Consumer Complaint Data Analysis and Classification

Presented By:

VAIBHAVI R GAONKAR

Presentation Outline

Consumer Financial Protection Bureau

1. Project Objective
2. Understanding Dataset
3. Exploratory data analysis (EDA)
4. Data Preprocessing
5. Model Building
6. Conclusion

Project Objective

In this presentation, we explore consumer complaints related to student and vehicle loans. Our objectives are to:

- Comprehend the data
- Build different visuals to understand the data from various perspectives
- Develop models to interpret the data
- Derive observations and conclusions

CFPB Consumer Complaint Database

The Consumer Complaint Database is a comprehensive collection of complaints about consumer financial products and services. These complaints are submitted by consumers and sent by the CFPB to companies for response.

Key Features

1. ***Source:*** Consumer Financial Protection Bureau (CFPB)
2. ***Content:*** Complaints about various consumer financial products and services.
3. ***Publication Criteria:*** Complaints are published after the company responds or after 15 days, whichever comes first.
4. ***Update Frequency:*** The database generally updates daily.

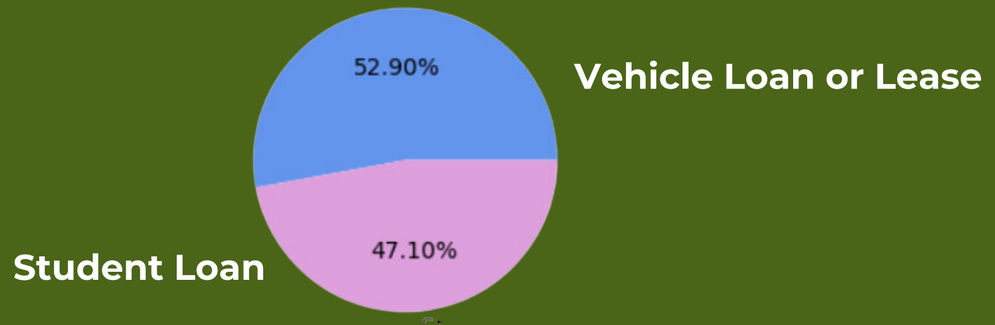
Focus of Our Project

From this database, we are specifically interested in two products:

- Student Loan
- Vehicle Loan or Lease

We extract complaints related to these products to use as our dataset for this project.

60,185 records and 18 columns



Overview of the Features

Date received

Sub-product

Tags

Issue

Sub-issue

Company public response

Company

Consumer consent provided

State

Submitted via

ZIP code

Timely response

Company response to consumer

Date sent to company

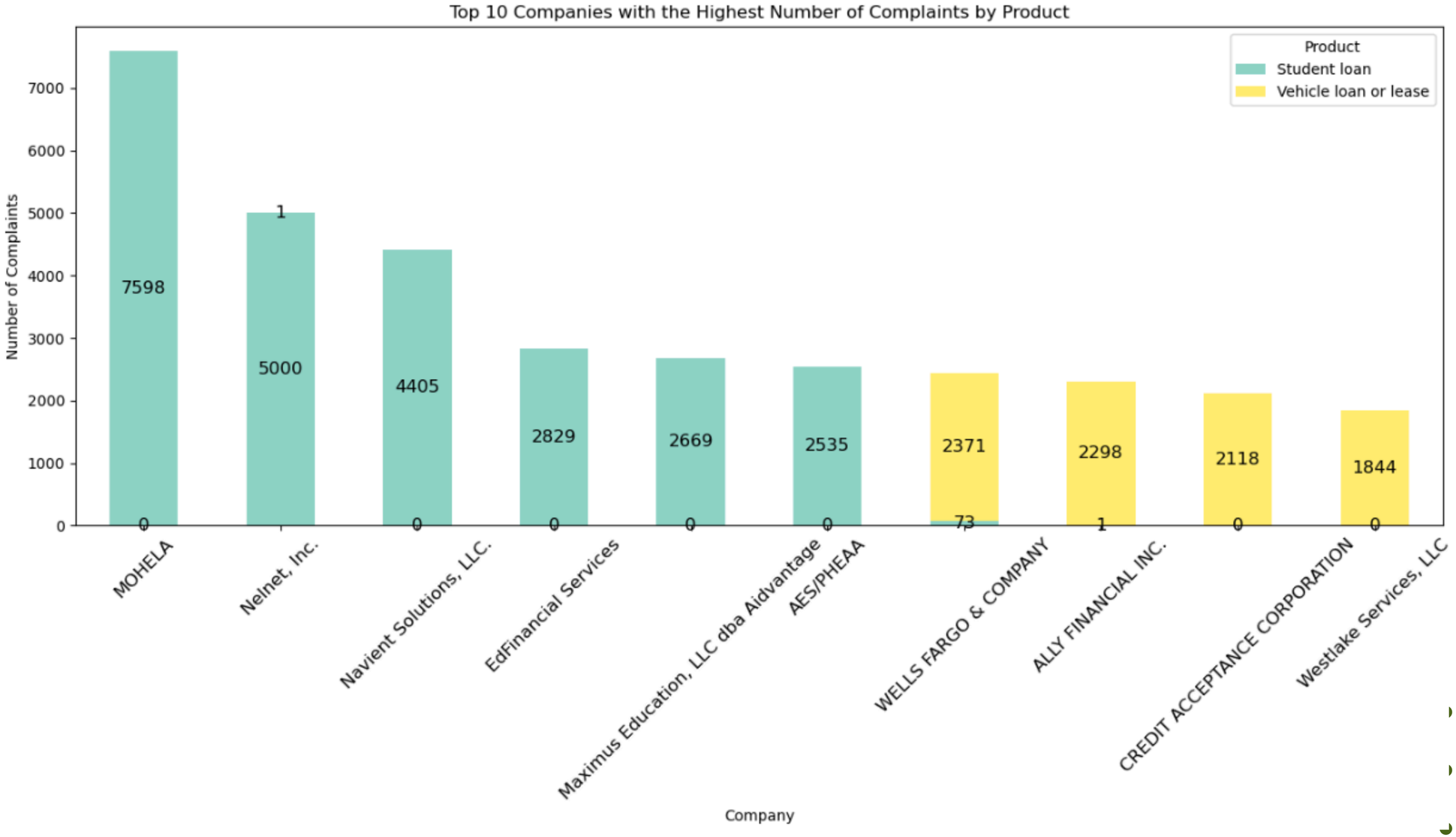
Consumer disputed

Complaint ID

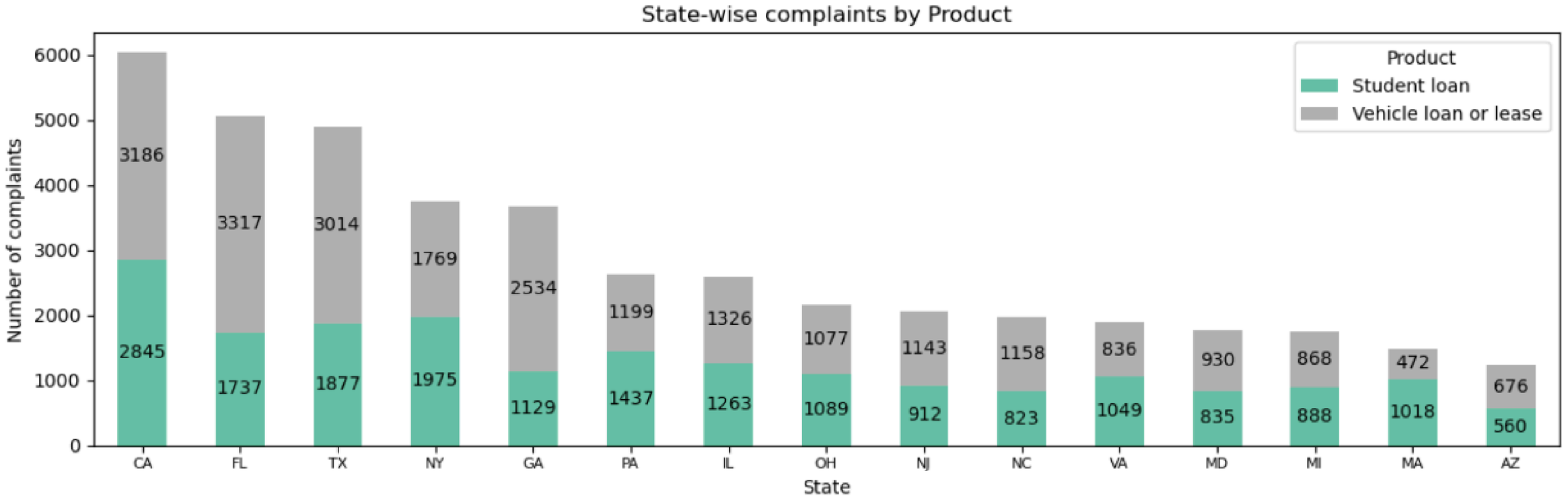
Consumer complaint narrative

Product

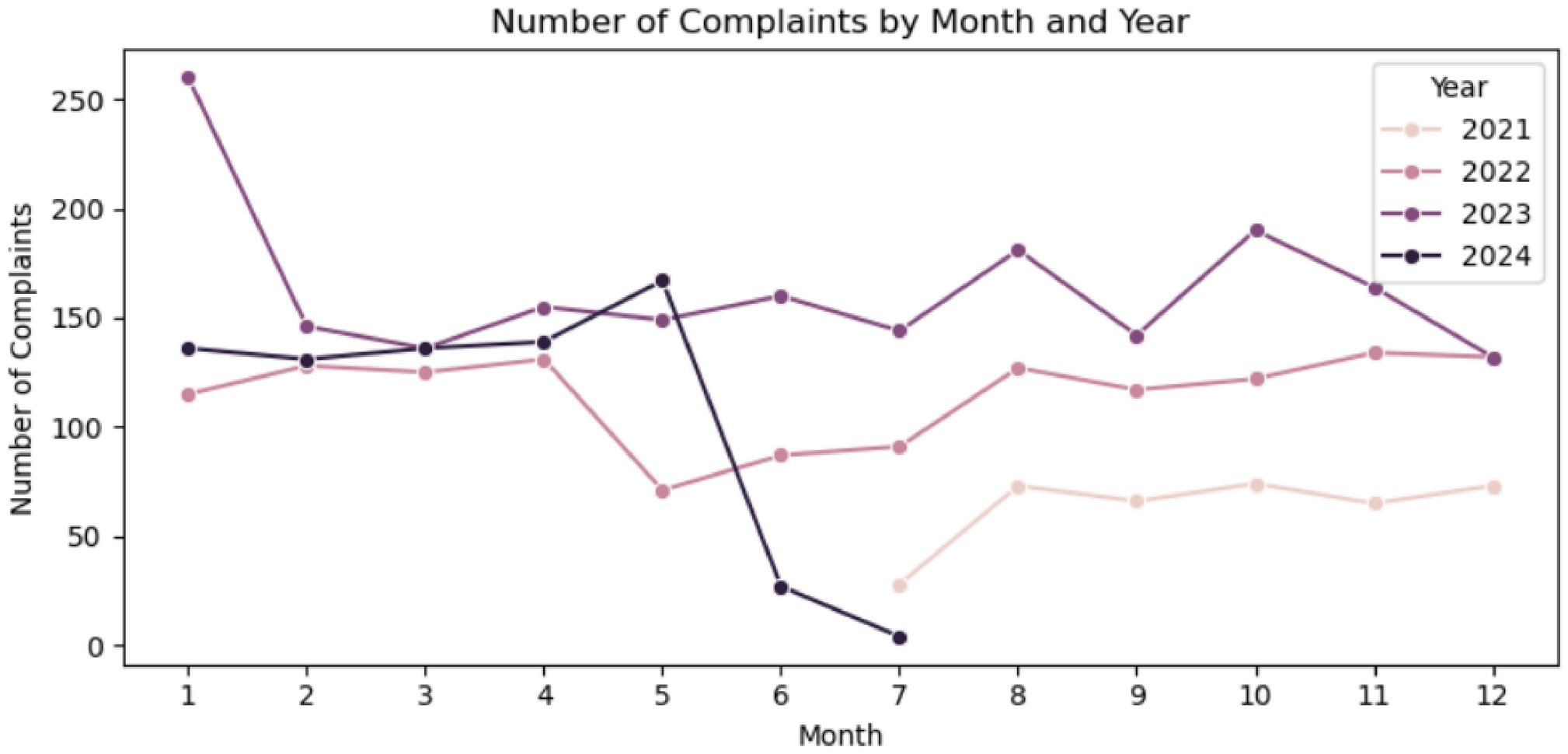
Exploratory data analysis



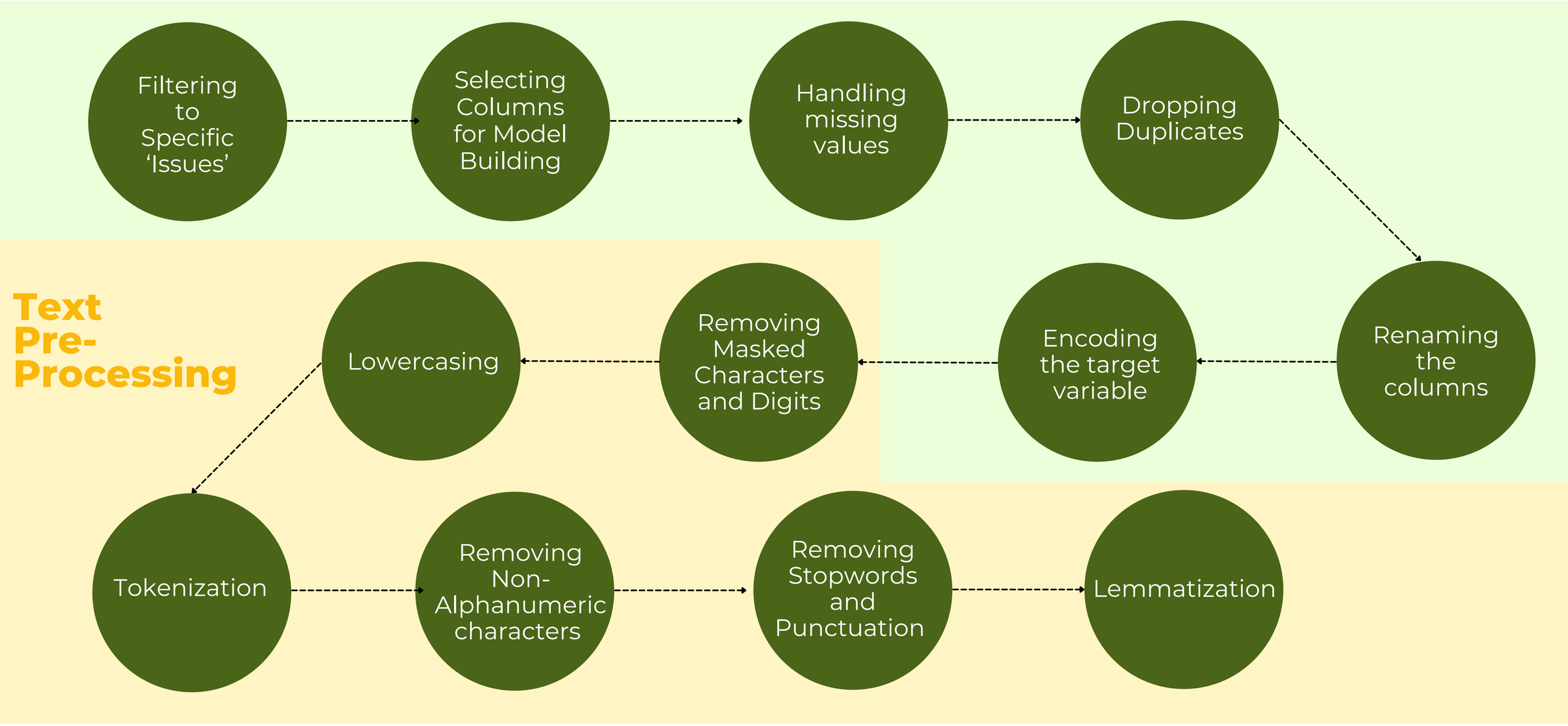
Exploratory data analysis



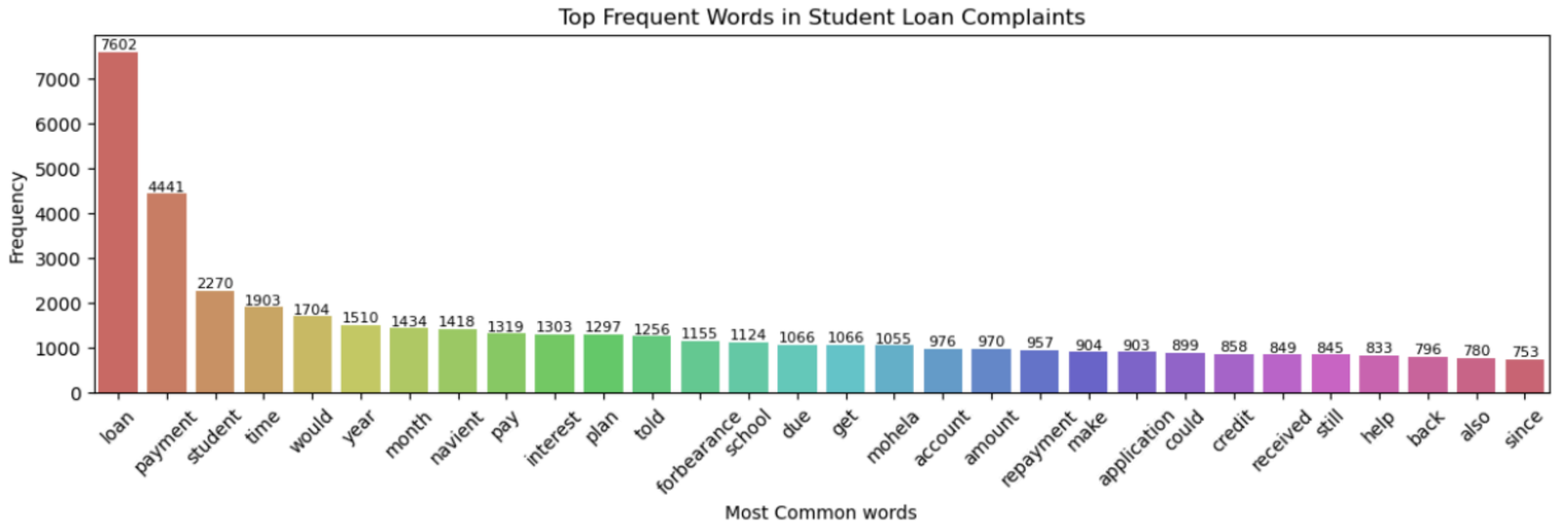
Exploratory data analysis

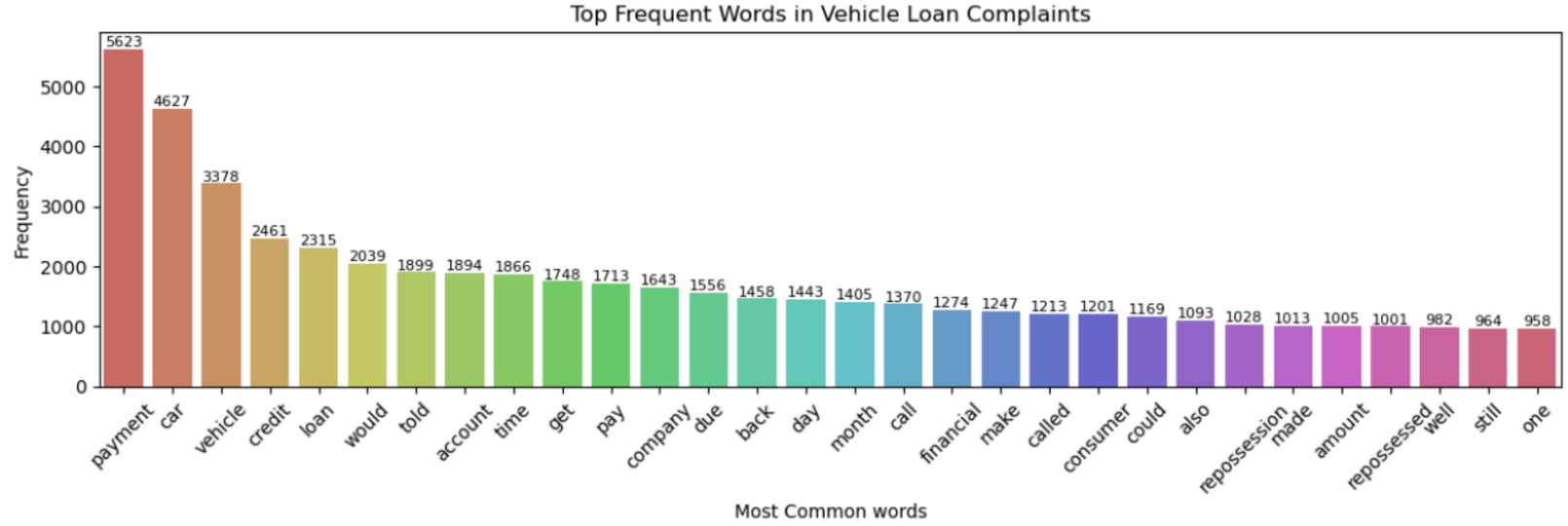


Data Preprocessing



Text data visualisation

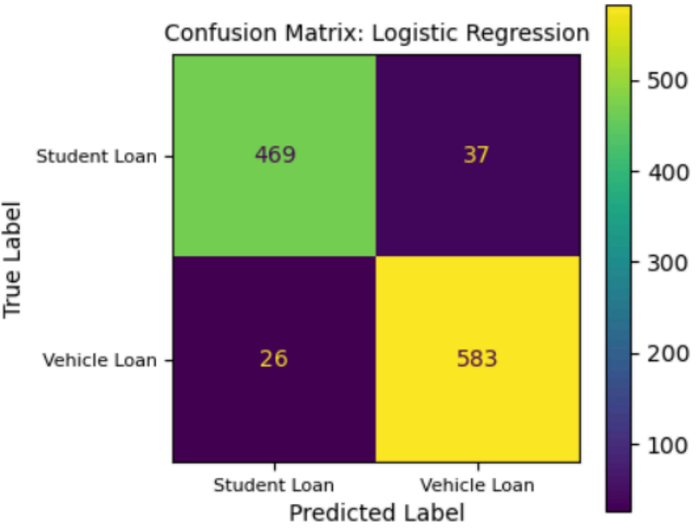




Model Building with Raw features

- Encoding the features: Date,'State','ZIP code'
- Text Vectorization using TFIDF
- Splitting the data

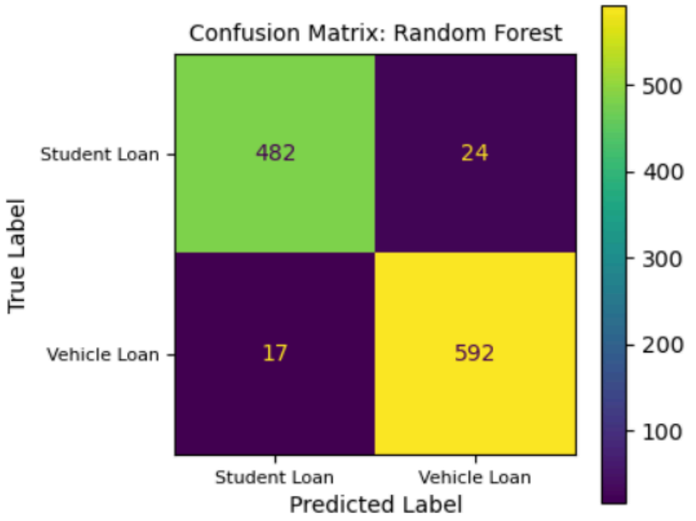
Logistic Regression



- Accuracy: 94.35%
- Precision: 94.03%
- Recall: 96%
- F1-Score: 95%
- AUC: 0.984

Random Forest

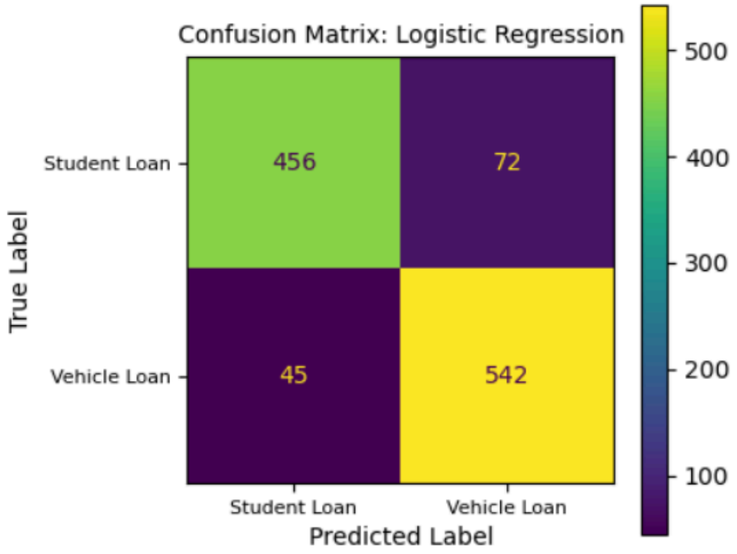
- Accuracy: 96.32%
- Precision: 96.10%
- Recall: 97%
- F1-Score: 97%
- AUC: 0.995



Model Building with engineered features

- Feature Engineering:'Char_count','Word_count','Sent_count','Year','Month','Day','DayOfWeek'
- Correlation map
- Encoding the features: Date,'State','ZIP code'
- Text Vectorization using TFIDF
- Splitting the data

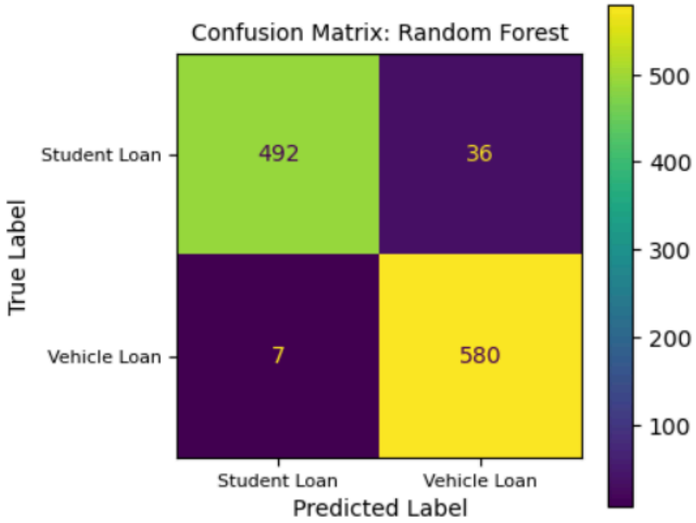
Logistic Regression



- Accuracy: 89.51%
- Precision: 88.27%
- Recall: 92%
- F1-Score: 90%
- AUC: 0.961

Random Forest

- Accuracy: 96.14%
- Precision: 94.15%
- Recall: 99%
- F1-Score: 96%
- AUC: 0.994



Conclusion

- Feature engineering had mixed effects on model performance:
  - Logistic Regression: Experienced a decline in performance metrics with feature engineering.
  - Random Forest: Maintained high performance with minimal variations.
- Overall Performance:
  - Random Forest consistently outperformed Logistic Regression.
  - Random Forest is the superior model for this classification task.

	Model	Accuracy	Precision	Recall	F1 Score	AUC
0	Logistic Regression (No FE)	94.35	94.03	96.23	94.90	0.9844
1	Random Forest (No FE)	96.32	96.10	97.04	96.57	0.9953
2	Logistic Regression (FE)	89.51	88.27	92.31	90.25	0.9616
3	Random Forest (FE)	96.14	94.15	98.81	96.33	0.9940

THANK YOU