**Capstone Project Submission**

| **Team Member’s Name, Email and Contribution:** |
| --- |
| 1. Arvind Krishna ([killerdude.arvind@gmail.com](mailto:killerdude.arvind@gmail.com))  * Data Wrangling * Data\_ted\_talks * Loading and Preprocessing * Structuring Data * Enriching Data * Data Validation * Data Mining * Data Analysis * Model Development * Decision Tree Regression * Random Forest Regression * XGBoost Regression * Random Forest Regression with gridSearchCV * Visualizations * Box Plot * Segmentation * Summarization * Observations * Conclusions  1. Sahil Ahuja ([s.ahuja38@gmail.com](mailto:s.ahuja38@gmail.com))  * Data Wrangling * Data\_ted\_talks * Structuring Data * Enriching Data * Data Mining * Data Analysis * Model Development * Linear Regression * Lasso Regression * Ridge Regression * Elastic Net Regression * Visualizations * Dist Plots and Sub Plots * Segmentation * Summarization * Observations * Conclusions  1. Keshav Sharma ([keshav1506sharma@gmail.com](mailto:keshav1506sharma@gmail.com))  * Data Wrangling * Data\_ted\_talks * Structuring Data * Enriching Data * Data Validation * Data Mining * Data Analysis * Model Development * Decision Tree Regression * Random Forest Regression * XGBoost Regression * Random Forest Regression with gridSearchCV * Visualizations * Tableau Visualizations * Segmentation * Summarizations * Observations * Conclusions  1. Jayesh Panchal ([jaypan290497@gmail.com](mailto:jaypan290497@gmail.com))  * Data Wrangling * Data\_ted\_talks * Structuring Data * Data Validation * Data Mining * Data Analysis * Model Development * Linear Regression * Lasso Regression * Ridge Regression * Elastic Net Regression * Visualization * Historical Bars * Segmentation * Summarization * Observations * Conclusion |
| **Please paste the GitHub Repo link.** |
| Github Link:-  https://github.com/Keshav1506/Ted-Talks-views-Prediction-Model  Drive Link:-  https://drive.google.com/drive/folders/13fBUu74uncgHX5sllNEedcXpu5QHQSN7?usp=sharing |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)**   | A TED talk is a recorded public-speaking presentation that was originally given at the main TED (technology, entertainment and design) annual event or one of its many satellite events around the world. TED is a nonprofit devoted to spreading ideas, usually in the form of short, powerful talks, often called "TED talks." TED is dedicated to researching and sharing knowledge that matters through short talks and presentations. Their goal is to inform and educate global audiences in an accessible way.  As a first step we explored through all the columns in the given dataset to understand the rubrics and content of dataset that we have. We tried to check the correlation among various features and found a good correlation between views and comments. But almost 17% of the values in comments were null values. So, we removed the outliers in comments and filled the nan values with the median.  Upon implementation of various algorithms, we were able to see that the linear algorithms were not performing optimally even with Gradient Boosting optimization, and the tree-based algorithms performed significantly better.  Out of the tree-based algorithms, the Random Forest Regressor was providing an optimal solution towards achieving our Objective. We were able to achieve an R2 score of 0.99 in the train split, and 0.92 in the test split. We also noticed that even in the case of Decision tree, we were able to achieve an R2 score of 0.89 in the test split.  We then implemented Grid Search Cross Validation on the Random Forest Regressor, to further optimize the model, and were able to achieve an R2 score of 0.99 in the train split, and 0.95 in the test split.  Finally, we conclude Random Forest with GridSearchCV to be the best model to achieve our objective. | | --- | |  | |
|  |