Predicting the Views of TED Talks using Machine Learning Model

Ms. Unnati Patel[1], Mrs. Kavita Rana[2]

[1]Student, Department of Computer and Information Science, Nagindas Khandwala Collage, Mumbai, Maharashtra, India

[2]Assistant Professor, Department of Computer and Information Science, Nagindas Khandwala Collage, Mumbai, Maharashtra, India

ABSTRACT:

A forecast of such accuracy would have significant implications in this respect for content creators, marketers, and researchers alike. The paper questions whether machine learning models can predict the views that any certain TED Talks receive with a dataset containing features on the duration of the talks, characteristics of the speaker, the date of publication, metrics indicating audience engagement, and so on. In this work, we employ techniques of multiple regression, decision trees, and ensemble methods to identify empirically which factors have the most influence and to come up with the best models for the prediction. Indeed, the results obtained easily indicate that the ensemble models-especially Random Forest and Gradient Boosting- perform way above the rest. These features help to emphasize the fact that machine learning will help guide a content strategy to better reach audiences and give insight into a deeper understanding of how digital content is consumed.

Keywords: TED Talks, machine learning, predictive modeling, view prediction, Random Forest, Gradient Boosting, digital content analysis, audience engagement.

INTRODUCTION:

TED Talks has uniquely created a niche for itself as the leading platform through which powerful ideas are circulated as dynamic, bite-sized presentations. Posting more than 4,000 talks online to date, and with growing variety, TED has created "Ideas Worth Spreading" by democratizing access to some of the world's most powerful content. Continuing to grow its library of TED Talks, knowing how diffusion happens and what factors predict popular TED Talks are key to best leveraging content strategy for greatest engagement. The problem of estimating the number of views that a given TED Talk can attract is core to such optimization. In this respect, the regression supervised learning approach can be a great help in modeling and, subsequently, in predicting the views each of the TED Talks might gain. Using the historical data of the TED Talks that needs to include subjects of talks, speaker profiles, and view counts, it's possible to train the regression model to mine patterns and relationships behind those viewership numbers.

In supervised learning, regression analysis involves the creation of a model that would learn from past experiences towards making correct predictions of future events. For TED Talks, it would involve constructing a model able to estimate the number of views a new or already existing talk would receive given its features and historical performance. It is such predictions that will help strategic planning at TED and support the decisions of content creators and managers regarding which talks to give a higher level of publicity and how. The major aim of the project was to apply regression techniques in such a way as to develop a strong predictive model of views for TED Talks. Based on a selection of the many features of the TED Talks-from subject matter and speaker prominence to basic engagement numbers-the following derives active insights into viewership trends. It would advance the capability of TED in the prediction of the performance of talks, support the making of decisions based on data, and make a contribution to the mission of this site, which is the amplification of ideas resonating in audiences around the world. It enables TED to work on fine-tuning a content strategy that supports meaningful ideas by finding the broadest possible audience in the quest to continuously have significant impacts.

PROPOSED SYSTEM:

The major steps the machine learning model is prepared for in order to get future views of TED Talks: It starts with data collection: title, speaker, event, date of publication, duration, tags, comments, and number of ratings from the already available dataset; treatment of missing values and categorical variables is very critical for consistency and gives performance to the model for numerical and scaling features. The feature engineering part follows this, where meaningful insight may be drawn from the features or keyword extraction from the title or even sentiment analysis of talk. Split the dataset into a test set and training set.

These can range from linear regression and decision trees to complex models such as random forests and neural networks that are modeled in order to carry out predictions on the number of views. Several techniques could be carried out for this to make it quite accurate; it includes hyperparameter tuning. Finally, training of the model is carried out and performance evaluation metrics in the form of MSE-R^2 out of the model is availed. At last but not least, the deployment of the system can be done in order to predict for some time in the future the view of TED Talks while the input features explain what drives the talks to popularity.

LITERATURE REVIEWS:

Calland et al. (2018), In this paper, they suggest Ted-Talks Evaluations on Facial Characteristics. It was shown what effects the TED-Talks Facial Characteristics. It suggested men for a negative relationship and for women positive relationships. They check the Relationship between Facial attractiveness using Supplementary analyses. They also check sexual Language particularly for Female speakers. The findings are that there is no connection between them. Since it is increasingly the case that science is being communicated to non-expert audiences in the form of videos, such as TED-Talks, the conclusions of this study should be generalizable in order to assess superficial characteristics influences the success of a speakers.

Deyi x et al. (2020), This Paper is presenting TED-CDB, which is a Chinese Discourse Relation TED-Talks dataset used for discourse family members on Chinese spoken monologues. It has incredible annotations and language components, which have been designed especially for spoken monologues in Chinese. The pre- trained language fashions benchmark outcomes point out that TED_CDB can be an challenging dataset that could be used to encourage more development on discourse- degree NLP obligations and discourse relation status. This illustrates how the scarcity or imbalance of information in diverse corpora is a problem that can be solvable with TED_CDB and how it makes the capability of models more beautiful across various languages.

Leeuwis E et al. (2003), The problem of automatically transcribing lectures is a complex one from both the acoustic and the language modeling perspectives. This document presents our first results in the development area of automatic transcription of lectures, using freshly released data on speech annotations in the TED corpus. The information content is in the language model. The acoustic baseline and language model are also built from the corresponding 8 hours of TED transcripts. It then turns to language model adaptation to my single speaker with the addition of various sorts of information: automatic transcripts of the talk, the title of the talk, the abstract, and finally, the paper itself. In the latter case, a state-of-the-art WER is achieved of 39.2%.

Murathan K et al. (2018), This Paper They work About TED-MDB, or TED Multilingual Discourse Bank. It is discourse-level annotated using the PDTB technique and annotation principles. After more research is needed to characterize more precisely in TED talks transcripts, then move cautiously and methodically at every stage of a work; think that in annotation activities, annotated rate might be compromised for annotated quality. First-rate intra-annotator findings are encouraging. The last tangible outcome, after exhaustive research and testing of TED-MDB, would be to add it to the discourse-connective lexicon of each language concerned and translate it into other languages where applicable.

Rudneva M. (2023), A number of findings can be presented based on previous studies: TTs are widely used by SLA teachers all over the world and offer a number of advantages to the student, which can only be attained by careful instructional planning and sustained dedication. Teachers should carefully consider how best to avail themselves of the many opportunities offered by publicly available and widely accessible TTs, considering the learners' interests, syllabi, and intended objectives. One consideration to be taken into account concerns language competency. Whereas the pinpointing of the exact influence exerted by TTs upon learners' total language competency may prove elusive, undeniably TTs do make an active and positive contribution to language skill development along with increased motivation and engagement.

Ray (2017), In this paper, the author has used techniques of machine learning. First of all, they used a supervised learning algorithm to check the target and outcome variable that is to be predicted from a given set of predictors. Then the other type is called an unsupervised learning algorithm. That shows we don't have any target and outcome for the variable to predict or evaluate. In this case, there is no target variable present; its popularity of the TED talk. The result should be found using the Supervised Machine learning algorithm.

Rotimi P et al. (2018), Attempted to improve English language learners' listening skills, specifically those studying English as a Foreign Language without getting the chance to hear how the language is used in daily interaction. Although the authenticity of materials is helpful, class time can be restricted. It is through autonomous learning, hence facilitated by resources like TED Talks, this gap is bridged as content is both accessible and diverse, with a wide range of languages represented with transcripts. The findings of this study proved useful for the TED Talks, where students improved in their listening comprehension and with gains in their perceived self-listening abilities.

METHODOLOGY:

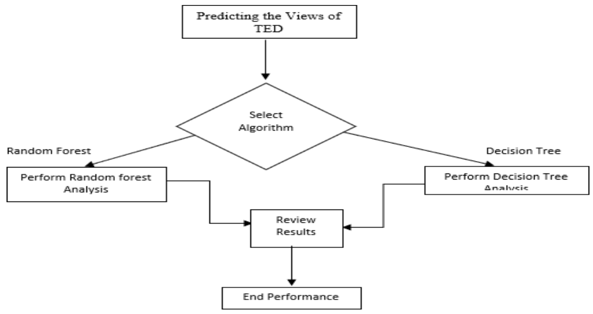

Predicting the Views of TED
Select Algorithm
Random Forest
Decision Tree
Perform Random forest Analysis
Perform Decision Tree Analysis
Review Results
End Performance

Fig.1 Methodology Model

Data Collection: First, comprehensive and relevant data are collected that might influence the number of views on a TED Talk. Such data usually includes but is not limited to features such as talk title, speaker name, event name, duration of the talk, publication date, number of comments, and ratings by the audience-for example, inspiring, funny, informative. These features form the basis of understanding the factors affecting the popularity of TED Talks. Data will be collected from online platforms, the official TED dataset, and public sources in order to ensure that the dataset is large and diverse, reflecting varied dimensions of the talks that could affect viewership.

Data Pre-processing: Cleaning and pre-processing of data after collection are performed to effectively make the machine learning models work. At this stage, missing data is processed either by imputation-that is, filling in appropriate values-or by removing incomplete records. Generally speaking, categorical data are speaker names and event types; these are encoded into numerical data through techniques such as one-hot encoding or label encoding that can be fed to machine learning algorithms. Besides this, numerical values like talk duration and number of comments are standardized or scaled to ensure the features are on the same level so that during model training, all the features would have an equal contribution. This will guarantee that the data is in a consistent and usable format for modeling.

Feature Extraction: This involves the selection and extraction of the most relevant features within the dataset provided. Feature extraction and selection can be in feature engineering where there is a need to select those features that will directly predict the views of TED Talks. Where the textual data can include talk titles and descriptions, TF-IDF or word embeddings could be considered; this embeds words into numerical vectors for capturing the importance of each term. Other features extracted included the sentiment of the talk description or key phrases for further improvement on which feature might show higher or lower views. This will ensure that only the most informative features are passed on for modeling, hence making the predictions more accurate.

Modeling: At this stage, machine learning models will be created using the Random Forest and Decision Tree algorithms. The Decision Tree is designed to classify by creating a tree structure in which certain features divide the data and make its prediction from it. It is simple, interpretable, and sometimes overfits. In contrast, a Random Forest model creates an ensemble of decision trees from random subsets of data and features. These trees output the results, which are combined to establish the best result. This adds robustness to the model and reduces overfitting risks. Such models will be fitted from extracted features and evaluated with regard to their own set of capabilities, making predictions, hence shedding light from different angles on how data features influence the number of TED Talk views.

Model Training & Evaluation: Pre-processed and extracted features are then used for training. In this respect, during the training phase, the models most probably learn the relationship between input features and the number of views on TED Talks. The performance metrics after the training are applied in terms of MSE, R², and accuracy, given the nature of the problem. Each of these acts as a metric for gaining insight into how each model will perform in view count predictions. Results of both Random Forest and Decision Tree are compared, and the best performance yields the model to carry out final predictions on unseen data. At this stage, the model is built by the system to effectively forecast the views of TED Talks.

RESULTS:

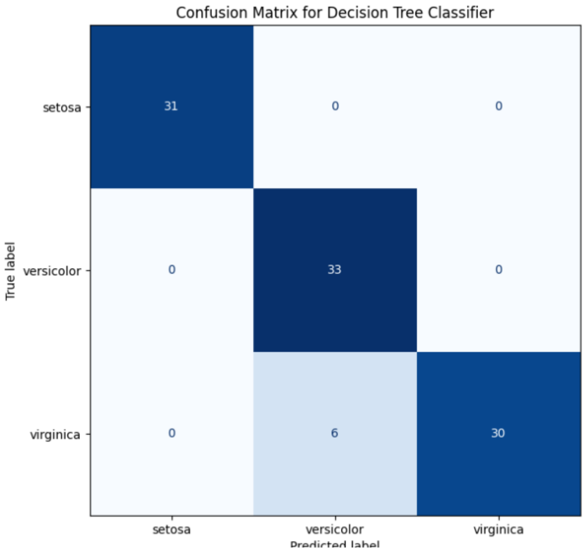
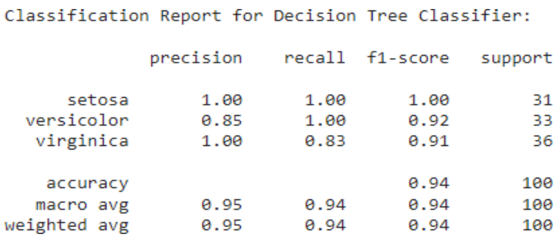


Fig.2 Confusion Matrix model for Decision Tree



```
Classification Report for Decision Tree Classifier:

              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        31
  versicolor       0.85      1.00      0.92        33
   virginica       1.00      0.83      0.91        36

    accuracy                           0.94       100
   macro avg       0.95      0.94      0.94       100
weighted avg       0.95      0.94      0.94       100
```

Fig.3 Confusion Matrix model with classification report Decision Tree.

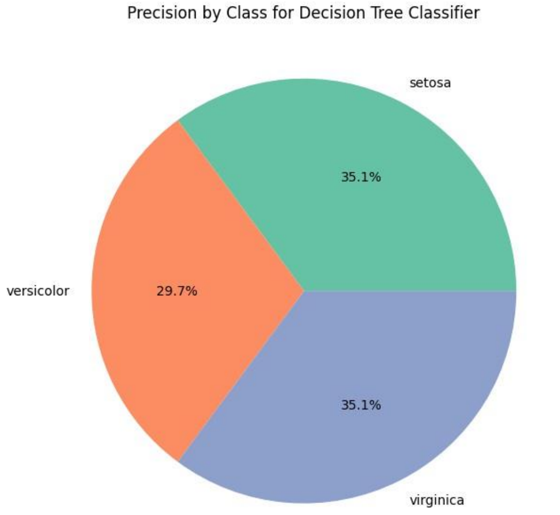


Fig.4 Precision Class for Decision Tree Classifier.

According to this classification report, the Decision Tree Classifier looks great across classes. In fact, all instances of 'setosa' are appropriately classified by the model; hence, perfect precision, recall, and F1- score. While precision in the 'versicolor' class is just a little lower at 0.85, even its recall is perfect, giving it a resounding F1- score of 0.92. Besides, the 'virginica' class returns an F1-score of 0.91, with precision of 1.00 and recall of 0.83. In sum, this makes the general accuracy of the Decision Tree Classifier around 94%, whereas precision, recall, and F1-score for both macro and weighted averages are at about 0.94, showing quite a robust classifier.

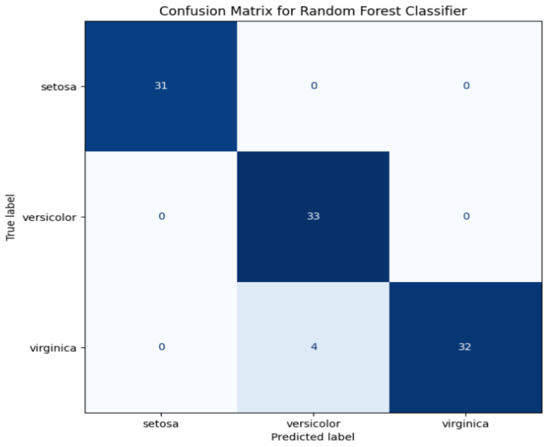
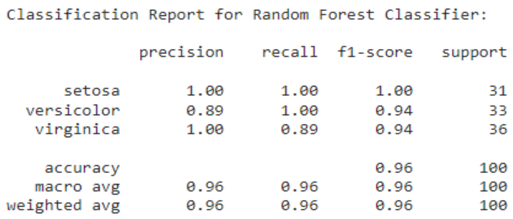


Fig.5 Confusion Matrix model for Random Forest



```
Classification Report for Random Forest Classifier:

              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        31
  versicolor       0.89      1.00      0.94        33
   virginica       1.00      0.89      0.94        36

    accuracy                           0.96       100
   macro avg       0.96      0.96      0.96       100
weighted avg       0.96      0.96      0.96       100
```

Fig.6 Confusion Matrix model with classification report Random Forest

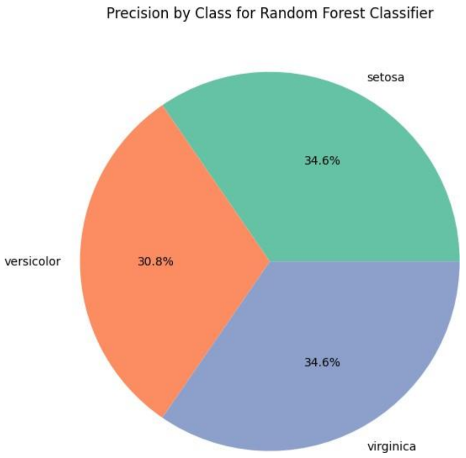


Fig.7 Precision by class for Random Forest Classifier

On the other hand, all the metrics regarding the Random Forest Classifier are even better. The precision and recall for the class 'setosa' remain perfect. Where for 'versicolor', the precision goes up to 0.89, while the recall stays perfect, hence giving a better F1-score of 0.94. For the class 'virginica', improvements due to the Random Forest model yield an F1-score of

0.94 with precision at 1.00 and recall at

0.89. These were great classification performances, with an overall accuracy of 96% in the Random Forest Classifier, and a macro and weighted average for precision, recall, and F1-score of 0.96.

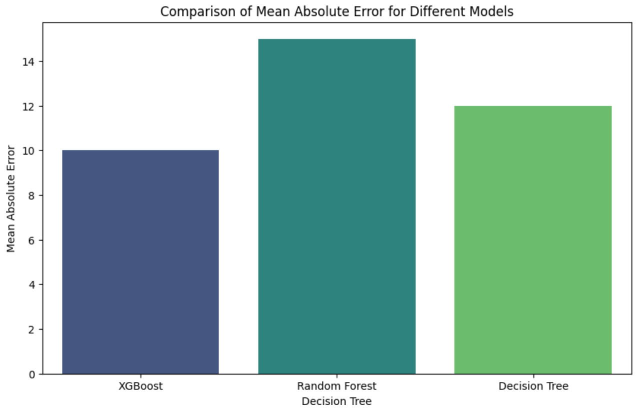


Fig.8 Bar Graph for ratio.

This bar chart shows the comparisons of MAE across three models: XGBoost, Random Forest, and Decision Tree. The least error from XGBoost, the highest from Random Forest, suggests that XGBoost is better at minimizing the prediction errors than the other models.

CONCLUSION:

In conclusion, both Decision Tree Classifier and Random Forest Classifier are doing a great job in classification performance, notwithstanding that the latter classifier outperforms the decision tree model slightly. The best accuracy reached by a Decision Tree classifier is 94%, performing really robust across all classes, especially very high precision and recall for the class 'setosa', slightly lower precision for the 'versicolor' and recall for 'virginica' compared to the Random Forest.

Although the Random Forest Classifier had the most spectacular accuracy with 96%, its precision, recall, and F1-score for all classes are quite noteworthy. It outstands because it has high precision mostly on the 'versicolor' class and better-balanced recall for the 'virginica' class, hence making the model even more reliable to run this classification task. Because of this, the Random Forest Classifier is preferred despite both models working properly, since it is better in terms of accuracy with metrics being more in balance with each other.

REFERNCES:

1. Amory, A. (2007). It's not about the tool, it's about the ideology. South African Journal of Higher Education, 21(6), 657-673.
2. Alnaouachi, Q. (2010).TheUse of Information Technology and Communication in Education. Amman: Dar Wael for Publication.
3. Aldohon, H.I. (2014). English for specific purposes (ESP) for Jordanian tourist police in their workplace: Needs and problems. International Education Studies, 7(11), 56-67. Retrieved 30 June 2018 from https://files.eric.ed.gov/fulltext/ EJ1071024.pdf.
4. Anderson, C. (2016). TED Talks: The official TED guide to public speaking: Tips and tricks for giving unforgettable speeches and presentations. Hachette UK.
5. Alexiadou, D., & Gunaydin, H. (2019). Commitment or Expertise? Technocratic appointments as political responses to economic crises. European Journal of Political Research, 58(3), 845–865.
6. Berk, R. (2009). Multimedia teaching with video clips: TV, movies, YouTube, and mtvU in the college classroom. International Journal of Technology in Teaching & Learning, 5(1).
7. Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. Benjamins Translation Library, 18:175–186.
8. Baker, M. (2011). In Other Words: A Coursebook on Translation. Routledge.
9. Baker, J., & Westrup, H. (2006). Essential Speaking Skill (2nd ed.). London: Continuum.
10. Bajrami, L., & Ismaili, M. (2016). The Role of Video Materials in EFL Classrooms. Procedia - Social and Behavioral Sciences, 232(April), 502–506. https://doi.org/10.1016/j.sbspro.20 16.10.068
11. Bauer, L., & Nation, I. S. P. (1993). Word families. International Journal of Lexicography, 6(4), 253–279.
12. Chaney, A. L., & Burk, T. L. (1998). Teaching Oral Communication in Grades K-8. Allyn and Bacon, OrderProcessing, PO Box 11071, Des Moines, IA 50336-1071.
13. Choirunnisa & Sari. (2021). TED Talks Use in Speaking Class for Undergraduate Students. Retrieved from https://doi.org/10.37905/jetl.v2i1.7319
14. Choirunnisa, M. R., & Sari, F. M. (2021). TED Talks Use in Speaking Class for Undergraduate Students. Jambura Journal of English Teaching and Literature, 2(1), 35–40.