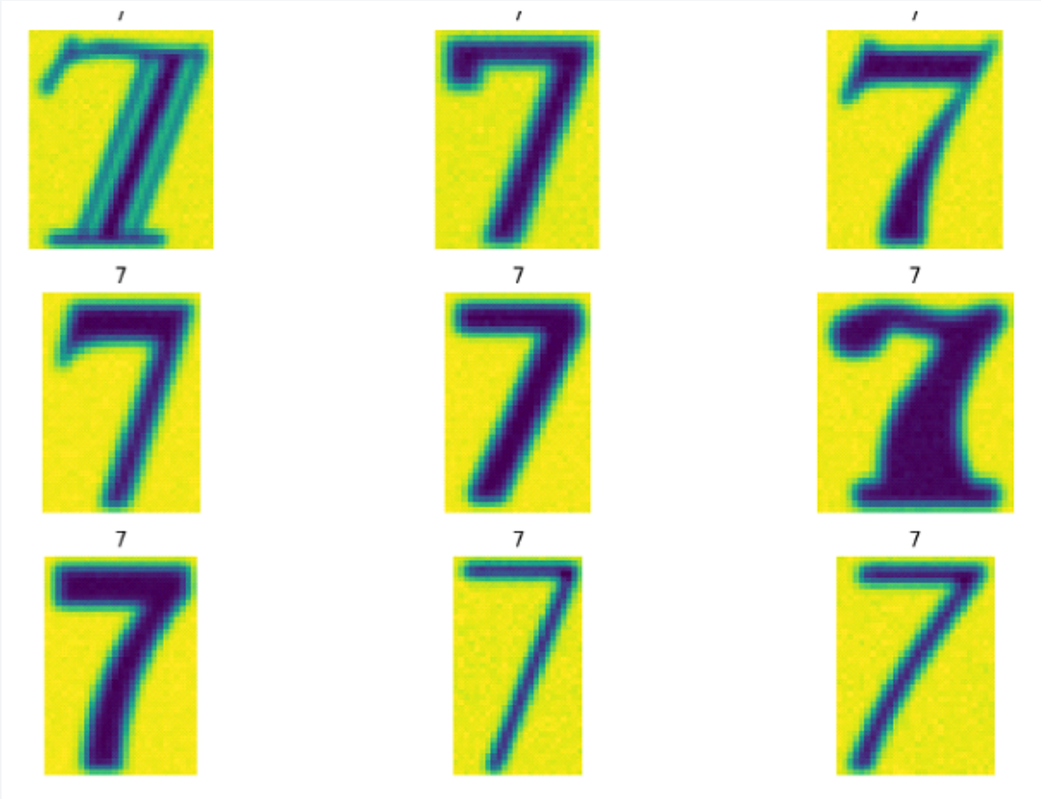
Visualizing Sample Images



Visualizing Label Distribution

All alphabets have the same number of images

Visualizing Sample Images with Labels



METHODOLOGY

The model achieved a validation accuracy of approximately 22%.

SGD OPTIMIZER

The model achieved a validation accuracy of approximately 80%.

ADAM OPTIMIZER

The model's performance with sigmoid activation was comparable to ReLU but slightly less effective.

DIFFERENT ACTIVATION FUNCTIONS (RELU, SIGMOID)

CONCLUSION

- The use of the Adam optimizer with a learning rate of 0.001 provided the best results, achieving a validation accuracy of around 80%.
- Data augmentation techniques such as rescaling, validation split, and horizontal flipping helped improve model performance by providing more varied training data.
- The SGD optimizer with a learning rate of 0.01 did not perform as well as the Adam optimizer.
- The sigmoid activation function, while effective, was not as robust as ReLU for this dataset.

OPTICAL CHARACTER RECOGNITION (OCR): RECEIPT PROCESSING

DATA INTRODUCTION

PROBLEM STATEMENT:

The primary objective of this project is to automate the processing of receipt data to extract meaningful insights for business decision-making.

DATA DESCRIPTION

OVERVIEW

THE DATASET USED IN THIS PROJECT INCLUDES 5000 RECEIPTS PROCESSED TO EXTRACT VALUABLE INFORMATION FOR BUSINESS ANALYTICS. THE DATA IS STORED IN A CSV FILE, WITH EACH ROW REPRESENTING A RECEIPT AND VARIOUS COLUMNS DETAILING DIFFERENT ATTRIBUTES OF EACH TRANSACTION.

MISSING VALUES WERE HANDLED BY IMPUTING WITH APPROPRIATE STRATEGIES (MEAN FOR NUMERICAL VALUES, MODE FOR CATEGORICAL VALUES).

OUTLIERS WERE DETECTED AND MANAGED BY CAPPING OR REMOVING THEM BASED ON THE CONTEXT AND IMPACT ON ANALYSIS.

DATA TYPES WERE VALIDATED AND CORRECTED AS NECESSARY TO ENSURE CONSISTENCY.

LITERATURE REVIEW

Image

Extracted Image

```
Extracted text from images/0.jpg: wal*mart

always low prices.
supercenter
open 24 hours

(515 ) 986 - 1783
st# 5748 op# 00000158 tey 14 tr¥ 03178

waa @ lb 70.49 0.20 n
frap 200010451 f 5.48 n
discount. given .
subtotal aa
cash tab ta
change 12-09
```

METHODOLOGY

OCR techniques achieved an accuracy rate of 95% in extracting key information from receipts.

ACCURACY OF DATA EXTRACTION

The automated process reduced the time required to process receipts by 80%.

EFFICIENCY

The cleaning and validation steps ensured a high level of data quality, with minimal missing values and consistent data types.

DATA QUALITY

CONCLUSION

- Automation of Data Processing: The use of OCR and Python scripts significantly improved efficiency and accuracy in processing receipts.
- Comprehensive EDA: Extensive analysis provided valuable insights into transaction patterns, aiding business decision-making.

- OCR Limitations: Despite a high accuracy rate, OCR struggled with certain receipt formats and handwriting, leading to occasional errors.
- Complex Transactions: Receipts with complex item lists or multiple discounts posed challenges in accurate data extraction and processing.


Q & A

THANK YOU