Interview _ mini project

Amity Lu (Dec 2023)

Project Overview

This project aimed to predict the number of likes on Instagram posts.

Data Preprocessing

- Handling missing values and converting the timestamp into a readable format.
- Applying log transformations to 'likes,' 'comments,' and 'follower count' to normalize their distributions.
- Creating additional features such as 'comments squared' and 'followers squared' to explore non-linear relationships.

Findings:
- This dataset contains posts with zero comments, but none of the posts have zero likes.
- The log transformation of comments showed a significant relationship with the number of likes, suggesting its importance as a predictive feature.

Exploratory Data Analysis

- Conducting correlation analysis.
- Gaining insights into variable distribution and relationships through visual analysis, including box plots and scatter plots.

Findings:
- Correlation analysis revealed a moderate relationship between the log-transformed number of comments and likes.
- Visual analysis highlighted the presence of outliers in the data and remove these outliers.

Model Selection and Performance

Lasso Regression

- Conducting a grid search to find the optimal alpha value for Lasso Regression.
- The best alpha value found was 100.
- Achieving a Mean Squared Error (MSE) of 14,264,294,499.14 and an R^2 score of 0.5924.

Linear Regression

- Applying Linear Regression as a baseline model.
- Obtaining a resulting MSE of 14,265,029,451.45, with an R^2 score of 0.5924, which is similar to the Lasso Regression.

K-Nearest Neighbors Regressor

- The KNR model was tested with various values of 'k'.
- Achieving the lowest MSE of 6,289,247,605.73 and a higher R^2 score of 0.8177 at k=3, indicating a better fit than Lasso and Linear Regression.

Findings:
- The K-Nearest Neighbors Regressor, particularly with k=3, outperformed Lasso and Linear Regression models in terms of MSE.
- The similar performance of Lasso and Linear Regression indicates that feature selection (done by Lasso) did not significantly impact the model's predictive ability in this context.