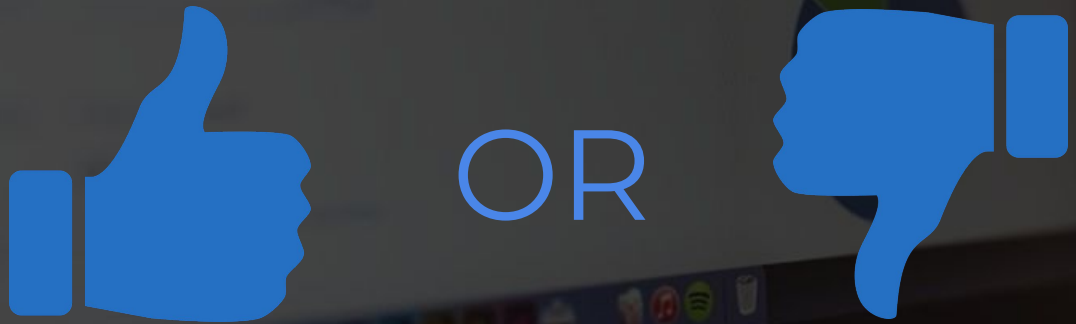
RUBIX

PRERIT JAIN

TARUN BHATTAR

PROBLEM STATEMENT

- SENTIMENT ANALYSIS OF EMPLOYEES REVIEWS FROM “GLASSDOOR.COM” USING MACHINE LEARNING APPROACH.

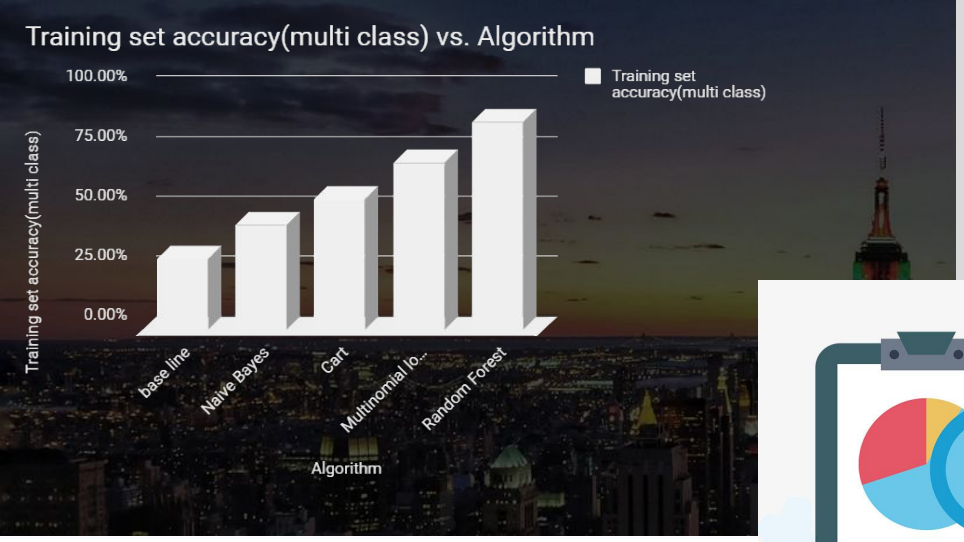


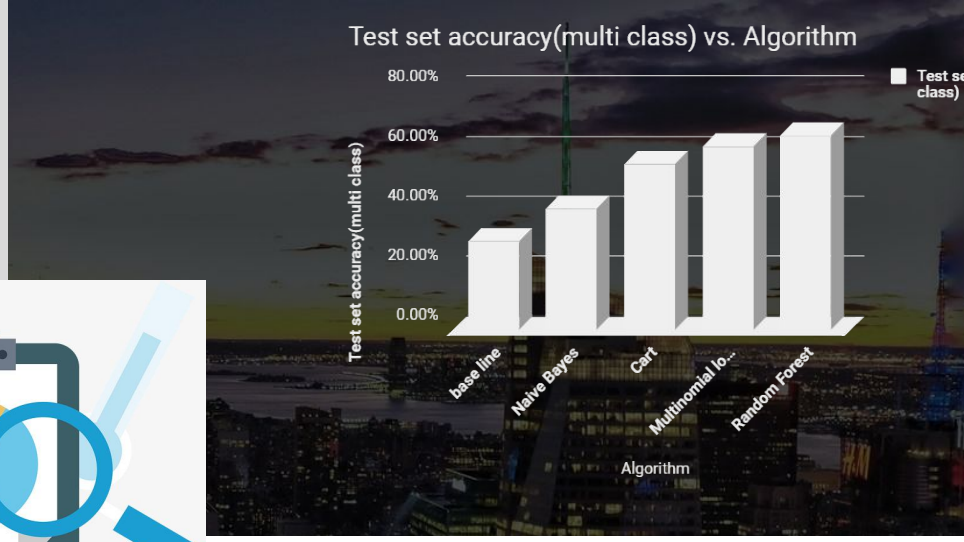
CHALLENGES

- DATA EXTRACTION
  - Scraping Glassdoor website using a google chrome extension.
- TRAINING DATASET CREATION
  - Manual Approach
  - Dictionary-Based Approach
- DECIDING APPROPRIATE MODELS
  - Trying Different Machine Learning Models.
  - Checking Accuracy.

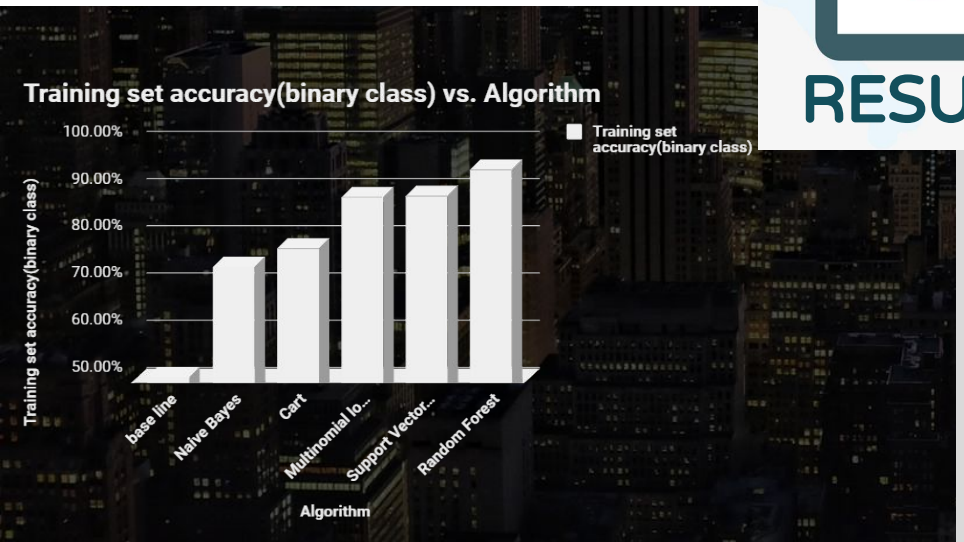
APPROACH

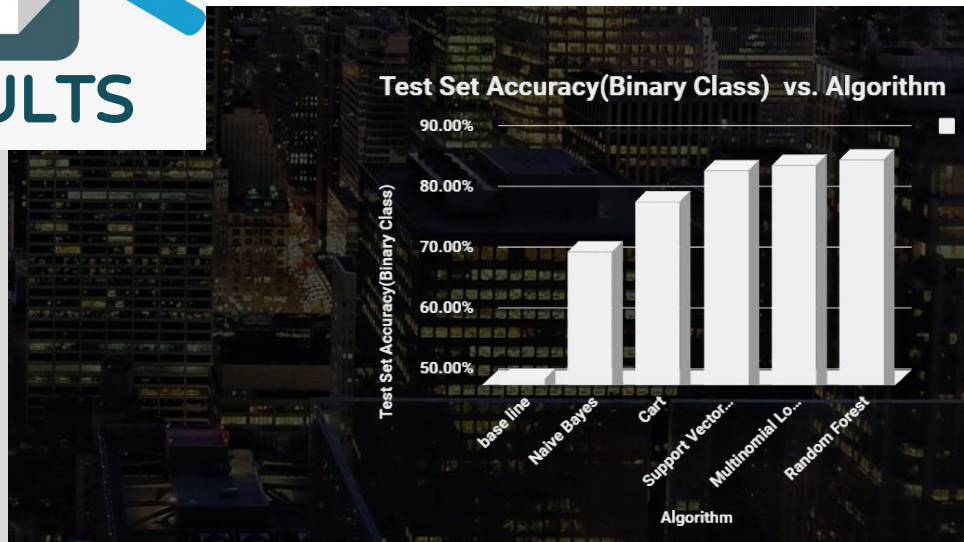
- SPLITTING THE DATASET.
- CREATION OF CORPUS.
- PREPROCESSING STEPS:
  - Removing Stop Words.
  - Removing Punctuations.
  - Stemming.
- CREATION OF DOCUMENT TERM MATRIX
- REMOVING SPARSITY
- MACHINE LEARNING MODELS APPLICATION.
  - Classification Trees.
  - Random Forests.
  - Naive Bayes Algorithm.
  - Multinomial Logistic Regression.
  - Support Vector Machines
- APPLYING ON TEST SET AND CALCULATING ACCURACY.


Training set accuracy(multi class) vs. Algorithm
Training set accuracy(multi class)
100.00%
75.00%
50.00%
25.00%
0.00%
Training set accuracy(multi class)
base line
Naive Bayes
Cart
Multinomial lo...
Random Forest
Algorithm


Test set accuracy(multi class) vs. Algorithm
80.00%
60.00%
40.00%
20.00%
0.00%
Test set accuracy(multi class)
base line
Naive Bayes
Cart
Multinomial lo...
Random Forest
Algorithm

RESULTS


Training set accuracy(binary class) vs. Algorithm
Training set accuracy(binary class)
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
Training set accuracy(binary class)
base line
Naive Bayes
Cart
Multinomial lo...
Support Vector...
Random Forest
Algorithm


Test Set Accuracy(Binary Class) vs. Algorithm
90.00%
80.00%
70.00%
60.00%
50.00%
Test Set Accuracy(Binary Class)
base line
Naive Bayes
Cart
Support Vector...
Multinomial Lo...
Random Forest
Algorithm

OVERALL RATING OF A COMPANY

BENCHMARK SCORE 1 =3.3

BENCHMARK SCORE 2 =3.9

OVERALL RATING =5*POS/(POS + NEG)

POS = No. of Reviews classified as positive
NEG= No. of Reviews classified as negative

- Overall rating > Benchmark score 2 — GOOD
- Benchmark score 1 < Overall rating < Benchmark score 2 — AVERAGE
- Overall rating < benchmark score 1 — BAD

LIMITATIONS

- DATASET SIZE
- SARCASM HANDLING:
- STATEMENTS HAVING MULTIPLE SENTIMENTS.
- LIMITED DICTIONARY
- MULTICLASS CLASSIFICATION

HOW OUR WORK IS RELATED TO RUBIX ?

SENTIMENT ANALYSIS	OVERALL RATING OF A COMPANY	AN INDEPENDENT VARIABLE FOR CREDIT RATING
Understanding Customer/Employer Perspective for a company through their sentiment expressed in reviews.	Overall rating by aggregating the individual review ratings.	Key Independent variable for analysing the behaviour of a company towards loan/credit.