Capstone Project Submission

Team Member's Name, Email and Contribution:

Aamir Sohail: aamirsohail23081998@gmail.com

- Exploratory data analysis – univariate and multivariate analysis.
- Data Wrangling – checking missing values, outliers, and features modification.
- Fitting Models – splitting the data, applying algorithms, evaluating, and model explanation.
- Presentation, Technical documentation.

Please paste the GitHub Repo link.

Aamir's Github Link:- https://github.com/Asohail115/Ted-Talk-Views-Prediction

A short summary of the Capstone project and its components.
the problem statement, approaches and conclusions.

TED is devoted to spreading powerful ideas on just about any topic. The dataset contains over 4,000 TED talks including transcripts in many languages Founded in 1984 by Richard Salman as a nonprofit organization that aimed at bringing experts from the fields of Technology, Entertainment, and Design together, TED Conferences have gone on to become the Mecca of ideas from virtually all walks of life. As of 2015, TED and its sister TEDx chapters have published more than 2000 talks for free consumption by the masses and its speaker list boasts of the likes of Al Gore, Jimmy Wales, Shahrukh Khan and Bill Gates. The data had variables such as talk_id, title, speaker_1, published_date, event, comments, duration, topics, related_talks and transcript.

The problem statement was to build a machine learning model that could predict the views of the videos uploaded on the TEDx website.

The first step in the exercise involved exploratory data analysis where we tried to dig insights from the data in hand. It included univariate and multivariate analysis in which we identified certain trends, relationships, correlation and found out the features who had some impact on our dependent variable.

The second step was to clean the data and also did some feature engineering. We checked for missing values and imputed them with the help of KNNimputer and also treated outliers. We also encoded the categorical variables.

The third step was to try various machine learning algorithms on our splitted and standardized data. We tried 3 different algorithms namely; Random Forest Regressor, XGBoost Regressor and Extra Trees Regressor. We did hyperparameter tuning and evaluated the performance of each model using various metrics. And also we implemented feature importance technique to understand which features were important for our prediction task.

The best performance was given by the Random Forest Regressor model where MAE was around 10% of target variable mean.

The model performed good and similarly we can also focus on future work like we can do dynamic regression time series modelling due to the availability of time features and also we can use topic modelling to tackle views in each topic separately.