Capstone Project on Regression

Ted talks views Prediction

Work done by- Sunil Kumar

Content

- Problem Statement.
- Data Summary.
- Project Story.
- Visualization.
- EDA-Top 5 speakers according to daily Views.
- EDA-Most Popular speakers.
- EDA-Data Cleaning- Steps and challenges.
- Correlation.
- Models.
- Feature Importance XGBoost and Random forest.
- Model Scores.
- Challenges.
- Conclusion.

Problem Statement

TED is devoted to spreading powerful ideas on just about any topic. These datasets contain over 4,000 TED talks including transcripts in many languages. Founded in 1984 The main objective is to build a predictive model, which could help in predicting the views of the videos uploaded on the TEDx website.

Data Summary

This dataset include the variables from 17 October 2017 to 20 April 2018

- **Talk id- Talk identification number provided by TED.**
- **title -Title of the talk.**
- **speaker_1 -First speaker in TED's speaker list.**
- **speakers -Speakers in the talk.**
- **occupations -Occupations of the speakers.**
- **About speakers *Blurb about each speaker dictionary.**
- **Views(Dependent Variable) Count of views.**
- **Recorded date -Date the talk was recorded.**

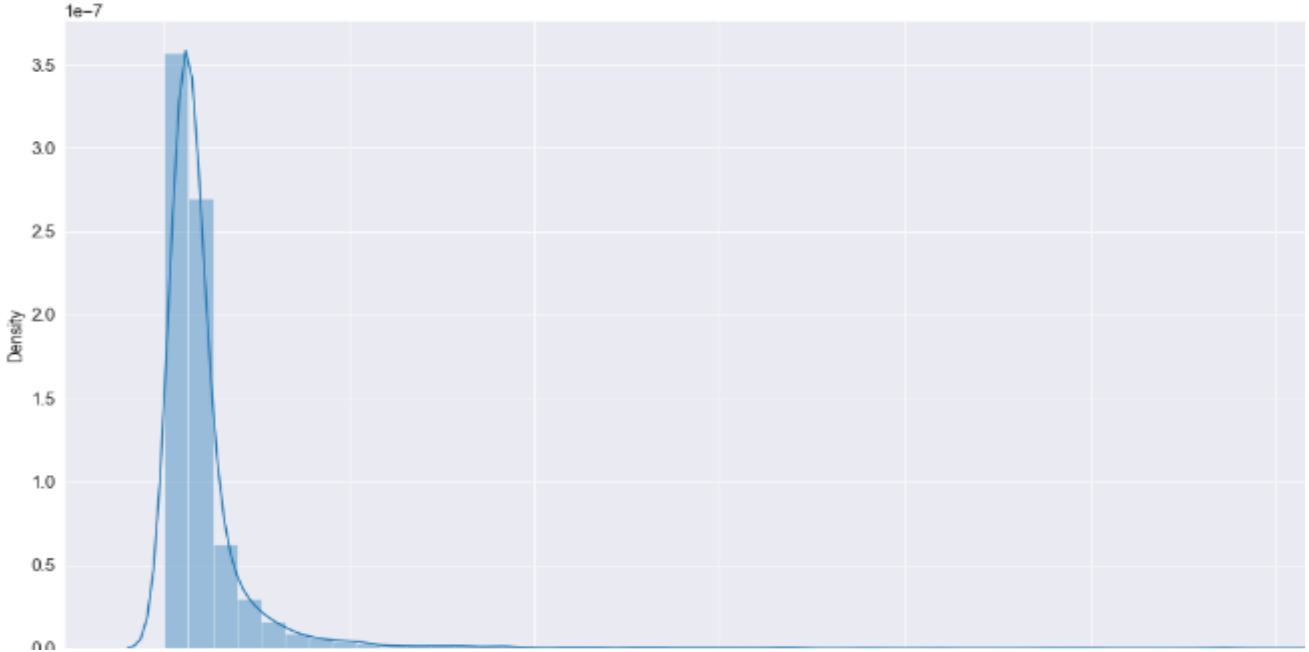
Data Summary continued..

- Available lang-All available languages (lang. code) for a talk .
- Comments-Count of comments.
- Topics-Related tags or topics for the talk .
- Related talks-Related talks (key='talk id', value='title') dictionary .
- Url -URL of the talk.
- Description-Description of the talk
- Transcript-Full transcript of the talk.
- Published date-Date the talk was published to TED.com
- event -Event or medium in which the talk was given.
- Native lang Language the talk was given.

Project Story

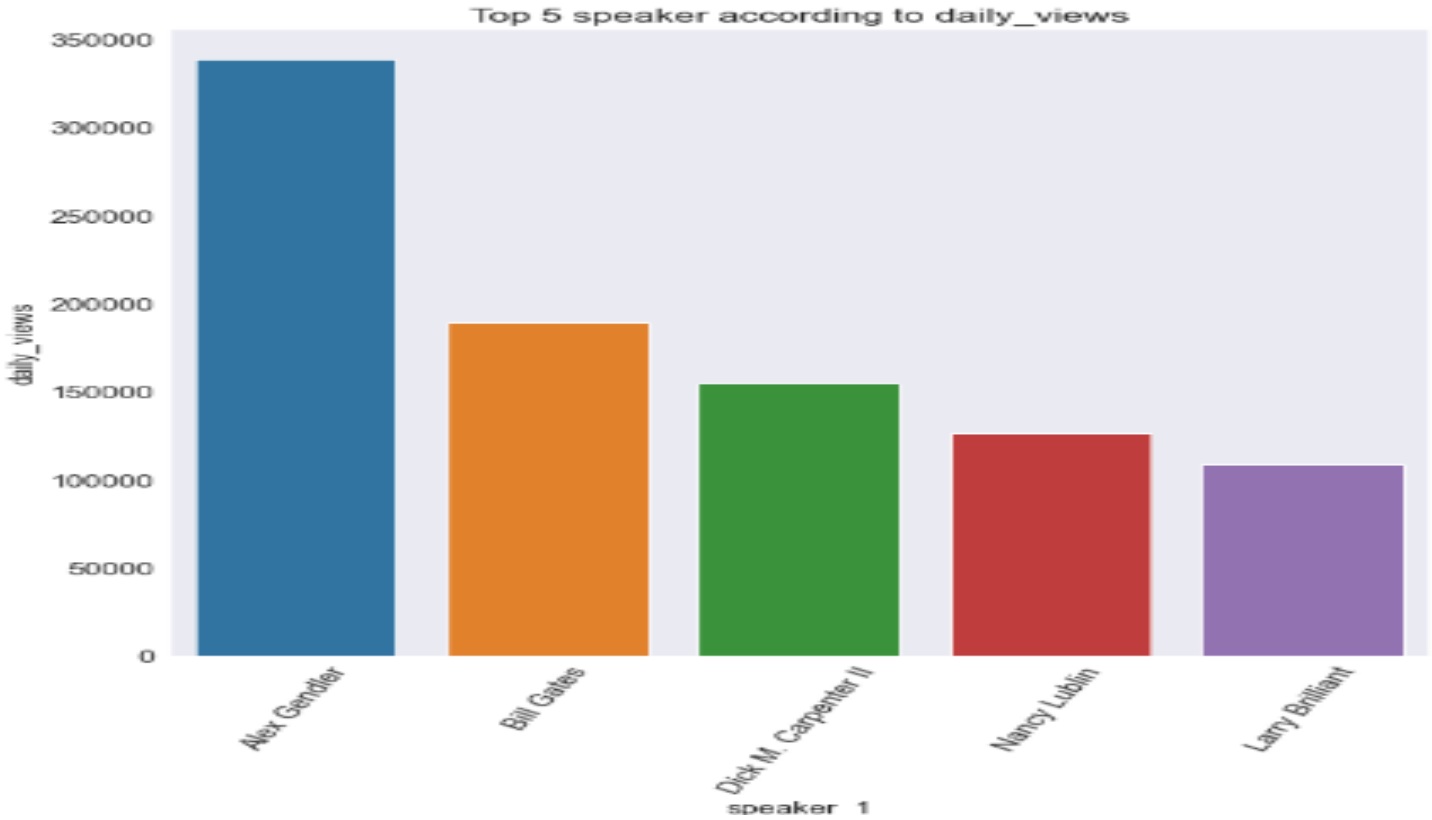
- Data collection and data Understanding.
- Loading to tool and importing Libraries.
- Data Preparation and Data cleaning.
- Data Transformation.
- Feature selection and removal.
- Modeling- importing models, Defining the dependent and independent.
- Model improvement- Regularization.
- Model improvement- Hyper parameter tuning.
- Model Evaluation and selection.
- Deployment.

Visualization



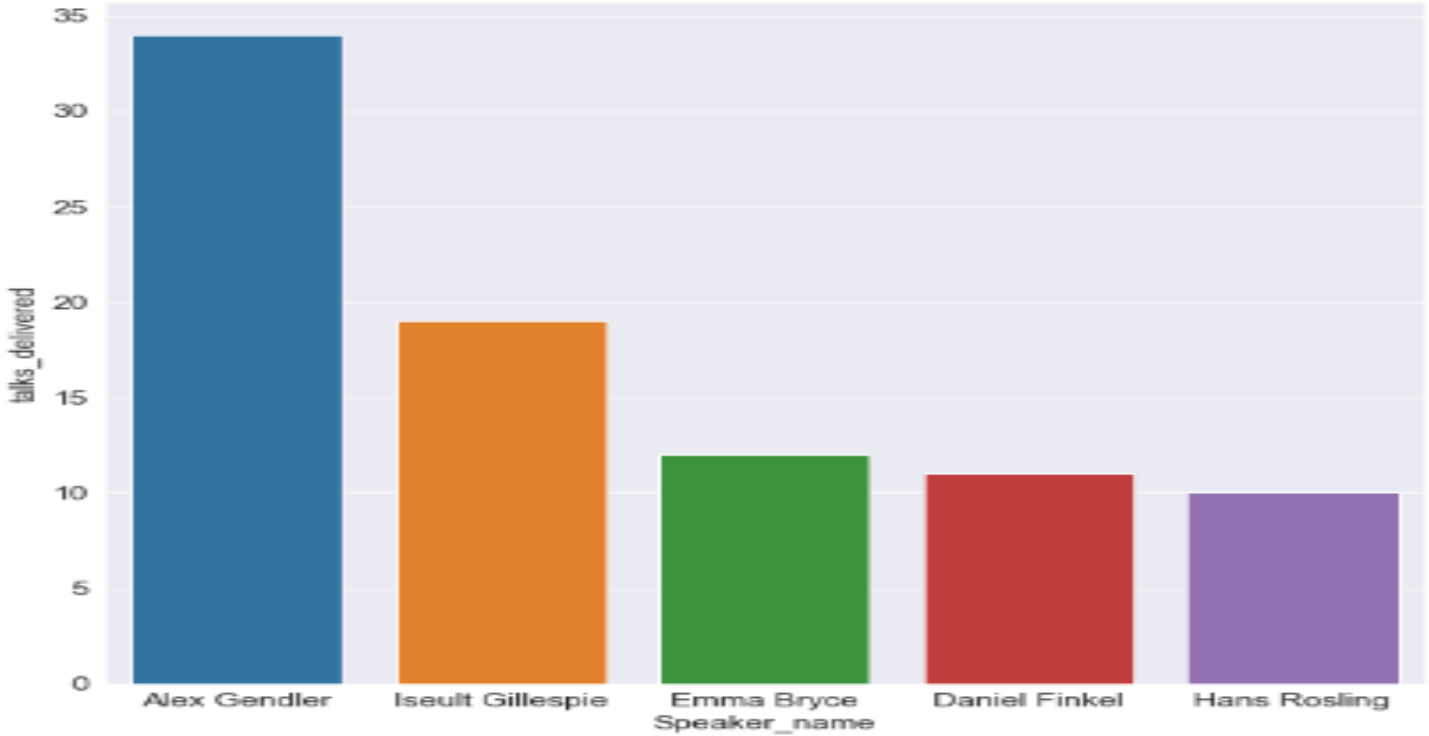
Views are right skewed. Views have positive skewed distribution of data. Mean is greater than mode. Mean is greater than median as well.

EDA-Top 5 speaker according to daily Views



EDA-Most Popular speakers

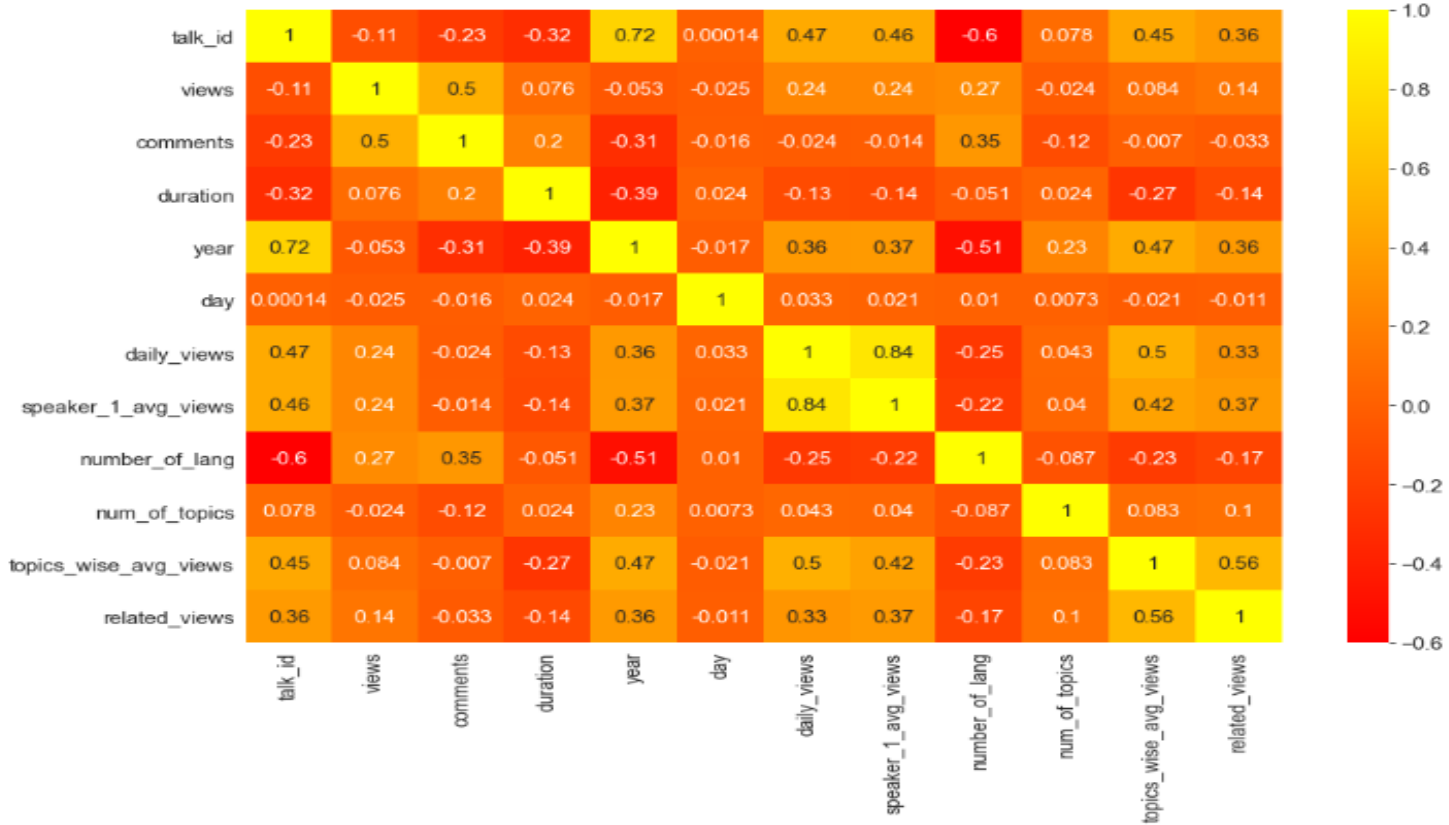
- More talks session delivered shows more views and popularity.



EDA-Data Cleaning- Steps and challenges

- Finding out missing values.
- Finding out duplicates.
- Finding out outliers.
- Data Transformation.
- Feature selection and feature dropping.

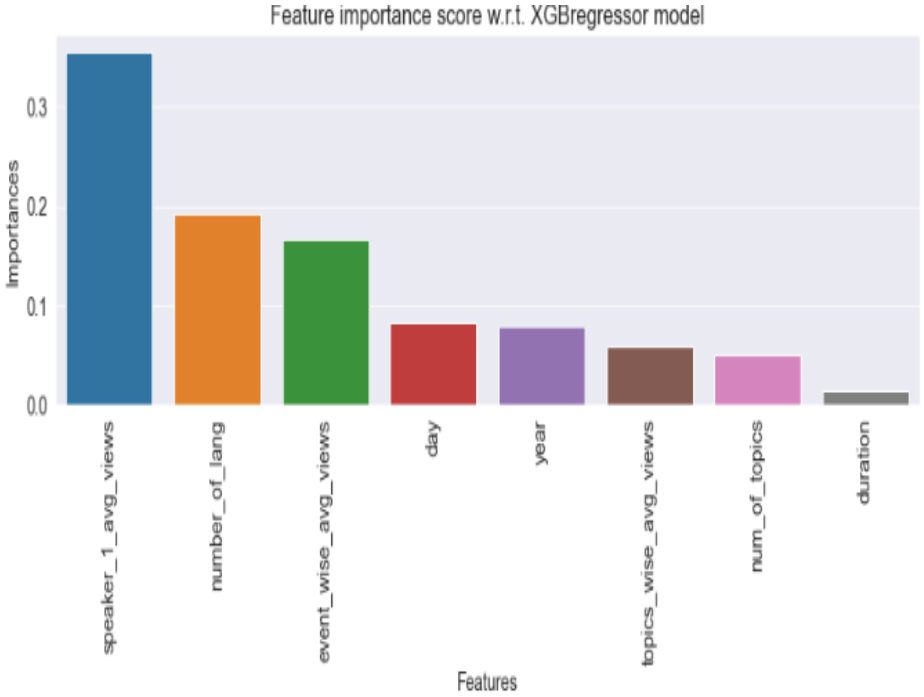
Correlations

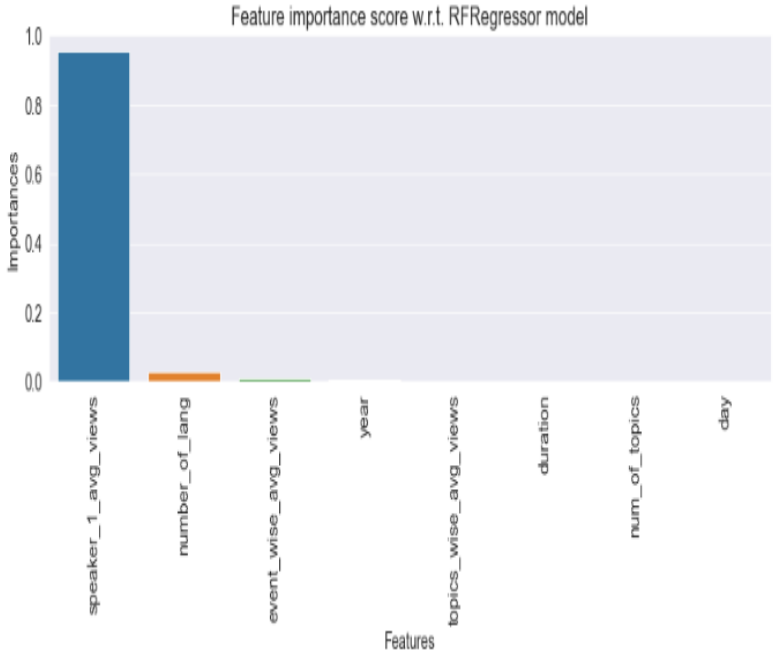


Models

- Linear Regression.
- Catboost.
- XGBoost.
- Random forest.
- Model improvement.

Feature Importance XGBoost and Random forest





Model scores

1)Linear Regression Scores.

- MSE on test is 0.1674690571782963
- RMSE on test is 0.40922983417426484
- Training MAE: 0.18
- Test MAE: 0.17

2)Xgb model Scores.

- xgb_r2_train = 0.9991119271428827
- Xgb_ r2_Test score = 0.8839532921183787
- MAE = 0.11754976639491378
- RMSE_test = 0.34471294420887577
- RMSE_Train = 0.029678836350722874

After tunning Scores

- Training MAE: 0.09
- Test MAE: 0.11
- MAE_train:0.077087
- MAE_test :110052
- MSE test :0.17713819580443496
- R2_Score_train :0.96376
- R2_Score_test : 0.810604
- RMSE_Score_train : 0.194288
- RMSE_Score_test : 0.412989

3)Random forest Scores

- Training MAE: 0.05
- Test MAE: 0.11
- Target mean: 0.00
- Target std: 1.00

After tunning scores

- Training MAE: 0.128898
- Test MAE: 0.11
- RMSE_Score_test : 0.691699
- RMSE_Score_train : 0.667746
- R2_Score_test : 0.468713
- R2_Score_train : 0.571932

4)Cat boost Scores.

- Cat boost R2 Test =0.8182368967755949
- Cat boost R2 Train = 0.994866406672067
- Train_MAE = 0.043066117420744836
- Test_ MAE = 0.14
- Test_MSE - 0.18611787605661576
- Test_MARE =0.1357737499098533
- Train RMSE =0.07135646085702779
- TEST RMSE = 0.43141381069295376

Challenges

- I have faced many challenges during this project starting with looking for converting the data into numerical and standardization also selecting the right amount of features and encoding including target encoding and one_hot_encoding.

- I have made a few observations that keeping right amount of features and right feature selection is very important and finally are responsible for your models betterment and good accuracy.
- Regularization and hyper parameter tuning.
- I have faced some of challenges during converting the str type to the dictionary type column.

Conclusion

I have started with data loading and importing the libraries and then I started with the exploring the data and looking into columns which were used for visualization. I have seen few columns for the null values and outliers as well.
I treated the data accordingly and since I have categorical data with me I did target encoding and one hot ending on given columns. Since some values were varying and causing the heteroskedasticity. I have done standardization accordingly.
I have explored the data and looked for the trend in bi-variate analysis.
I have checked that views are correlated with the speakers as shown top 5 speaker with daily views exceeding 100000.
More talks sessions delivered is showing that more views and popularity also duration is not much influenced when there are popular speakers.
Finally I started modeling and saw XGBoost , cat boost and random forest performed well. For score improvement I have used regularization and hyper parameter tuning. I was able to predict views correctly more than 80% of time.

Thank you..