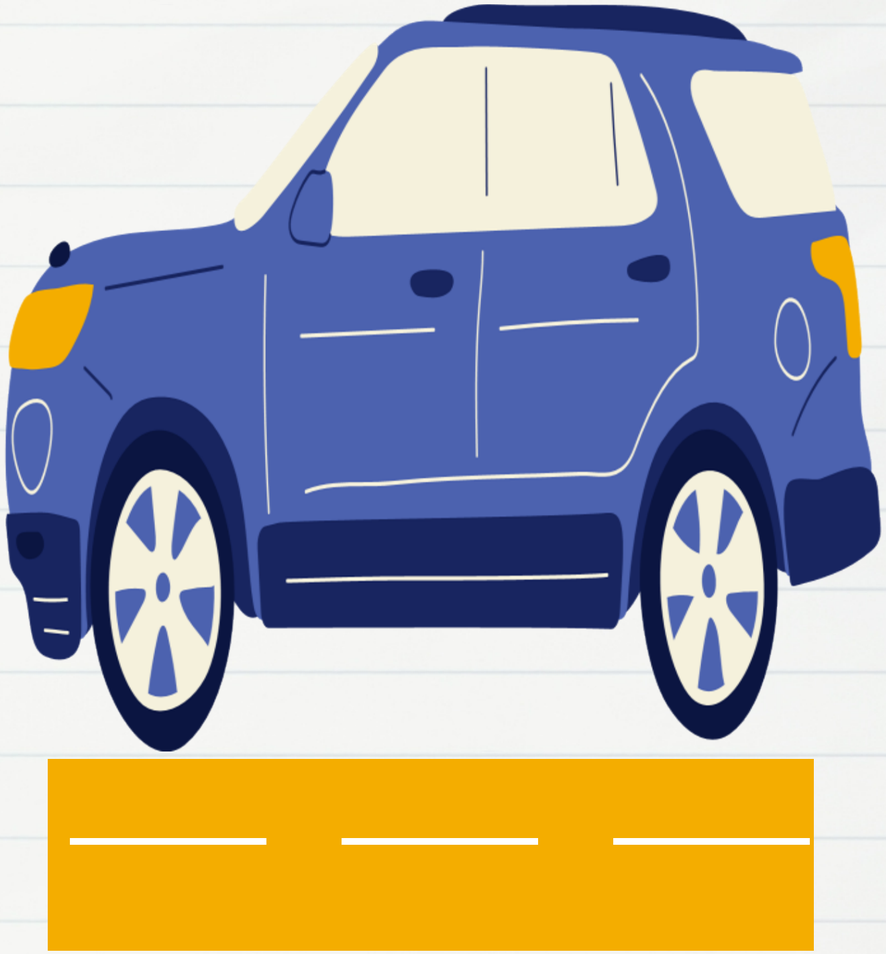
Car Reviews Sentiment Analysis



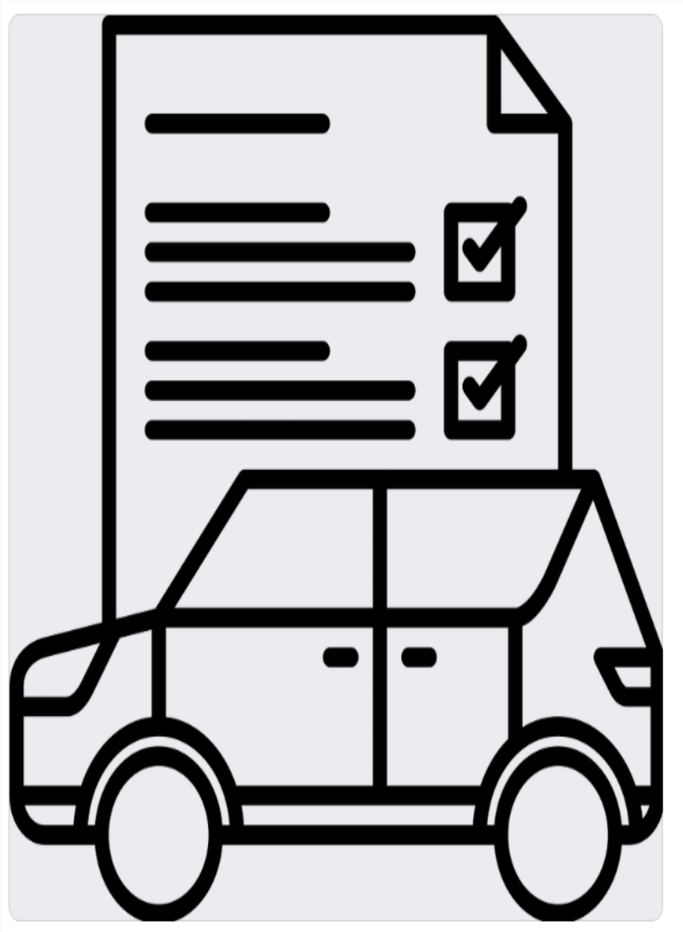
Presented by Team 2:
Rohit Karre,
Navaneeth Gopireddy,
Anish Kolaparthi,
Deepthi Subramanyam

Agenda

- Introduction
- Dataset Description
- Methodology
- Code Review
- Results
- Team Responsibilities
- Conclusion

Introduction

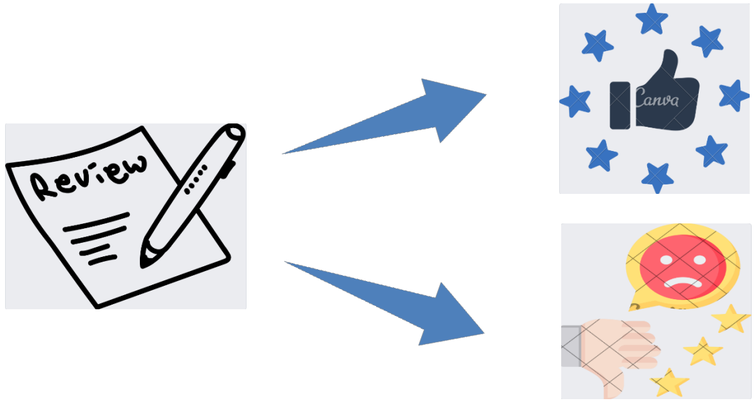
- The automotive industry generates a vast array of reviews spanning websites, forums, and social media.
- Understanding sentiment in these reviews is crucial for manufacturers, dealerships, and consumers.
- This project utilizes NLP and machine learning to analyze sentiment.



Introduction

- Objective:

Categorize reviews into positive or negative sentiments



Dataset Description

- Dataset: Edmunds-Consumer Car Ratings and Reviews (Kaggle repository)

- 284715 records of different brands and models included in the dataset
- 50 unique car brands are present in the data.
- After removing null values, 215386 records are preserved.
- The variables that are present in the dataset are review_date, author_name, vehicle_title, review_title, rating, and review.
- Rating column would be used to generate the target variable
- Additional features would be generated from review column that are ingested in the model as feature set

Methodology

Sentiment Polarity

Count words,
Count Unique words,
Count Letters,
Mean Word Length

Target Sentiment
Rating>2.5 : 1
Rating <2.5 : 0

Casing:
Lower casing the alphabets

Removing Stop words:
Removing words which doesn't contribute to the context of review

Stemming:
Reducing the words to their roots

Polarizing:
Converting the words to their polarized values using TextBlob

Methodology

Models

Supervised Models

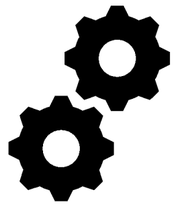
- Logistic Regression
- Decision Tree Classifier
- K-Nearest Neighbors
- Support Vector Machine

Ensembled Models

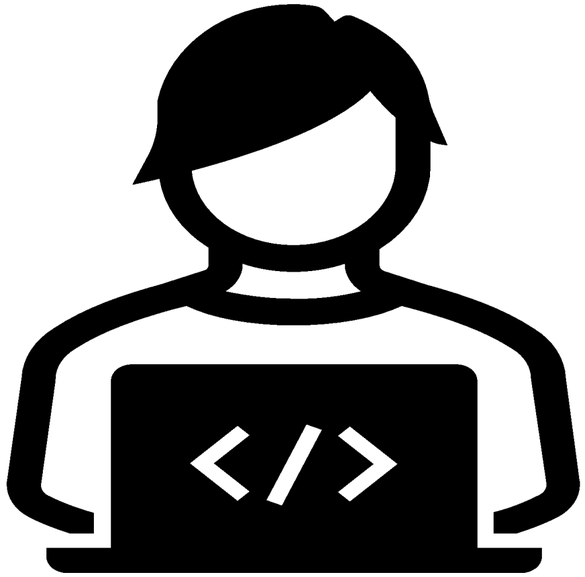
- Random Forest Classifier
- XG Boost
- Light GBM

Neural Networks

- Basic ANN with input, hidden layer with ReLU activation and an output layer with Sigmoid activation

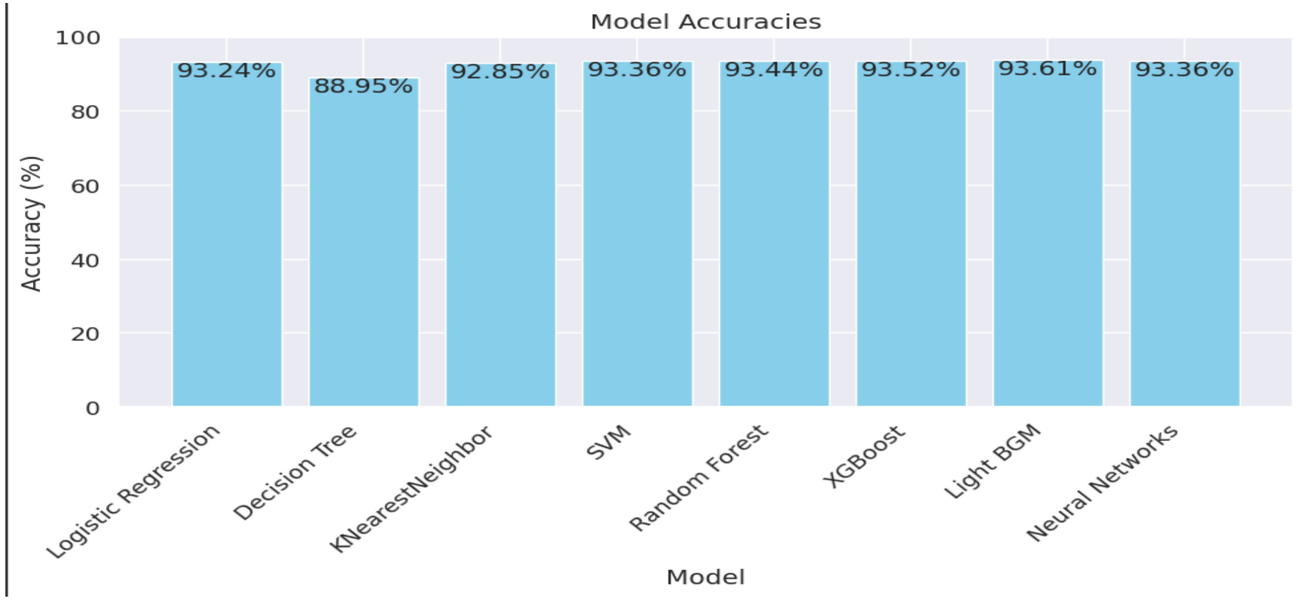


Code Review



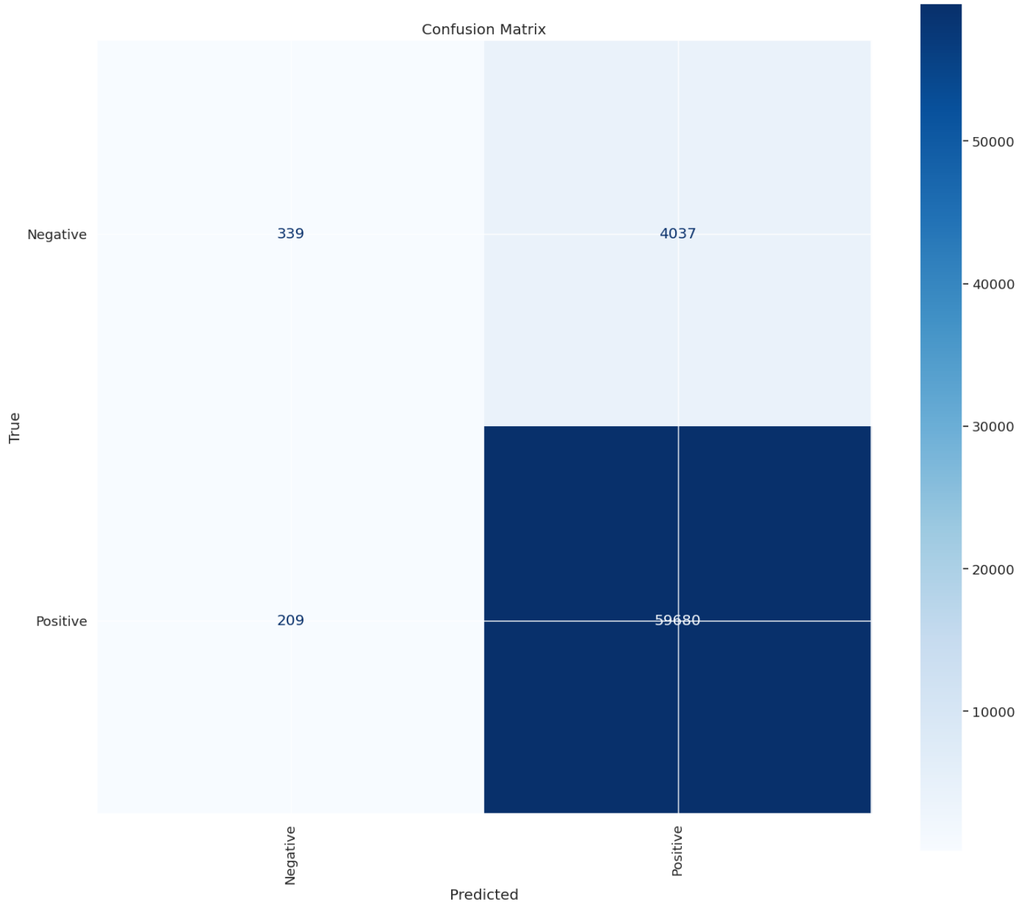
Results

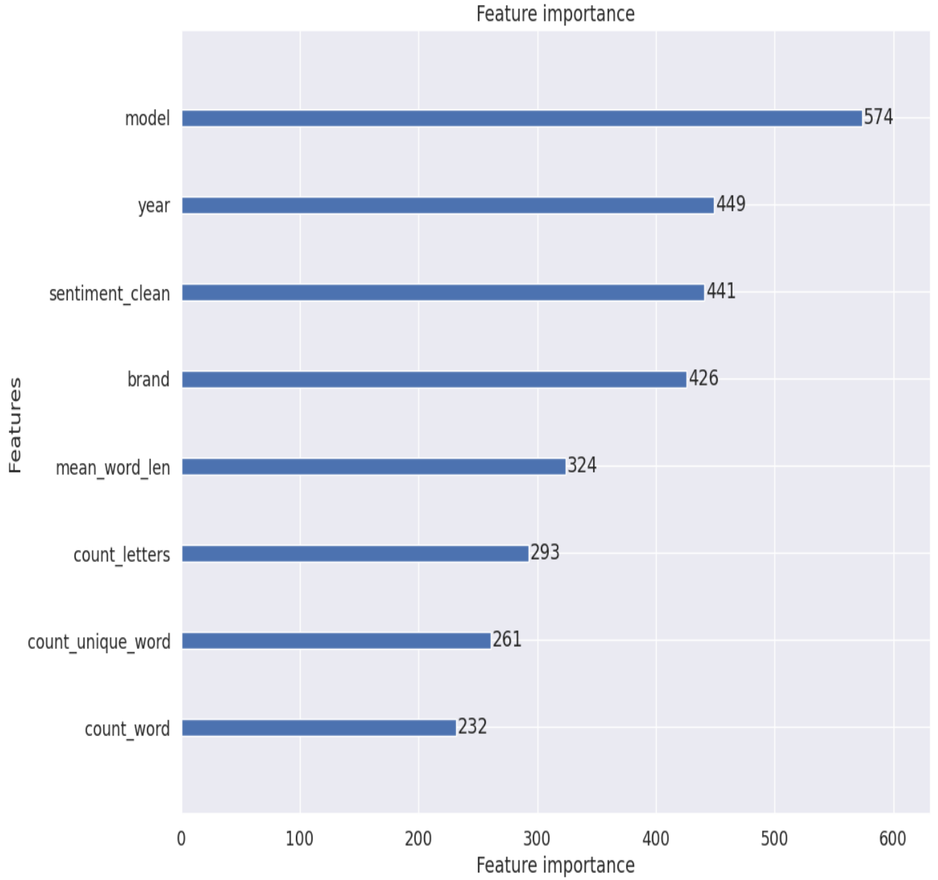
- 30% of the records are used for testing which comprises 65,435 records
- 4,909 Negative and 59,353 Positive records
- Accuracy is used as the metric for Model Evaluation



Results

Finalized Model : Light GBM





Responsibilities

Rohith Karre

- Data Visualization
- Data Preprocessing
- Assistance in EDA
- Assistance in Supervised Learning Models
- Presentation

Navaneeth Gopireddy

- Exploratory Data Analysis
- Data Cleaning
- Supervised Models
- Neural Networks
- Presentation

Anish Kolaparthi

- Correlation Matrix
- Confusion Matrix
- Documentation
- Ensembled Models
- Assistance in Ensembled Models

Deepthi Subramanyam

- Algorithm and Research
- Ensembled Models
- Documentation
- Assistance in Neural Networks

Conclusion

Best Performing Model

- All the models performed similarly well, but Light GBM was the most efficient of them all because of its efficient feature bundling and gradient one sided sampling.

Feature Importance

- For the model Light BGM certain features such as model, year, and polarity value were more influential in predicting sentiments of customer reviews.

Questions?

Thank You