Predicting News Popularity with Supervised Machine Learning

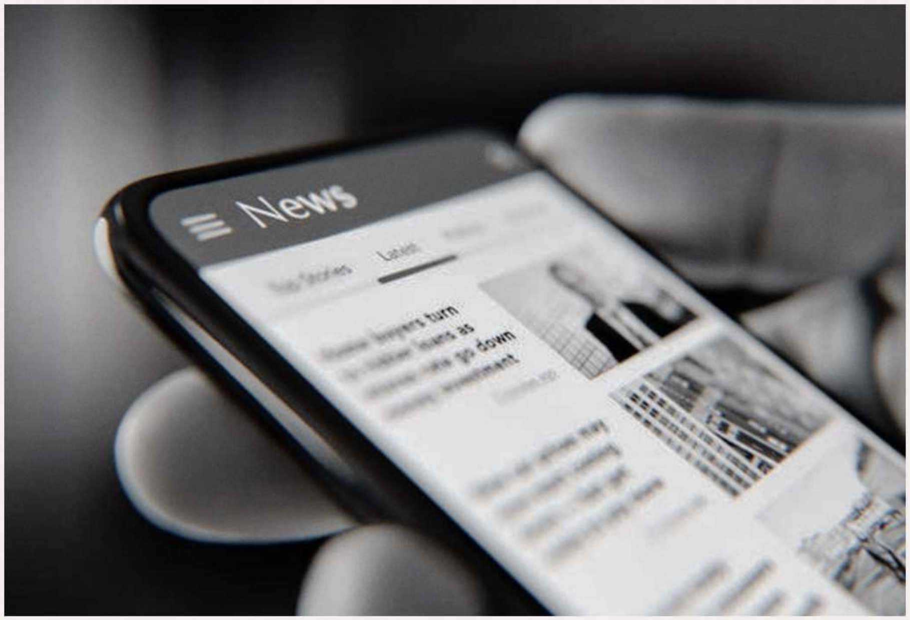
Monday 12/09/2024

Team 2 Section B

Contributors: Achinthya Sreedhar, Neha Shastri, Chaitali Deshmukh, Aryan Sehgal

09.03.2020 - 15.03.2020
Week 11
SCI DAY
Nº157
BREAKING NEWS
Theoretical mechanics makes a breakthrough! This is unbelievable.
TEXT 1
TEXT 2
TEXT 3
TEXT 4
— Arthur Baker, special correspondent
45%

Problem Statement



Problem: Accurately predicting the success of newly published news articles by forecasting their popularity (clicks/impressions ratio)

Stakeholders: News platforms and publishers

Why: To optimize content selection, promotion strategies, and personalized recommendations, maximizing reader engagement and platform success

Data Source

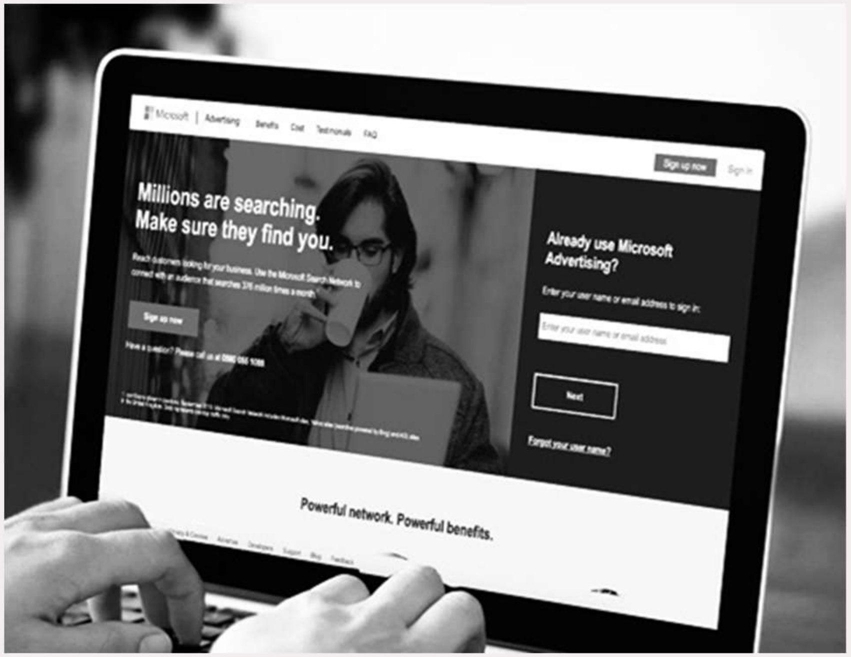
Microsoft News Dataset (MIND)

Behaviors:

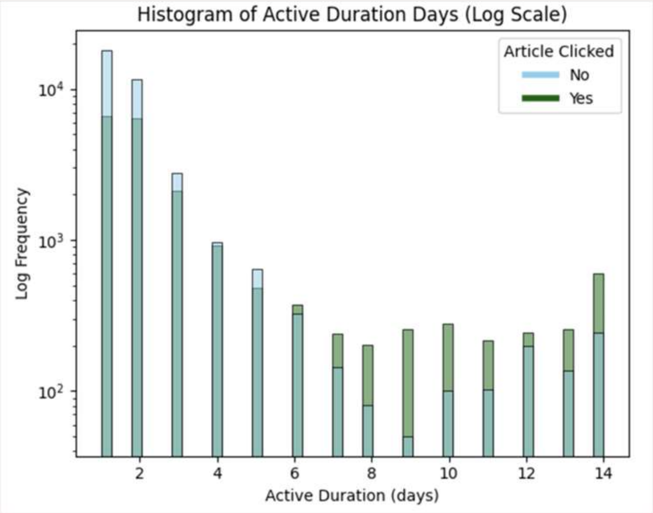
- Contains click histories and impression logs of users
- Used to analyze user interaction patterns and define the target variable, click percentage

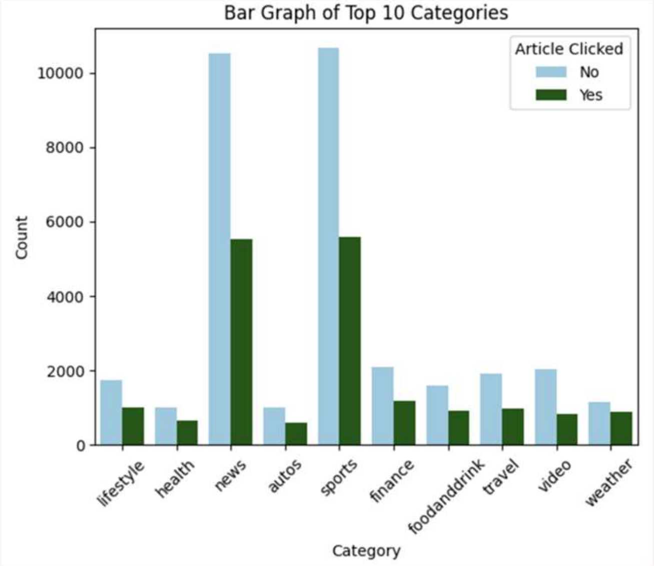
News:

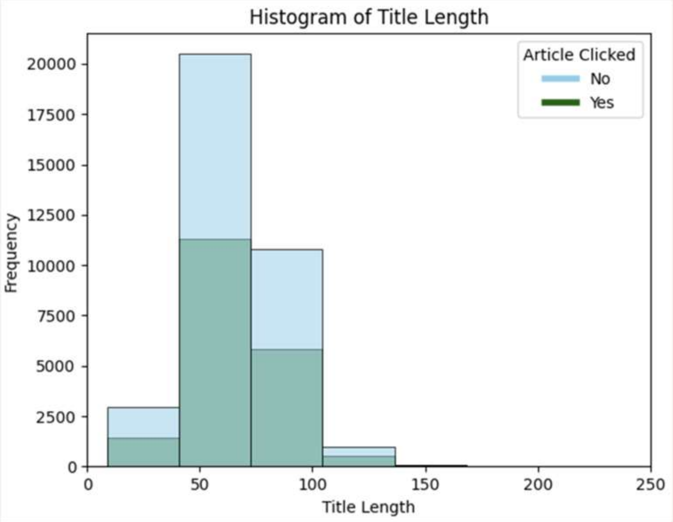
- Contains information of news articles
- Used to extract features such as title length and category for predictive modeling



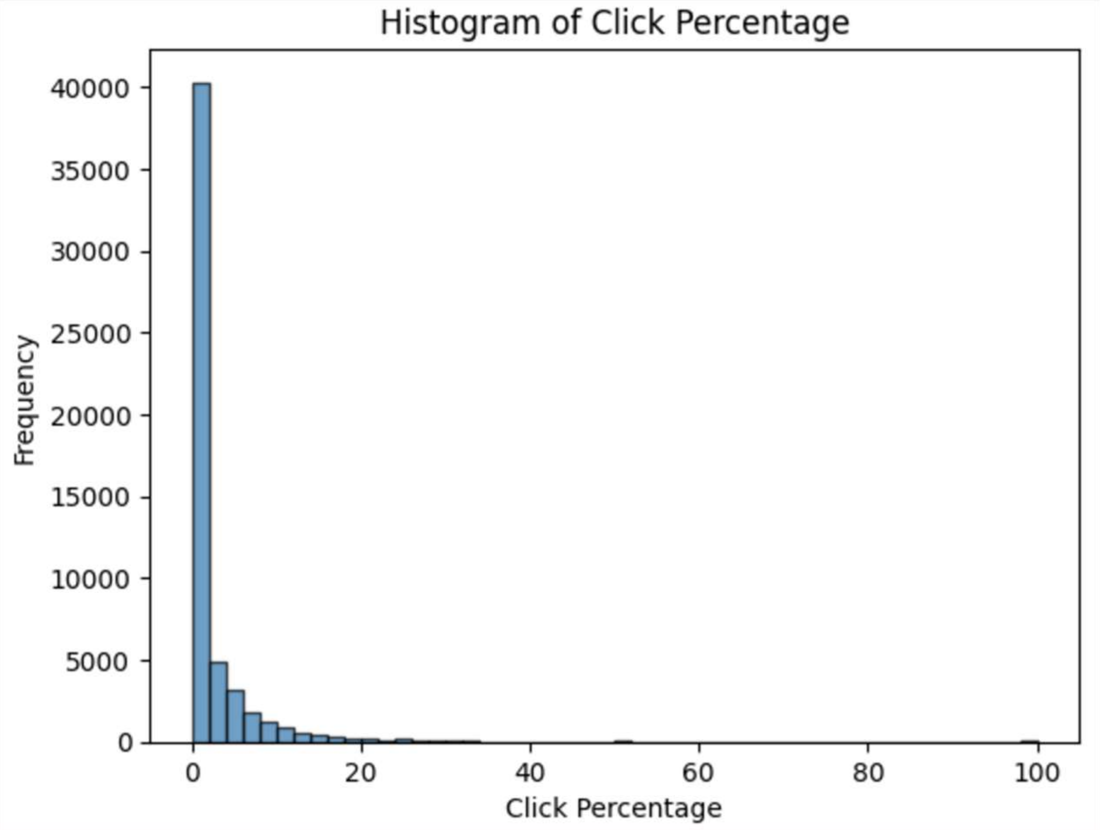
Descriptive Analysis and Insights







Distribution of Target Variable

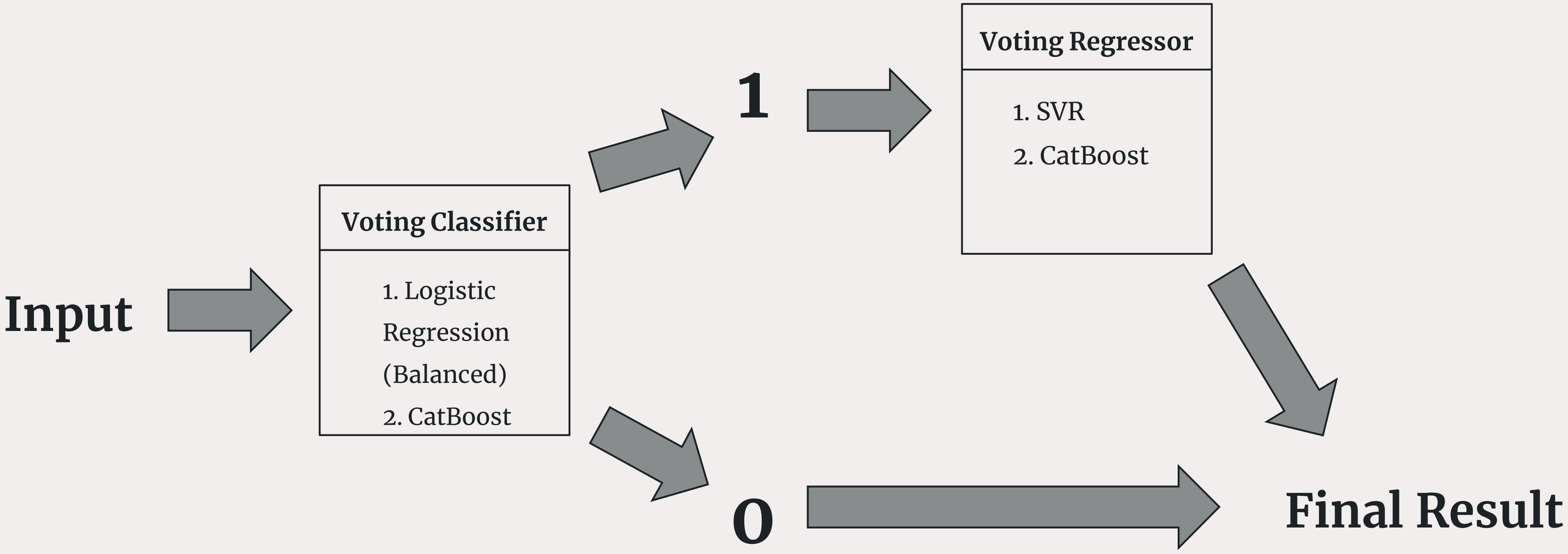


Most click percentage values are 0

Indicates the dataset is:

- Highly imbalanced
- Positively skewed

Machine Learning Methods Applied



Input

Voting Classifier

1. Logistic Regression (Balanced)
2. CatBoost

1

Voting Regressor

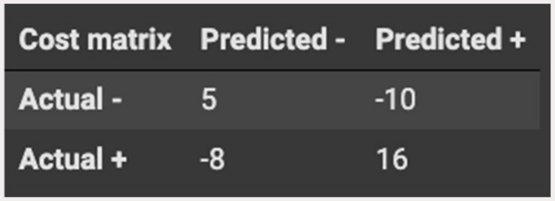
1. SVR
2. CatBoost

0

Final Result

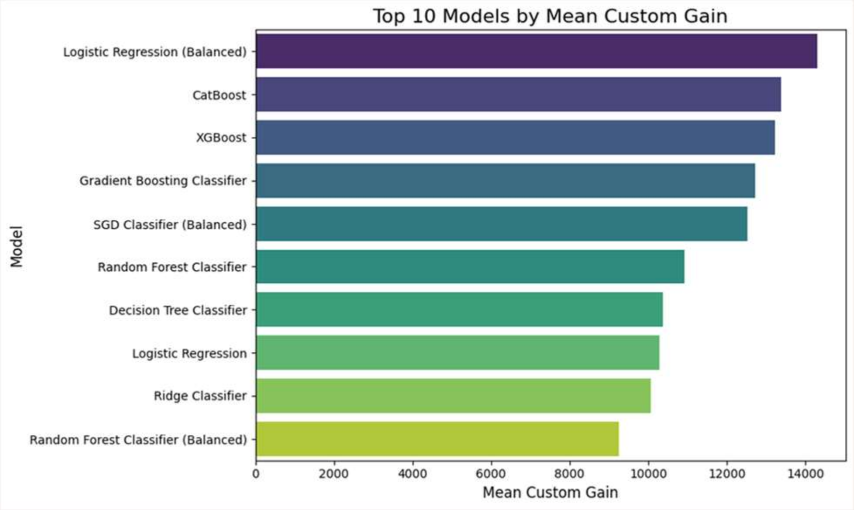
Classification Model

1. Cost Function

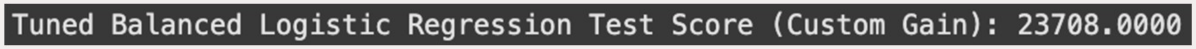


Cost matrix	Predicted -	Predicted +
Actual -	5	-10
Actual +	-8	16

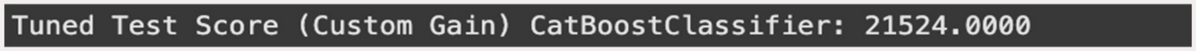
2. Top 10 Models During Cross Validation



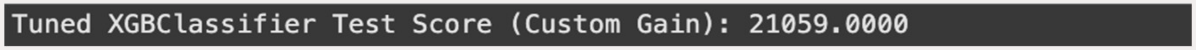
3. Value of Top 3 Models After Tuning



```
Tuned Balanced Logistic Regression Test Score (Custom Gain): 23708.0000
```

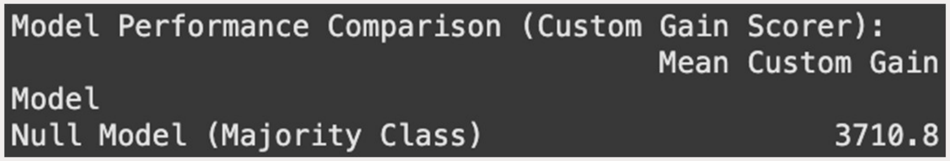


```
Tuned Test Score (Custom Gain) CatBoostClassifier: 21524.0000
```

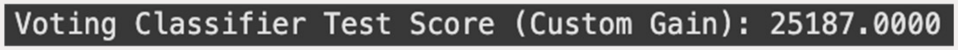


```
Tuned XGBClassifier Test Score (Custom Gain): 21059.0000
```

4. Voting Model Performance Improvement Over Null Model



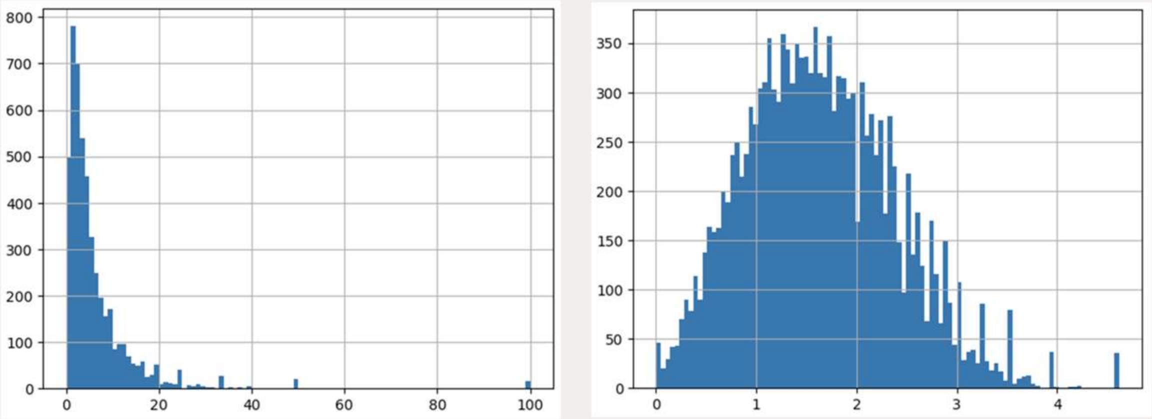
```
Model Performance Comparison (Custom Gain Scorer):
                                          Mean Custom Gain
Model
Null Model (Majority Class)                         3710.8
```



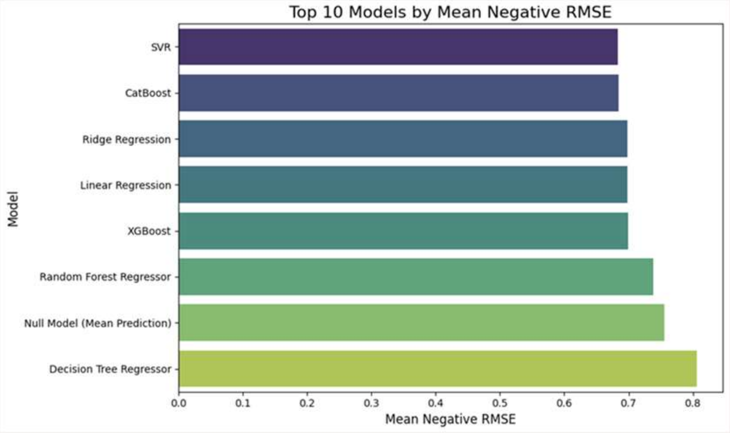
```
Voting Classifier Test Score (Custom Gain): 25187.0000
```

Regression Model

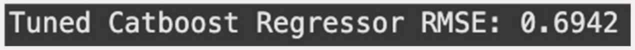
1. Histogram Before and After Applying Log



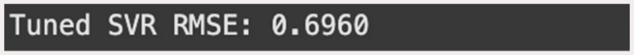
2. Top 10 Models During Cross Validation



3. RMSE of Top 2 Models After Tuning

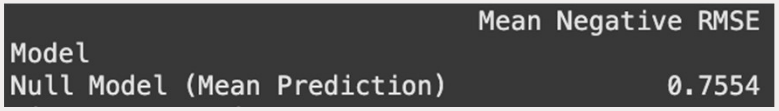


```
Tuned Catboost Regressor RMSE: 0.6942
```

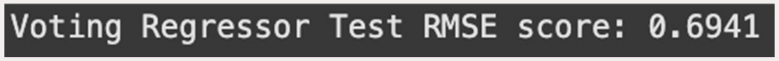


```
Tuned SVR RMSE: 0.6960
```

4. Voting Model RMSE Over Null Model

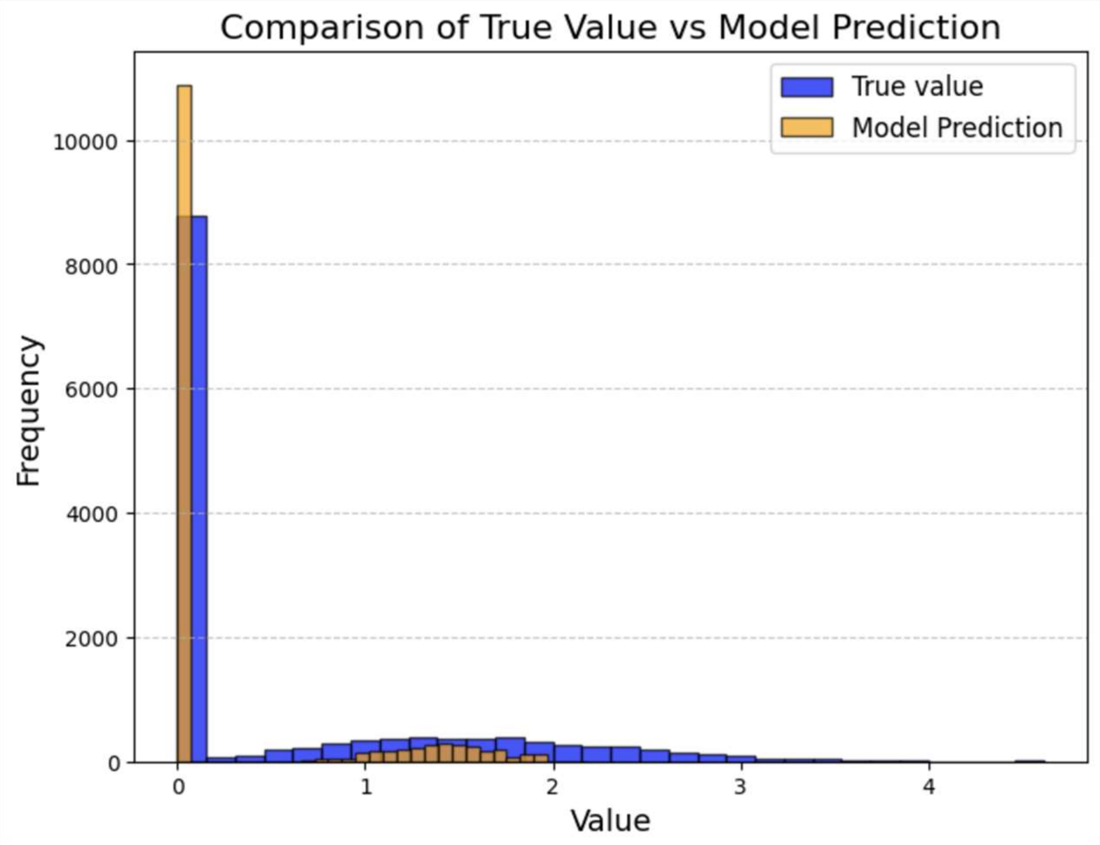


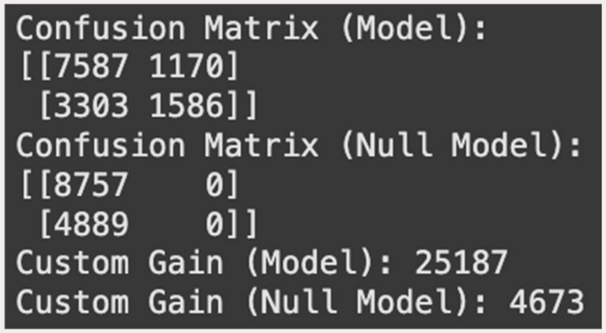
```
                                   Mean Negative RMSE
Model
Null Model (Mean Prediction)                   0.7554
```



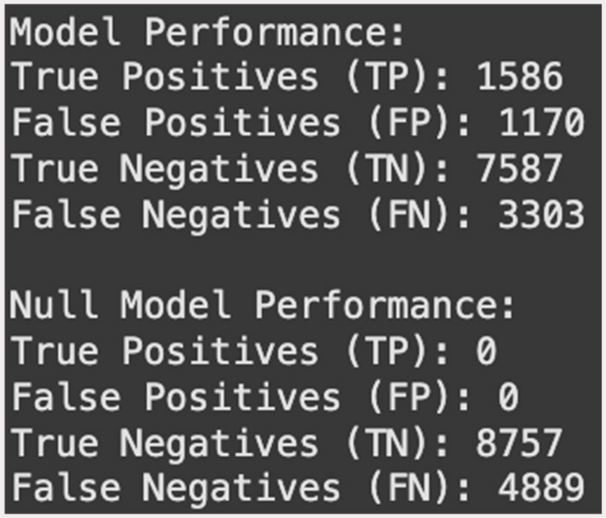
```
Voting Regressor Test RMSE score: 0.6941
```

Combined Model





```
Confusion Matrix (Model):
[[7587 1170]
 [3303 1586]]
Confusion Matrix (Null Model):
[[8757    0]
 [4889    0]]
Custom Gain (Model): 25187
Custom Gain (Null Model): 4673
```



```
Model Performance:
True Positives (TP): 1586
False Positives (FP): 1170
True Negatives (TN): 7587
False Negatives (FN): 3303

Null Model Performance:
True Positives (TP): 0
False Positives (FP): 0
True Negatives (TN): 8757
False Negatives (FN): 4889
```

Challenges & Key Takeaways

1. Extensive Data Preprocessing
2. Skewed Dataset
3. Google Trends Integration
4. Managing Narrowly Distributed Non-Zero Click Percentages
5. Slow Performance of Certain Models
6. Stacking Model Challenges
7. Classifier Voting Limitations

Conclusion & Future Steps

TODAY'S NEWS
HOME
POLITICS
ECONOMY
WORLD
LIFESTYLE
SPORTS
HEALTH
TECHNOLOGY
CULTURE
WEATHER
Cover Story: the Truth About Our Current Affairs
How to Buy Fashion Brands on a Budget
Addressing the Health Accessibility

- The regression models used were relatively weak learners.
- Explore ensemble methods like boosting for better performance.
- Integrate search terms with trend velocity to refine predictions.
- Expand applications to ads, YouTube videos, and other digital content.

Link to Final Colab Notebook

Predicting News Popularity With Supervised Machine Learning