CS 412 Project Report
Hande Altunbaş - 32643
Cansu Temizkan - 30710
Mert Seçen - 22532
Efe Yağız Kılıçkaya - 30619
Defne Koçulu-31292

Round 1

Classification Task

Objective

The aim of this classification task was to predict whether a social media user belongs to the "Food" category or "Not Food" based on their profile features and post captions. This binary classification problem was addressed using RandomizedSearchCV for hyperparameter tuning and a Random Forest Classifier.

Methodology

Model Used

- Random Forest Classifier Dataset
- Training and testing datasets were prepared using user profiles and their posts:
    - Features included:
        - Textual Features: Combined captions preprocessed and vectorized using TF-IDF.
        - Numeric Features: Follower count, following count, average likes, comments, and food-related word counts.
    - Labels were binary: "Food" (1) or "Not Food" (0).

Hyperparameter Tuning

- Used RandomizedSearchCV with cross-validation for hyperparameter optimization.

Results

```
Test Set Performance:
Accuracy: 0.9379562043795621
              precision    recall  f1-score   support

    Not Food       0.96      0.97      0.96       238
        Food       0.79      0.72      0.75        36

    accuracy                           0.94       274
   macro avg       0.87      0.85      0.86       274
weighted avg       0.94      0.94      0.94       274


Predictions saved to 'predictions.json'.
```

Conclusion

The classification model achieved an accuracy of 94 on the test set, demonstrating strong performance in identifying "Food" users based on extracted features. Further improvements could include addressing class imbalance and exploring advanced models like XGBoost for potentially better results.

Regression Task

Methodology

- Model Used: Linear Regression
- Dataset: Regression dataset with 3,000 records.
- Process:
    - Extracted features such as word count, media type, and engagement metrics (e.g., like count).
    - Processed text data using TF-IDF for captions.
    - Split the data into training and testing sets.
    - Evaluated the model using Mean Absolute Error (MAE) and R-squared metrics.

Findings

- Performance
    - The regression model achieved an R-squared value of 0.65, indicating moderate correlation between features and the target variable.
    - The Mean Absolute Error (MAE) was relatively high, suggesting variability in predictions.

- Strengths
    - Captions with trending hashtags contributed positively to prediction accuracy.
    - Media type was a significant predictor, with carousel albums generally yielding higher like counts.

- Challenges
    - High Variability: Posts with very high or very low like counts were not well-predicted, indicating that the model struggled with outliers.
    - Limited Features: Key factors influencing engagement, such as the time of posting and follower activity, were not available in the dataset.
    - Non-linear Relationships: The linear regression model may not have captured complex interactions between features, such as caption tone and media type.

Conclusion

While the regression model provided moderate predictive power, its performance was limited by high error rates and the linearity of the approach. Incorporating more comprehensive features and exploring advanced models could significantly improve prediction accuracy for engagement metrics like counts.

Round 1 Regression task was attempted by all team members, with Defne and Hande achieving acceptable results. Among which Hande's was selected as the regression code to be executed with Defne's fine tuning.

For the classification task, Efe, Mert and Cansu came up with a solution. Similarly, classification codes were compared against each other with Efe's code having the best performance for classifying "Food" category, it was selected to be worked further on. Mert and Cansu helped fine tune Efe's code.

Everyone debugged/helped format each other's code and we all helped each other navigate the project, the demands and the expectations thus in each step of the way, we collaborated.

Round 2

Classification Task

- **Model Used**: Random Forest
    - Submitted predictions using this model.
- **SMOTE**: Attempted to address class imbalance using SMOTE, but it did not improve results.
- **XGBClassifier**: Tried using XGBoost for classification, but performance was not satisfactory.
- **SVC with Linear Kernel**: Evaluated the model, but accuracy was relatively low.
- **Cross-Validation**: Performed to evaluate model reliability and detect overfitting.

Regression Task

This task was carried out collaboratively by two individuals. Both approaches were thoroughly evaluated, and the final model, Linear Regression, was selected based on its lower Mean Squared Error (MSE) score compared to the alternative model. However, we also experimented with the Random Forest Regressor as part of our implementation. The Random Forest model demonstrated robust performance by leveraging its ensemble approach to handle both numerical and categorical data effectively. Although it was not chosen as the final model due to its higher MSE, it provided valuable insights and served as a benchmark during the experimentation phase.

Methodology

Model Used:

- Random Forest

Dataset:

- A regression dataset containing 3,000 records.

Process:

1. Extracted features such as word count, media type, and engagement metrics (e.g., like count).
2. Processed text data using TF-IDF for captions.
3. Preprocessed data, including standardizing numerical features and encoding categorical features.

4. Split the dataset into training and testing subsets.
5. Evaluated the model using metrics such as Mean Absolute Error (MAE) and R-squared.

Findings

Performance:

- The Random Forest Regression model achieved an R-squared value of 0.53, indicating a moderate correlation between features and the target variable.
- The Mean Absolute Error (MAE) was relatively high, suggesting some variability in predictions.

Strengths:

- Captions with trending hashtags contributed positively to prediction accuracy.
- Media type was a significant predictor, with carousel albums generally yielding higher like counts.

Challenges:

1. High Variability: Posts with very high or very low like counts were not well-predicted, indicating the model struggled with outliers.
2. Limited Features: Key factors influencing engagement, such as posting time and follower activity, were unavailable in the dataset.
3. Non-linear Relationships: The Linear Regression model may not have captured complex interactions between features, such as caption tone and media type.

Conclusion

While the regression model provided moderate predictive power, its performance was limited by high error rates and the linear nature of the approach. To improve prediction accuracy for engagement metrics like counts, future work could incorporate more comprehensive features and explore advanced, non-linear models such as ensemble or deep learning techniques

In Round 2, the role allocation was done to maximize the efficiency of the team by allocating members with parts of the project each were familiar with. Thus, Hande and Defne worked on

the Regression and Efe, Mert and Cansu worked on the classifier, with every team member helping debug/format the code of the fellow team members regardless of their own task.

In the 1st round, the classifier could only predict the “Food” category, albeit with great accuracy. To implement the rest, we have focused more on the classifier this time and thanks to the team’s collaborative effort, we managed to improve the classifier.

Round 3

Regression Task

This task was carried out collaboratively by two individuals. Multiple models were evaluated to identify the best-performing approach. The CatBoost Regressor was trained and evaluated as part of the experimentation process. While it demonstrated strong performance with advanced feature interaction handling, the final model was selected based on a comparison of error metrics across all evaluated models.

Methodology

Model Used:

- CatBoost Regressor

Dataset:

- A regression dataset containing 3,000 records.

Process:

1. Extracted features such as word count, media type, and engagement metrics (e.g., like count).
2. Processed data using appropriate feature transformations, including interactions and encoding.
3. Trained the CatBoost Regressor, leveraging its capability to automatically handle categorical features and explore feature interactions.
4. Split the data into training and testing sets.

5. Evaluated the model using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

Findings

Performance:

- Mean Absolute Error (MAE): 46.32
- Root Mean Squared Error (RMSE): 93.66

These metrics indicate that the CatBoost Regressor provided reasonably accurate predictions with a strong capacity for capturing feature interactions. The relatively low MAE and RMSE suggest that the model performed well on the dataset.

Strengths:

- Feature Interaction Handling: CatBoost's ability to model complex feature interactions contributed positively to the overall performance.
- Automatic Categorical Feature Handling: The model efficiently processed categorical features without extensive preprocessing.
- Speed and Scalability: Despite the dataset size, the training process was manageable, completing in approximately four minutes.

Challenges:

1. High Variability in Targets: Posts with extremely high or low like counts introduced prediction challenges, as evidenced by the RMSE.
2. Limited Feature Set: While feature interaction was utilized, the absence of additional contextual features, such as time of posting or follower activity, limited further performance improvements.
3. Overhead in Tuning: Optimizing CatBoost parameters required additional computation and experimentation compared to simpler models.

On Round 3, we decided to adopt a more collaborative approach. Since everyone familiarize themselves with both tasks of the project, we decided to run a little competition among each other to see whose code performed the best. After narrowing down the finalists, we all worked together on the final codes, with remarkable improvements.

Conclusion

The CatBoost Regression demonstrated robust performance in the regression task, effectively leveraging feature interactions and handling categorical variables. Its MAE and RMSE scores indicate strong predictive capability. However, further improvements could be achieved by incorporating additional features or fine-tuning hyperparameters. While CatBoost provided competitive results, the final model was chosen based on a comparative evaluation of multiple models, including error metrics. Future work could explore a hybrid approach that combines CatBoost with other models to enhance prediction accuracy.