Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Pushkar Srivastava

Email- pushkarsrivastava999@gmail.com

Contribution-

1. Data Wrangling
2. Data validation
3. Strategizing Roles for the analysis
4. Contributed to strategizing the exploration process
5. Contributed to strategizing the model creation process
6. Completed the presentation of the final report

Rahul Pandey

Email- rpanday2661997@gmail.com

1. Data Wrangling
2. Data Preprocessing and Feature engineering.
3. Confirmed the OLS assumptions and outlier removal.
4. Created Models for prediction.
5. Computed the error metrics to compare different models.
6. Worked on the presentation of the final report.

Hritik Sharma

Email-hritik.2.sharma@gmail.com

Contribution-

1. Data Wrangling
2. Data validation
3. Strategizing Roles for the analysis
4. Contributed to strategizing the exploration process

5. Contributed to strategizing the model creation process 6. Worked on presentation of the final report.

Please paste the GitHub Repo link.

GitHub Link:-
https://github.com/hritiksharma11/Capstone-project-ted-talk-views-prediction

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions.
(200-400 words)

Problem Statement

The main objective was to build a predictive model, which could help in predicting the views of the videos uploaded on the TEDx website. TEDx is a nonprofit organisation that aimed at bringing experts from the fields of Technology, Entertainment, and Design together.

Our Approach

We started the project with Data Exploration in which we faced many challenges as the dataset included data in many different data structures which was a bit complicated to extract but the most challenging part of EDA was dealing with categorical variables, it took a lot of research to deal with such a large number of categorical variables. We used visualizations to explain our findings mainly through matplotlib and seaborn libraries in python. Following EDA we performed feature engineering, data cleaning, target encoding and one hot encoding of categorical columns, feature selection, standardization and then model building. We dealt with missing data and outliers. That's a lot of work that Python helped us make easier. Then we checked our model for overfitting by comparing it with the Lasso Regression model, Ridge Regression model, and ElasticNet Regression model. And performed hyperparameter tuning with the help of grid search. We found that our original base model was overfitting and Lasso Regressor has the best accuracy. based on the MAE error metric.

Conclusion

In all of these models error has been around 34 % but we will consider the median error percentage, which is 13 %, to measure the accuracy because of the outliers. That implies we have been able to correctly predict views 87 % of the time.
After hyperparameter tuning, we have prevented overfitting and decreased errors by regularizing. Given that only 13 % errors, our models have performed very well on unseen data due to various factors like effective EDA, feature selection, and correct model selection.
Among all the features speaker_1_avg_views is the most important this implies that speakers are directly impacting the views.