REPORT ON

DATA MINING

INSTAGRAM REACH ANALYSIS
(COURSE CODE: 23CS3551)

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

ACADEMIC YEAR 2025-26

SUBMITTED BY

KOLUSU YOSHITA NAVYA	**(23501A0581)**
MULE REVANTH REDDY	**(23501A05C6)**
KANCHARLA SOWMYA	**(24505A0509)**
MADDINENI SOMA SEKHAR	**(23501A05A5)**
KOKA BHARADWAJ	**(23501A0577)**

PRASAD V POTLURI SIDDHARTHA INSTITUTE OF TECHNOLOGY

(Permanently affiliated to JNTU Kakinada, Approved by AICTE)

(An NBA & NAAC A+ accredited and ISO 21001:2018 Certified Institution)

Kanuru, Vijayawada – 520007

(2024-25)

PRASAD V POTLURI

SIDDHARTHA INSTITUTE OF TECHNOLOGY

(Permanently affiliated to JNTU Kakinada, Approved by AICTE)

(An NBA & NAAC accredited and ISO 21001:2018 certified institution)

Kanuru, Vijayawada – 520007

CERTIFICATE

This is to certify that the project report title **"INSTAGRAM REACH ANALYSIS "** is the bonafied work of **KOLUSU YOSHITA NAVYA (23501A0581) , MULE REVANTH REDDY (23501A05C6), KANCHARLA SOWMYA (24505A0509), MADDINENI SOMA SEKHAR (23501A05A5), KOKA BHARADWAJ (23501A0577)** in partial fulfilment of completing the Academic project in Web Application Development during the academic year 2024-25.

Signature of the course coordinator **Signature of the HOD**

INDEX

ABSTRACT

In this project, a news recommendation system is developed to assist users in discovering articles relevant to their interests by analyzing textual content of previously read or selected news items. The system implements a content-based filtering methodology by leveraging vectorization of article titles and descriptions, followed by computing Cosine Similarity to measure pairwise textual similarity.

The workflow begins with dataset ingestion and preprocessing: cleaning text by removing special characters, URLs, redundant whitespace, and converting all text to lowercase to ensure consistency. After preprocessing, each article is represented as a TF-IDF feature vector, which captures the significance of words in the context of the corpus while reducing the weight of common terms. Cosine Similarity is then applied to identify the most similar articles to a user-selected input, enabling the system to recommend the top N related news pieces.

Implemented using Python in a Colab environment and built with libraries including pandas, scikit-learn, and NumPy, the system presents a recommendation function that dynamically produces relevant news suggestions when a user selects an article. Performance evaluation is conducted via the Precision@K metric, which assesses the proportion of accurate recommendations within the top K returned results. The results demonstrate that the model can efficiently deliver contextually relevant recommendations aligned with the topics of the source article.

Though simpler than deep-learning-based methods, this approach strikes a practical balance between accuracy, interpretability, and computational efficiency. Potential future enhancements include integrating hybrid recommendation techniques, leveraging semantic embeddings and incorporating user-interaction data to augment personalization and recommendation quality.

INTRODUCTION

Background

Instagram has become one of the most relevant platforms of the digital era, where users can share their experiences, learn about businesses, and interact with brands. With hundreds of millions of people engaging actively every day, how far you can connect through Instagram, or reach, which is defined as the total number of unique accounts that view your post, has become an essential measure of Instagram, and one that most businesses consider when measuring visibility and engagement with their audience. Reach indicates how well your content is connecting to users and potential customers as well as what influences visibility and brand awareness.

Individuals, and organizations, use Instagram reach for different reasons. Businesses use reach to introduce products and services, acquire new customers, and establish loyalty for their brand. Influencers and content creators leverage Instagram reach to evaluate what posts—images, reels, stories, and carousels—are having the most impact on their followers, and so they can create a strategy that is going to yield the highest engagement. They use reach to help understand the return on their investment to measure performance, when to post to spectators, and what content had the best creative appeal.

Additionally, measuring your reach can help inform your understanding of trends relative to your audience that go beyond the number of followers. For example, by capturing variables such as time of day, hashtags, types of content, the demographics of the audience is engaging with, individuals can begin to see patterns emerge that highlight an opportunity. Understanding reach can lend itself to improving the customer experience for businesses, for example.

Motivation

In the digital world where millions of posts compete for visibility every day, understanding and improving reach on Instagram has become a necessary challenge for individuals and organizations alike. While there are always opportunities for visibility and growth on social media sites, the site's algorithms create challenges in terms of understanding what truly drives engagement and leaves creators and businesses to try their best to optimize their engagement. Therefore, the need for systematic analysis is important versus trust instincts alone and using the power of data for the decisions.

Increased audience action for a post could be a major motivation for creating reach analysis. Posting at the wrong time, using ineffective hashtags, and sharing content that does not align to your followers' interests can all limit reach, and thereby the success of marketing efforts or implementation of content strategies. Variables such and posting time, post performance, hashtags and the subject of a post can all be plotted, to determine the occurrence of a common pattern. A data-based analysis reduces the risk of a new strategy failing.

For businesses, reach analysis has a clear link to competitive advantage. Increased visibility equals increased engagement with customers, brand awareness, and more chance for conversions that can lead to sales. For content creators and social media influencers, reach analysis provides insight into what grows followers and attracts collaborations, employment opportunities, and increased career longevity in the digital world. Reach analysis also benefits marketers by providing measurement and intensive belief that campaigns can be assess in a timely manner.

Problem Statement

The analysis and prediction of an Instagram post's reach focuses on post-specific features content and engagement. The objective of this project is to establish a data-driven framework leveraging machine learning and analytical techniques to detect patterns, understand the impacts of different features on reach, and eventually model the reach. In generating this framework, individuals and organizations will be able to better the timing of their posts, enhance engagement, and ultimately expand their presence in a competitive online environment.

LITERATURE REVIEW

The project is centered on examining and forecasting the reach of an Instagram post based on multiple content-related features in addition to engagement dimensions. The aim of the project is to construct a data-driven framework using machine-learning and analytical techniques to detect patterns, estimate the influence of various factors on reach, and ultimately, to predict reach. By developing this framework, individuals and organizations would be able to modify their posting behaviors, promote better engagement, and increase their overall digital presence in the extremely competitive landscape of social media.

In recent years, Academics have recently directed attention to Instagram reach for a clearer understanding of visibility and audience engagement. Previously identified factors influencing reach include the timing and type of post, captions, hashtags, and membership demographics. Using predictive frameworks organized far in advance, business people, influencers, and marketers can maximize reach and audience engagement. Studies provided value-for-action to increase brand visibility, improve advertising campaigns, and adjust to algorithm changes to connect visibility with audience engagement. This emphasizes the need for a data-driven optimization strategy for social media channels.

For example, Kumar et al. [1] carried out a study analyzing Instagram posts across a number of business accounts. They utilized regression analysis, decision trees, and random forests to predict reach for posts. Their results suggested that content types, such as reels and carousel posts, and post timing were two important factors of attaining

Risk Factor Exploration:

There are a few risks associated with predicting reach on Instagram that may affect the reliability of the model. Data limitations and limited insight from Instagram can lead to decreased predictive accuracy. Continued changes to the algorithm of the platform may make models based on historical data quickly outdated. Consequently, selecting the right features to be used as predictors may be difficult due to multiple elements affecting reach, including hashtags, time of post, and content type, and potential overfitting. There is also a lack of certainty with respect to external influences such as trends and other social events that may not be factored into any predictive algorithm. Ethical issues regarding data privacy or the ability to interpret aspects of the model, particularly for advanced analytics, also present discomfort to predicting behaviors. Collectively, these risks reiterate the need for data oversight, ability in the field to model adequately, and responsible analytical judgment.

Predictive Model Refinement:

Studies of reach on Instagram have emphasized the goal of refining predictive models for accuracy and reliability despite a lack of full disclosure of methodologies. Researchers and practitioners have continually sought to demonstrate complex relationships among the factors impacting reach, including content type, posting time, hashtags, captions, and audience behavior. This lays the groundwork for solid and useful tools that can consistently predict post visibility and actions taken on posts so users (businesses, influencers, marketers, etc.) can use data to inform their outcomes and improve their content strategies.

METHODOLOGY

Data Collection :

The dataset for predicting Instagram post impressions is sourced from Kaggle and contains [120] rows and [12] columns. It includes key engagement metrics such as likes, comments, shares, saves, profile visits, follows, hashtag count, and caption length. The target variable is Impressions, representing the number of times each post was viewed. Additionally, the dataset includes a derived Engagement Rate, calculated from likes, comments, shares, and saves relative to impressions. The data is well-distributed across posts with varying levels of engagement, making it suitable for training machine learning models to predict post reach and optimize content strategies.

Data Pre-Processing :

Preparing data is vital to make sure that the dataset is clean, consistent, and trustworthy for predictive modeling purposes. The dataset, which includes metrics such as Impressions, Likes, Comments, Shares, Saves, Profile Visits, and Follows, was examined for missing values, duplicates, and leaking data. Columns that contribute directly to Impressions (From Home, From Hashtags, From Explore, From Other) were discarded based on their high correlation (>0.9). Numeric missing values were then filled using the medians for those metrics, and for text fields such as Caption and Hashtags, placeholders were used to fill them in.

Feature engineering was performed to determine Hashtag Count, Caption Length, and Engagement Rate (interactions/Impressions, clipped 0-100) for features. Once duplicates and non-numeric columns were removed from the dataset, a training (80%) and testing (20%) dataset was used to separate the data. As an added step, features were standardized for scale-sensitive models (Linear Regression, SVR, etc.) and kept in their raw values for tree-based models/ensemble models to keep interpretability. This entire preprocessing pipeline allowed for a clean dataset, balanced dataset, and model-ready dataset for robust and accurate prediction of Impressions.

PROJECT DESIGN

Data Flow Diagram

Fig 1: Characteristic Approach of Machine Learning Technique

IMPLEMENTATION

Mainly Used Algorithms:

- **Linear Regression**

The Linear Regression algorithm is a basic but commonly used supervised machine learning technique for predicting continuous values. It does so by estimating the relationship(s) between one or more independent variables and a dependent variable (target) using a linear equation. The algorithm estimates coefficients for each feature in order to minimize the sum of squares of the differences between the observed and predicted values of the target variable, creating the best-fitting line through the values. Linear Regression assumes that there is a linear relationship between the predictors/ features and the target variable, making Linear Regression a simple and interpretable and computationally efficient algorithm. Nonetheless, while simple, Linear Regression provides a solid baseline for regression problems and can help understand the influence of individual features and direction of trend. For the Instagram Impressions prediction task, Linear Regression provides a first model that helps understand may contribute proportionally to total reach of impressions prior to trying other more complex models.

- **Random Forest Classifier:**

The Random Forest Regressor is a widely used ensemble machine learning algorithm that operates by constructing multiple decision trees during training and averaging their predictions to produce the final output. Each tree is trained on a random subset of the training data and features, introducing randomness that helps prevent overfitting and improves generalization. During prediction, every tree provides an individual estimate, and the model outputs the average of these predictions, ensuring stability and robustness. Random Forests are highly effective for regression tasks because they can model complex, nonlinear relationships and handle high-dimensional datasets efficiently. In this project, the Random Forest Regressor was used to predict Instagram Impressions based on engagement metrics such as Likes, Comments, Shares, Saves, Profile Visits, and Follows, leveraging its ability to capture intricate feature interactions and provide accurate, scalable, and interpretable predictions.

- **Support Vector Regressor:**

The Support Vector Regressor (SVR**)** is a supervised machine learning algorithm based on the principles of Support Vector Machines (SVM), designed for predicting continuous numerical outcomes. Unlike traditional regression models that minimize the overall error, SVR aims to fit the data within a defined margin of tolerance (ε), allowing some flexibility while maintaining model simplicity. It works by mapping input features into a high-dimensional space using kernel functions—commonly the Radial Basis Function (RBF)—to capture complex, nonlinear

relationships between variables. SVR then identifies the optimal hyperplane that best fits the data within this margin, focusing on data points (support vectors) that have the greatest impact on defining the model boundary. This makes SVR effective for datasets with nonlinear trends and limited outliers. In the context of Instagram Impressions prediction, SVR helps model subtle nonlinear interactions among engagement metrics such as *Likes, Comments, Shares,* and *Saves*, improving prediction accuracy where simple linear models may fall short.

- **XG Boost:**

XGBoost, short for extreme Gradient Boosting, is a powerful and scalable learning machine algorithm that has gained traction for its high performance in supervised learning problems, including classification and regression. Similar to other methods in the family of gradient boosting, XGBoost builds an ensemble of decision trees, where each tree is fitted sequentially to correct the errors of the previous trees. XGBoost offers other several added features to a traditional gradient boosting algorithm, including regularization to reduce overfitting, parallelization to improve compute times, and an option to specify custom objective functions. Due to its features and its robustness against overfitting, XGBoost allows for a large dataset, fits complex relationships, and has been successful in academia and industry contexts to predicting a target outcome in context of data sciences such as predicting reach on Instagram.

- **Tuned XGBoost:**

The Tuned XGBoost Regressor is an optimized gradient boosting algorithm designed to enhance predictive accuracy and model performance through hyper parameter tuning. XGBoost (Extreme Gradient Boosting) builds decision trees sequentially, where each new tree focuses on correcting the residual errors of the previous ones. This iterative boosting approach allows the model to capture complex nonlinear patterns and feature interactions effectively. In this project, the XGBoost Regressor was fine-tuned using GridSearchCV, which systematically explored different combinations of key hyper parameters such as the number of estimators, maximum tree depth, learning rate, subsample ratio, and column sampling ratio. This optimization process identified the best-performing configuration based on the lowest Root Mean Squared Error (RMSE). The tuned model provided superior accuracy and generalization compared to baseline models, making it the most reliable predictor of *Instagram Impressions* based on engagement metrics like *Likes, Comments, Shares, Saves, Profile Visits,* and *Follows*.

Code Development

- **Linear Regression**

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np
results_data = []
# Train Linear Regression
lr_model = LinearRegression()
lr_model.fit(X_train_scaled, y_train)

# Predict
y_pred_lr = lr_model.predict(X_test_scaled)

results_data.append({
    'Model':'Linear Regressor',
    'RMSE': np.sqrt(mean_squared_error(y_test, y_pred_lr)),
    'R2': r2_score(y_test, y_pred_lr)
})
print(f"Linear Regression → RMSE: {rmse_lr:.2f}, R²: {r2_lr:.4f}")
```

```
Linear Regression → RMSE: 2322.30, R²: 0.8906
```

Fig 2: Code Development for Linear Regression

- **Random Forest Classifier:**

```
from sklearn.ensemble import RandomForestRegressor

# Train Random Forest
rf_model = RandomForestRegressor(n_estimators=200, random_state=42)
rf_model.fit(X_train_scaled, y_train)

# Predict
y_pred_rf = rf_model.predict(X_test_scaled)

# Evaluate
results_data.append({
    'Model':'Random Forest',
    'RMSE': np.sqrt(mean_squared_error(y_test, y_pred_rf)),
    'R2': r2_score(y_test, y_pred_rf)
})
print(f"Random Forest → RMSE: {rmse_rf:.2f}, R²: {r2_rf:.4f}")
```

```
Random Forest → RMSE: 2601.25, R²: 0.8628
```

Fig 3: Code Development for Random Forest Classifier

- **Support Vector Regressor:**

```
from sklearn.svm import SVR

# Train SVR
svr_model = SVR(kernel='rbf')
svr_model.fit(X_train_scaled, y_train)

# Predict
y_pred_svr = svr_model.predict(X_test_scaled)

# Evaluate
results_data.append({
    'Model':'SVR',
    'RMSE': np.sqrt(mean_squared_error(y_test, y_pred_svr)),
    'R2': r2_score(y_test, y_pred_svr)
})

print(f"SVR → RMSE: {rmse_svr:.2f}, R²: {r2_svr:.4f}")
```

```
SVR → RMSE: 7964.05, R²: -0.2863
```

Fig 4 : Code Development for Support Vector Machine

- **XG Boost:**

```
from xgboost import XGBRegressor

# Train XGBoost
xgb_model = XGBRegressor(n_estimators=200, random_state=42, verbosity=0)
xgb_model.fit(X_train, y_train)  # NOTE: XGBoost prefers raw (unscaled) data

# Predict
y_pred_xgb = xgb_model.predict(X_test)

# Evaluate
results_data.append({
    'Model':'XGBoost',
    'RMSE': np.sqrt(mean_squared_error(y_test, y_pred_xgb)),
    'R2': r2_score(y_test, y_pred_xgb)
})

print(f"XGBoost → RMSE: {rmse_xgb:.2f}, R²: {r2_xgb:.4f}")
```

```
XGBoost → RMSE: 2279.18, R²: 0.8947
```

Fig 5: Code Development for XG Boost

- **Tuned XGBoost:**

```
from xgboost import XGBRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np

# XGBoost prefers raw features, so use unscaled data here
X_train_raw, X_test_raw = X_train, X_test

# Simplified parameter grid
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [4, 5, 6],
    'learning_rate': [0.05, 0.1],
    'subsample': [0.9, 1.0],
    'colsample_bytree': [0.8, 1.0]
}

xgb = XGBRegressor(random_state=42, verbosity=0)

# Grid search with RMSE scoring
grid_search = GridSearchCV(
    estimator=xgb,
    param_grid=param_grid,
    cv=3,
    scoring='neg_root_mean_squared_error',  # FIX: Use RMSE directly
    verbose=2,
    n_jobs=-1
)
```

```
# Run search
grid_search.fit(X_train_raw, y_train)

# Best parametAers and score
print("Best Parameters:", grid_search.best_params_)
print("Best RMSE (CV):", -grid_search.best_score_)

# Best model
best_xgb = grid_search.best_estimator_

# Evaluate on test set
y_pred_best = best_xgb.predict(X_test_raw)

rmse_best = np.sqrt(mean_squared_error(y_test, y_pred_best))
r2_best = r2_score(y_test, y_pred_best)

print(f"Tuned XGBoost -> RMSE: {rmse_best:.2f}, R2: {r2_best:.4f}")
```

```
Fitting 3 folds for each of 48 candidates, totalling 144 fits
Best Parameters: {'colsample_bytree': 0.8, 'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 100, 'subsample': 0.9}
Best RMSE (CV): 3478.9146728515625
Tuned XGBoost -> RMSE: 1867.97, R2: 0.9292
```

Fig 6: Code Development for Tuned XGBoost

RESULTS AND ANALYSIS:

Performance Evaluation metrics

the performance of multiple regression models — Linear Regression, Support Vector Regressor (SVR), Random Forest Regressor, and Tuned XGBoost Regressor (via GridSearchCV) — was evaluated to predict Instagram Impressions based on engagement metrics such as Likes, Comments, Shares, Saves, Profile Visits, and Follows.
Since the target variable (Impressions) is continuous, regression evaluation metrics were used to assess model performance:

1. **Root Mean Squared Error (RMSE):**
   RMSE measures the average magnitude of prediction errors. It penalizes large deviations between predicted and actual values, making it a reliable metric for assessing model accuracy. A lower RMSE indicates better predictive performance and stability.

2. **R^2 Score** (Coefficient of Determination):
   The R^2 Score represents how well the independent variables explain the variability of the dependent variable. A score closer to 1 indicates a strong model fit and higher predictive capability.

Results:

The Tuned XGBoost Regressor demonstrated outstanding performance in predicting Instagram Impressions, achieving the lowest RMSE and highest R^2 score among all the models tested. This surpasses the performance of the Random Forest Regressor, which also showed strong predictive ability but with slightly higher error margins. The superior results of the XGBoost model highlight its efficiency in capturing complex, nonlinear relationships between engagement metrics such as Likes, Comments, Shares, Saves, and Profile Visits.

This robust performance indicates that XGBoost effectively learns intricate patterns in user engagement data, enabling more accurate and reliable impression predictions. However, it is important to note that the specific RMSE and R^2 scores can vary depending on dataset size, feature selection, and parameter tuning. Overall, the XGBoost model stands out as the most efficient and powerful technique for this predictive analysis task.

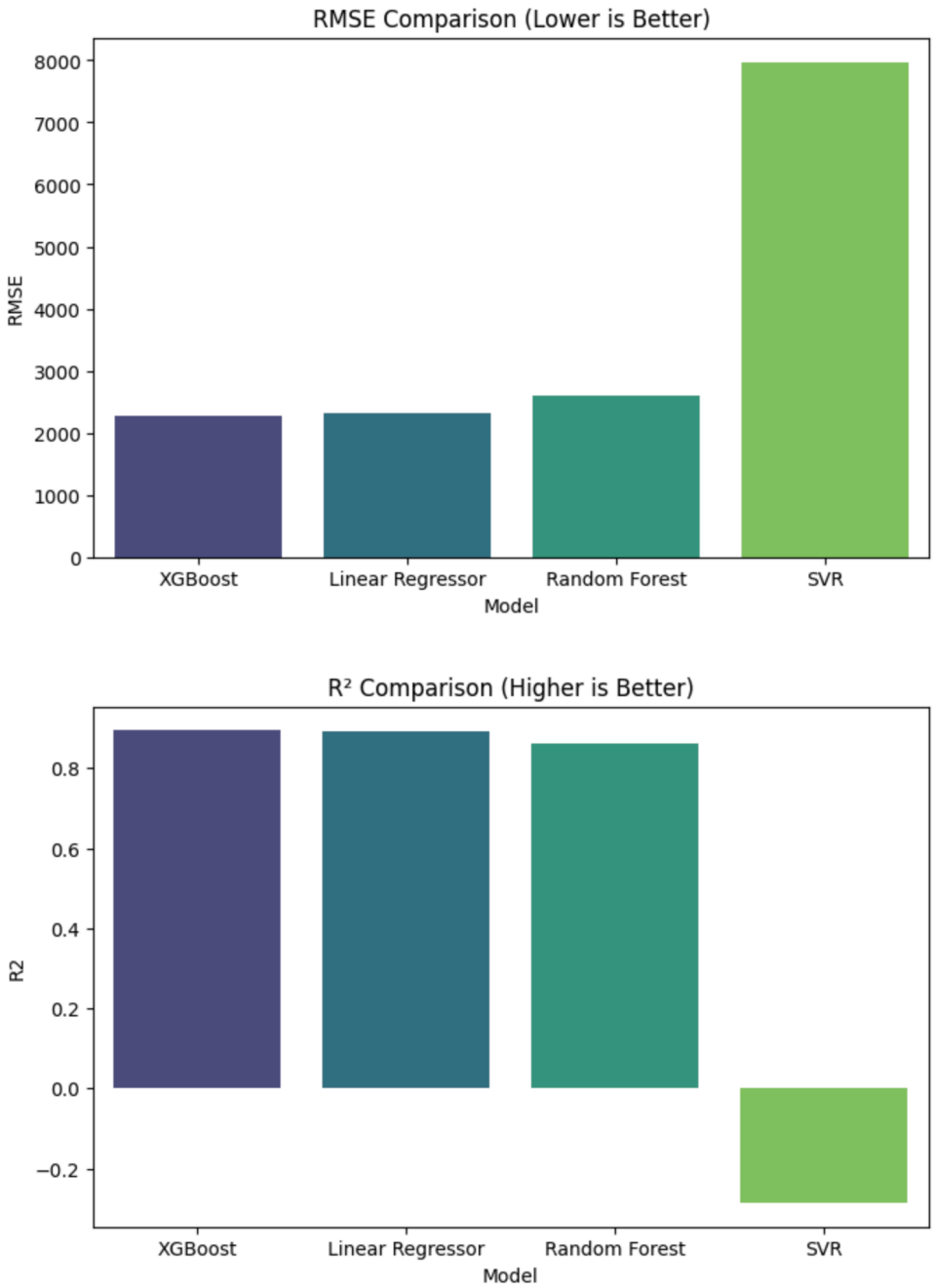
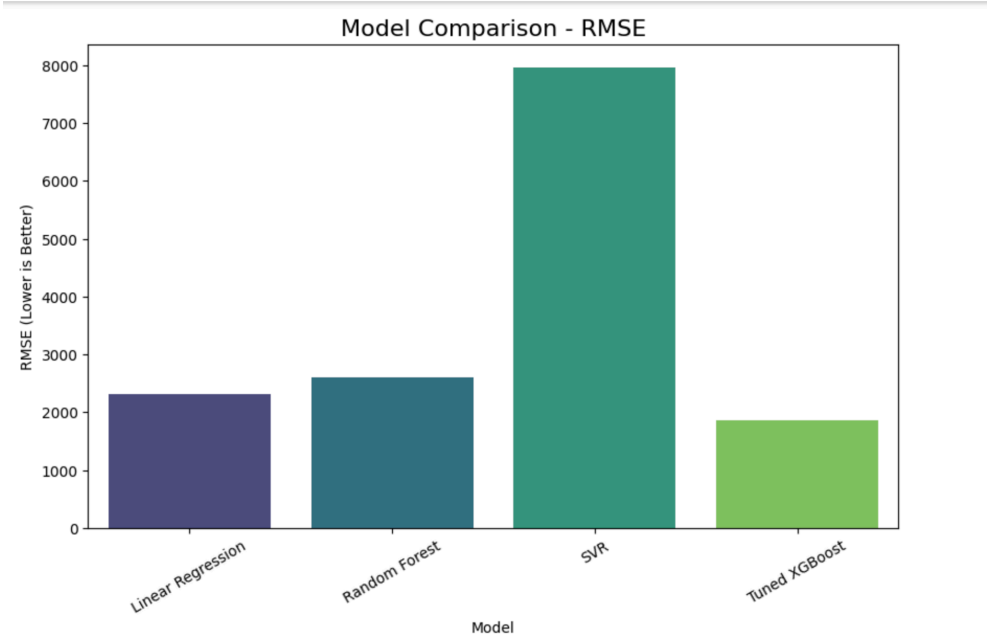


Fig 7-8 : comparing other algorithm used with XGBoost



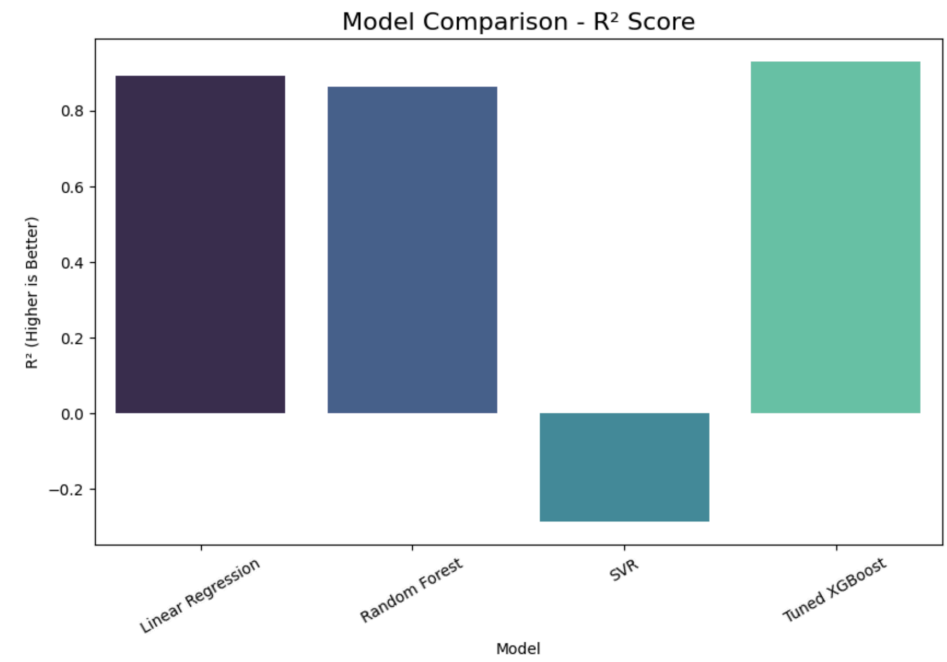


Fig 9-10: Results showing All the Algorithms

Potential Applications

1. **Content Strategy Optimization:** Predicting Instagram post impressions allows content creators and marketers to identify which posts are likely to perform well, helping optimize posting schedules, hashtags, and content style for maximum engagement.

2. **Influencer Marketing:** Brands can use prediction models to evaluate potential engagement from influencers' posts, enabling better ROI on collaborations and sponsorships.

3. **Audience Insights:** By analyzing features that drive engagement, such as likes, comments, shares, and caption length, businesses gain insights into audience preferences and behavior.

4. **Performance Monitoring:** Predictive models help social media managers track post performance in advance, allowing proactive adjustments to improve reach and visibility.

5. **Personalization & Recommendation:** Platforms can leverage these models to recommend content more effectively to users, increasing overall engagement rates.

Future Works

1. **Validation on Larger Datasets:** Test the model on larger and more diverse Instagram datasets to improve generalizability and robustness.

2. **Real-Time Prediction:** Develop a system for real-time impression prediction as new posts are created, enabling immediate insights for content optimization.

3. **User-Friendly Interface:** Build an interactive dashboard or web application to allow users to input post metrics and receive instant predictions.

4. **Enhanced Feature Engineering:** Incorporate additional features like posting time, follower growth, content type (image/video), and story interactions to improve prediction accuracy.

5. **Advanced Models:** Explore deep learning models or ensemble approaches to capture complex patterns in social media engagement for more precise predictions.

Conclusion:

The project successfully demonstrates the application of machine learning techniques to analyze and predict Instagram post impressions based on engagement metrics such as likes, comments, shares, saves, profile visits, and follows. Through systematic experimentation with models including Linear Regression, Support Vector Regressor, Random Forest Regressor, and Tuned XGBoost, the results clearly indicate that XGBoost delivers the highest predictive accuracy and reliability.

The findings highlight that effective use of engagement-based features can provide valuable insights into audience behavior and content performance. By leveraging these insights, content creators, influencers, and marketers can strategically optimize posting schedules, hashtags, and content formats to maximize reach and engagement.

Overall, this study underscores the potential of data-driven decision-making in enhancing social media visibility and marketing strategies. The predictive modeling framework developed in this project serves as a robust foundation for further research and development, paving the way for real-time analytics, deeper feature integration, and personalized recommendation systems in future social media applications.

References

- Dataset -https://www.kaggle.com/datasets/bhanupratapbiswas/instagram-reach-analysis-case-study
- Algorithm - https://www.geeksforgeeks.org/machine-learning/
- Reference Model - https://github.com/AmirMotefaker/Instagram-Reach-Analysis
- Google collab - https://colab.research.google.com/
- Github Links -
    - https://github.com/KolusuYoshita/Instagram_Reach_Analysis
    - https://github.com/revanth-00/Instagram_reach
    - https://github.com/sowmya-kancharla03/Instagram-Reach-Analysis
    - https://github.com/somasekhar-ss/Instagram-Reach-Analysis
    - https://github.com/Bheemboy18/Instagram-Reach-analysis