Predict the Popularity of a TED Talk

Shijie Liu
Computer Science Department
Hood College, Frederick, Maryland 21701, USA
sl23@hood.edu

ABSTRACT

TED (Technology, Entertainment, and Design) Talks are being posted by TED Conference LLC for free on their website and YouTube Channel under the slogan of "ideas worth spreading". TED Talk offers a wide range of topics within the research and practice of science and culture, and often through storytelling. It has a variety of presenters as well, such as Writer, Researcher, and Scientist. TED also gives people the right to raise their own local TEDx Event. How could TED filter and determine which talk shall be published? What an organizer could do to make his/her talk more popular? In this project, I would like to apply Exploratory Data Analysis on all the available attributes and help people to have a better understanding of what are the variables that could affect the popularity of a TED Talk. In addition, applying Natural Language Processing technique on the TED Talk transcript, and then train a classification model with algorithms that could predict the popularity of TED Talk.

CCS CONCEPTS

Computing methodologies → Machine learning → Machine learning approaches → Classification and regression trees

KEYWORDS

Exploratory Data Analysis, Visualization, Classification, Natural Language Processing

1. INTRODUCTION

TED Talks include talks on scientific, cultural, and academic topics, and the speakers also widely spread with different roles, such as, scientists, education researchers, businessmen, artists, etc. [1] Until October 2018, there are approximately 2,900 TED Talks freely available on the TED website [2]. Like Charlie Rose said, who is an American television journalist and former talk show host, TED Talks has become one of the most powerful platform because they are spreading ideas through the stories of remarkable people and they could be supported world widely with different languages for transcripts [3], which is another reason why TED Talks are so popular.

Today, the speed of data growth is extremely fast. According to IBM, there is 2.5 quintillion bytes of data created every day [4]. Everything is in a format of data, as well as TED Talks. A lot of people have already done researches on data of TED Talk. Hong et al. did a visual analysis of TED Talk topic trends, which visualize the relationship between talks and playlist, also used keywords to show the talks' relativity. [9] Another research being conducted by Oh et al. they built a recommender based on speech transcript by applying TF-IDF analysis and applying *Dos2vec* of the *Gensim* package to derive vectors of transcripts [10]. Another interesting project on TED talk is from Cullen and Harte [18], they built a predictive model that could predict the viewer impression on a talk based on video thin slicing. They pointed out that visual features are important for both audience engagement and emotion perception that they used algorithms to track face and hand movement, then trained a linear SVM to predict the viewers' impression. [15] [16] Therefore, when the data is large and multidimensional, it is practically impossible to get a potentially interesting and actionable insight without the help of suitably designed machine learning algorithms [5]. More importantly, machine learning is everywhere in today's Natural Language Processing, the goal of deep learning is to explore how computers can take advantage of data to develop features and representations appropriate for complex interpretation tasks [6].

2. BACKGROUND

TED Talk is my favorite program because of its diverse contents that cross different fields and creative ideas. One day, when I was browsing TED Talk's website, there was a question pop-up automatically, "what interests you?" With following choices: Technology, Science, Innovation, and Humanity, etc. After selecting, another question came up, "what you're looking for?" With following answers: Professional Growth, Inspiration or motivation, and Smart entertainment, etc. I was so curious about why they are asking me those questions.

After selecting my interested topic and idea, there was one recommended TED talk pop-up to me along with a sentence saying, "This idea offers 'professional growth' and matches your interest in 'innovation'". I was so satisfied that I can just let the website know my interest and then the website will find me a recommended video that matches my interest! I feel like I have a 'free' assistant. From the TED website, I also learned that people have the right to raise Local TEDx Event as they wish. How TED filter the talk topics to be published? How organizer could make their talk more popular? I really want to use Exploratory Data Analysis technique to extract information from large dataset and present it in a comprehensible way. And I want to use classification algorithms to train a model that could predict the future of an unpublished TED Talk.

3. OBJECTIVES

- **What makes a TED Talk more popular? What could an Organizer do to get more views on their talks?**
- **What are the most popular topics in TED Talks?**
- **Predict the popularity of an un-published TED Talk based on given transcripts.**

4. RELATED WORK

4.1 Data Collection:

4.1.1 Data Selection:

The TED Talk datasets I am using for this project were downloaded from Kaggle, which has approximately 2500 talks available. According to the dataset uploader, these datasets contain information about all audio-video recordings of TED Talks uploaded to the official TED.com website until September 21st, 2017. The TED Talk main dataset contains 17 columns, including number of views, number of comments, descriptions, speakers and titles, etc. The TED Talk transcripts dataset contains 2 columns, including the URL and the available transcripts [7].

4.1.2 Data Description:

4.1.2.1 TED Main Dataset feature descriptions:

- **Int64 - comments**: *The number of first level comments made on the talk*
- **Object - description**: *A blurb of what the talk is about*
- **Int64 - duration**: *The duration of the talk in seconds*
- **Object - event**: *The TED/TEDx event where the talk took place*
- **Int64 - film_date**: *The Unix timestamp of the filming*
- **Int64 - languages**: *The number of languages in which the talk is available*
- **Object - main_speaker**: *The first named speaker of the talk*
- **Object - name**: *The official name of the TED Talk. Includes the title and the speaker*
- **Int64 - num_speaker**: *The number of speakers in the talk*
- **Int64 - published_date**: *The Unix timestamp for the publication of the talk on TED.com*
- **Object - ratings**: *A stringified dictionary of the various ratings given to the talk (inspiring, fascinating, jaw dropping, etc.)*
- **Object - related_talks**: *A list of dictionaries of recommended talks to watch next*
- **Object - speaker_occupation**: *The occupation of the main speaker*
- **Object - tags**: *The themes associated with the talk*
- **Object - title**: *The title of the talk*
- **Object - url**: *The URL of the talk*
- **Int64 - views**: *The number of views on the talk*

4.1.2.2 TED Transcript Dataset feature descriptions:

- **Object - transcript**: *The official English transcript of the talk.*
- **Object - url**: *The URL of the talk*

4.1.3 Dataset Sample Data:

4.1.3.1 TED Main Dataset:

comments	description	duration	event	film_date	languages	main_speaker	name	num_speaker	published_date	ratings	related_talks	speaker_occupation	tags	title	url	views
4553	Sir Ken Robinson makes an entertaining and pro...	1164	TED2006	1140825600	60	Ken Robinson	Ken Robinson: Do schools kill creativity?	1	1151367060	[{'id': 7, 'name': 'Funny', 'count': 19645}, {...	[{'id': 865, 'hero': 'https://pe.tedcdn.com/im...	Author/educator	['children', 'creativity', 'culture', 'dance', ...	Do schools kill creativity?	https://www.ted.com/talks/ken_robinson_says_sc...	47227110

Figure 1: Sample Data Entry in the TED Main Dataset

4.1.3.2 TED Transcript Dataset:

transcript	url
Good morning. How are you?(Laughter)It's been great, hasn't it? I've been blown away by the whole thing. In fact, I'm leaving.(Laughter)There have been three themes running through the conference which are relevant to what I want to talk about. One is the extraordinary evidence of human creativity in all of the presentations that we've had and in all of the people here. Just the variety of it and the range of it. The second is that it's put us in a place where we have no idea what's going to happen, in terms of the future. No idea how this may play out.I have an interest in education. Actually, what I find is everybody has an interest in education. Don't you? I find this very interesting. If you're at a dinner party, and you say you work in education — Actually, you're not often at dinner parties, frankly.(Laughter)If you work in education, you're not asked.(Laughter)And you're never asked back, curiously. That's strange to me. But if you are, and you say to somebody, you know, they say, "What do you do?" and you say you work in education, you can see the blood run from their face. They're like, "Oh my God," you know, "Why me?"(Laughter)"My one night out all week."(Laughter)But if you ask about their education, they pin you to the wall. Because it's one of those things that goes deep with people, am I right? Like religion, and money and other things. So I have a big interest in education, and I think we all do. We have a huge vested interest in it, partly because it's education that's meant to take us into this future that we can't grasp. If you think of it, children starting school this year will be retiring in 2065. Nobody has a clue, despite all the expertise that's been on parade for the past four days, what the world will look like in five years' time. And yet we're meant to be educating them for it. So the unpredictability, I think, is.....	https://www.ted.com/talks/ken_robinson_says_schools_kill_creativity

Figure 2: Sample Data Entry in the TED Transcript Dataset

4.2 Data Understanding:

Technique: *Exploratory Data Analysis*

For this project, the technique being used to understand all the features of the provided dataset is Exploratory Data Analysis (EDA), which involves a number of graphical techniques, such as, Box Plot, Histogram, Multi-vari chart, and Scatter Plot, etc.[11]. The primary aim with Exploratory Data Analysis is to examine the data for distribution, outliers, and anomalies. It also

provides hypothesis generation by visualizing and understanding the data usually though graphical representation. [12].

Although not all algorithms will fail with missing values, it is still recommended to understand and mark where those missing values are and handle missing values accordingly. Some missing values can be replaced with different values, while some rows shall be dropped from the dataset. Figure 3 is a missing data visualization for TED main dataset that being presented with Library – *missingno*, from MIT [13].

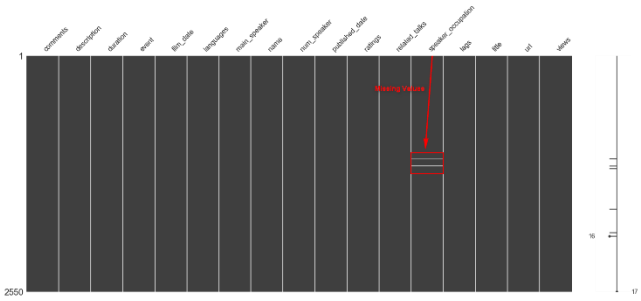

Figure 3: Missing Data Visualization in TED Main Dataset

Since the objective of this project is to predict the popularity a TED talk and TED.com published their "The most popular talks of all time" at TED.com, which is sorted by number of views. Consequently, I would also use "views" as the basis definition of the popularity. Figure 4 shows the distribution of 'views'.

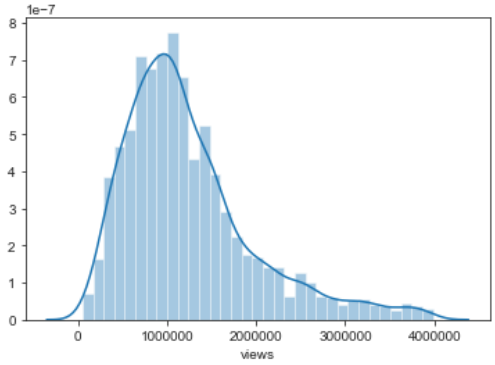

Figure 4: Distribution of number of views for each talk

According to Figure 4, 'views' is really wide dispersed that predicting the exact number of views can be super difficult. However, the density is pretty high for views around 1 million, it would be more sense to discretize the views and make it a categorical or binary value. For this project, Figure 5 defines how the popularity is being defined. In addition, it is possible that celebrity charm may attract additional views, for example, Bill Gates had given several talks at TED.com. Therefore, I used *barplot* in Figure 6 to show the TOP 10 talks and perform spot check on the speakers to make sure they are not celebrities.

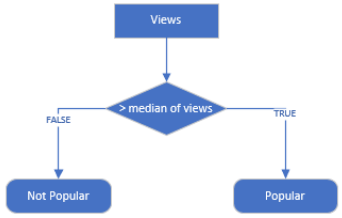

Figure 5: Define the popularity

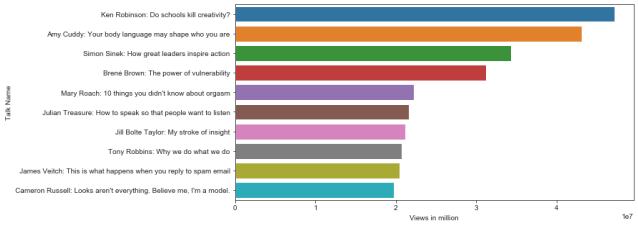

Figure 6: TOP 10 TED Talks (2017)

Figure 7 below shows the occurrences of unique values based on *main_speaker* from TED main dataset, some speakers have more than 1 TED talk being published, hypothetically, these occurrences is also kind of experiences, therefore, I introduced a new feature – *number_of_attendences* – to our TED main dataset.

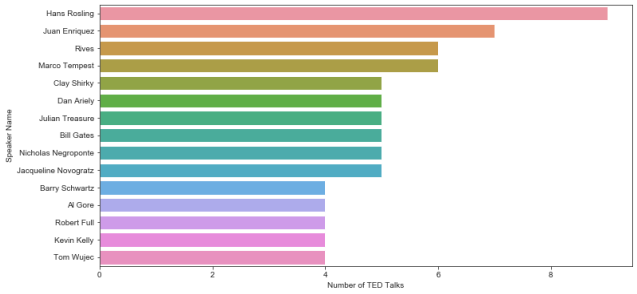

Figure 7: Speaker Occurrences on TED Talk

Figure 8 below shows how number of views is related with *speaker_occupation* with boxplot. *Speaker_occupation* is an object feature in TED main dataset. Converting this feature to a categorical feature would be necessary while building classification model.

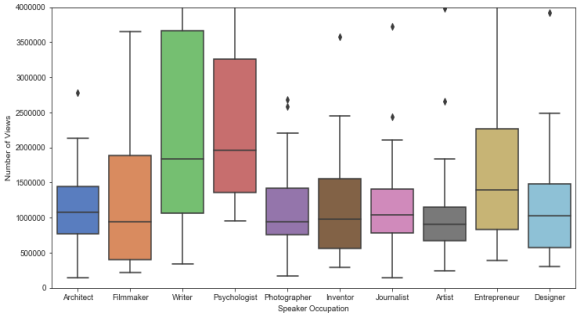

Figure 8: Boxplot for Speaker Occupation and Views

4.3 Data Pre-Processing:

4.3.1 Data cleaning:

As we identified in Data Understanding Section, there are 6 missing values in *speaker_occupation* column and this column is an Object type, the way how this project handles this missing value is to fill missing values with "**Unknown**".

4.3.2 Data Transformation:

In the TED main dataset, there are several columns could not be used directly in a meaningful way. For example, date time is being presented in UNIX Epoch Format, ratings matrix is embedded as one object. Tags is also presented as one string contains a list array. In order to understand those features properly and use them in training the classification model, I have taken following actions to transform the format of those 'meaningless' values. Figure 9 shows how the film_date and published_date being transformed, and Figure 10 shows how the ratings being converted from a string to a matrix table.

```
1 tedMain.loc[0:5,['film_date','published_date']]
```

	film_date	published_date
0	1140825600	1151367060
1	1140825600	1151367060
2	1140739200	1151367060
3	1140912000	1151367060
4	1140566400	1151440680
5	1138838400	1151440680

	film_date	published_date
0	2006-02-24	2006-06-26
1	2006-02-24	2006-06-26
2	2006-02-23	2006-06-26
3	2006-02-25	2006-06-26
4	2006-02-21	2006-06-27
5	2006-02-01	2006-06-27

Figure 9: Convert UNIX Timestamp to Human date

Figure 10: Convert ratings from string object to matrix table

After formalizing the *dates*, *ratings*, and *tags*, I extracted Published/Film year, Published/Film month and Published/Film weekday from date. Figure 11 below shows the number of talks being published/filmed per year, month, and weekday. Further, I also did boxplots on views for publish weekday (Figure 12), publish month (Figure 13), and topic (Figure 14).

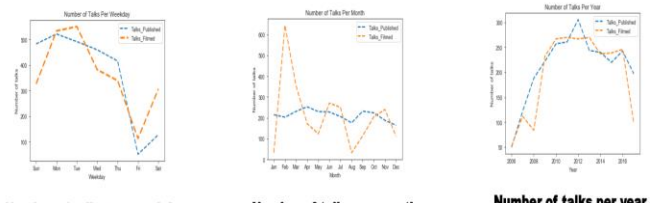

Figure 11: Number of Talks per Year, Month, and Weekday

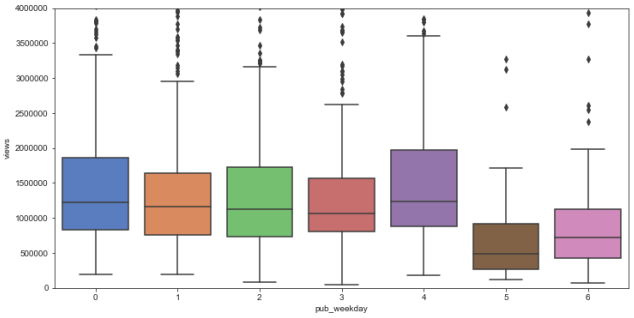

Figure 12: Boxplot for Publish Weekday and Views

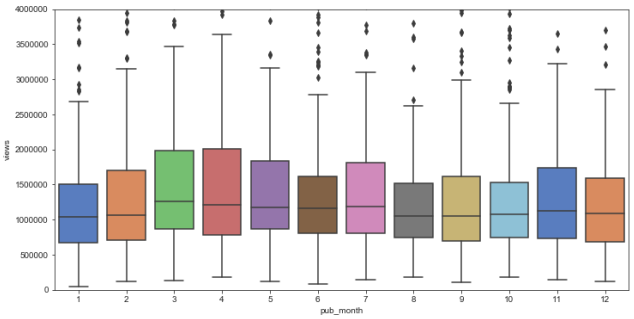

Figure 13: Boxplot for Publish Month and Views

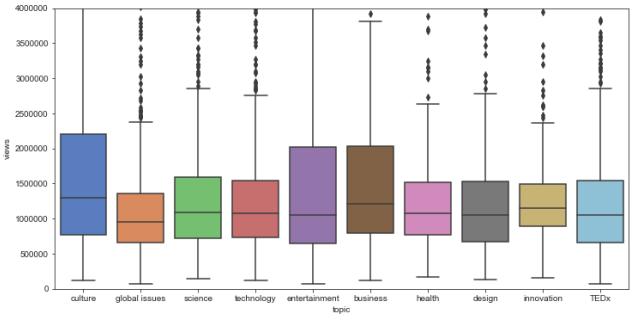

Figure 14: Boxplot for Topic and Views

In addition, there are also columns need to be normalized in the TED main dataset. For example, the duration is the length of each talk, which is in seconds that the value distribution of this feature is widely dispersed (it has 1083 unique values out of 2467 talks). However, the standard measurement shall be minutes, I have converted all duration from seconds to minutes in this project. Another feature needs normalization is event. According to TED Wikipedia, TED Talk has different event types, such as TED Conference, TED Global, TEDx, and TED Women etc. It is reasonable to categorize the event column to a categorical class. Figure 15 shows how *event* being converted.

Figure 15: Categorize the event type

In the TED Transcript dataset, the transcript is just the raw English transcript of the TED, all the greetings and background voices in the talk are all part of the transcript, so it will definitely

need clean-up on this column. The general text preprocessing includes tokenization, stop-word removal, lowercase conversion and stemming. [20] Figure 16 is a code snippet that being used to clean up the transcript document that removes special characters and word lemmatization, *CounterVectorize* is being used to convert a text document to a token matrix that each word is represented as a token and number of occurrences in that document is the token value and English stop words are also get eliminated at the same time.

```
# function to clean up the transcript
def textCleanup(document):
    stemmer = SnowballStemmer("english")

    # Remove all the special characters
    document = re.sub(r'\W', ' ', str(document))

    # remove all single characters
    document = re.sub(r'\s+[a-zA-Z]\s+', ' ', document)

    # Substituting multiple spaces with single space
    document = re.sub(r'\s+', ' ', document, flags=re.I)

    # Removing prefixed 'b'
    document = re.sub(r'^b\s+', '', document)

    # Converting to lowercase to reduce duplicates
    document = document.lower()

    # Lemmatization
    document = document.split()

    # Lemmatization
    document = [stemmer.stem(word) for word in document]
    document = ' '.join(document)

    # remove those background words
    document=document.replace('(laughter)', ' ')
    document=document.replace('(applause)', ' ')

    return document
```

Figure 16: Text Document Clean-up

4.3.3 Data Reduction and Data Integration:

In the TED main dataset, there is one redundant column can be removed because it is a combination of another two columns. For example, *name = main_speaker: title*, see Figure 17 for details. Further, both TED main dataset and TED transcript dataset have the column *url*, these two datasets are joined to one while training the classification model for transcript.

```
In [15]: tedMain.loc[0:3,['name','main_speaker','title']]
```

Out[15]:

	name	main_speaker	title
0	Ken Robinson: Do schools kill creativity?	Ken Robinson	Do schools kill creativity?
1	Al Gore: Averting the climate crisis	Al Gore	Averting the climate crisis
2	David Pogue: Simplicity sells	David Pogue	Simplicity sells
3	Majora Carter: Greening the ghetto	Majora Carter	Greening the ghetto

Figure 17: Redundant Columns

4.4 Model Building:

4.4.1 Feature selection:

Based on the Exploratory Data Analysis and the data type, there are columns would not be used for current project that would be dropped from the dataset while training the model. Such as*description*, *main_speaker*, *speaker_occupation*, and *url*, etc. Realistically, more columns shall be used for model training, but that part has been included as part of Future Work due to the time constraints.

Figure 18 is the heat-map being generated based on remaining columns to check the correlation between columns. Further, I also visualized the correlation between columns with regression visualization, which can be found in Figure 19.

4.4.2 Dataset spilt

The technique being used for splitting dataset is *train_test_split* with 80% as training dataset and 20% as test dataset. The training set contains a known output and the model learns on this data in order to be generalized to other data later on [14].

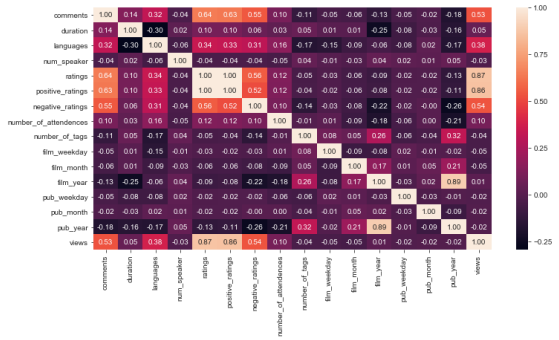

Figure 18: Heat-map - EDA selected features

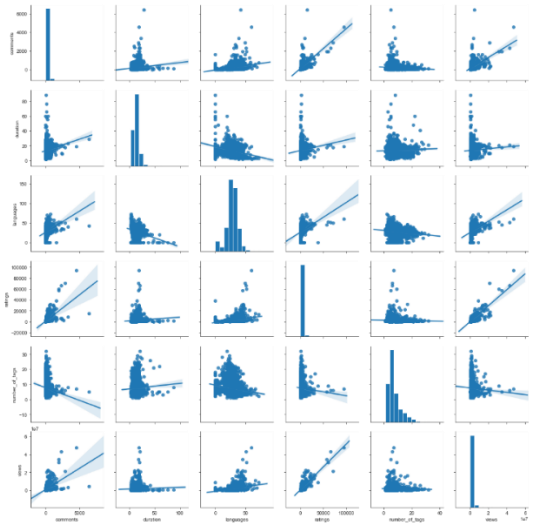

Figure 19: Regression Visualization - EDA selected features

4.4.3 Algorithms:

Logistic Regression:

According to Couronné, Logistic Regression is being considered as a standard approach for binary classification. [23]

Decision Tree Classifier:

The decision tree classifier is one of the most well- known machine learning techniques. A decision tree contains decision nodes and leaf nodes, each decision node corresponds to a single attribute with a number of branches as input data and each leaf node represents a class that is the result of decision for a case. [23]

Random Forest Classifier:

Random forest (RF) is an ensemble machine learning method based on the construction of multiple decision trees. In each decision tree, a data point falls into a particular leaf depending on its features and is assigned a prediction. The predictions of the data points are then averaged. RF has a built-in feature selection system and allows for joint features, making it not only an additive model but also a multiplicative one. [19]

Multinomial Naïve Bayes:

Naïve Bayes is a highly practical Bayesian learning method and is particularly suited to high dimensional tasks. It is often used as a baseline classifier and despite its simplicity often outperforms more sophisticated methods. [24] Multinomial Naïve Bayes also capture the information of the number of times a word occurs in a document. [21]

4.5 Model Selection:

4.5.1 TED main Classification model:

There are three algorithms being used for this Classification Model and the model evaluation metrics being used are confusion matrix, classification report, and accuracy score. Table 1, 2, 3 below are being used for each algorithm.

Table 1: Logistic Regression - 77.6%

Classification Report	*Precision*	*Recall*	*F1-Score*	*Support*
False	0.80	0.80	0.80	279
True	0.75	0.75	0.75	231
Average/Total	0.78	0.78	0.78	510

Confusion Matrix	*Predicted*	
Actual	1	0
1	222	57
0	57	174

Table 2: Decision Tree Classifier – 72.5%

Classification Report	*Precision*	*Recall*	*F1-Score*	*Support*
False	0.77	0.72	0.74	279
True	0.71	0.74	0.71	231
Average/Total	0.73	0.73	0.73	510

Confusion Matrix	*Predicted*	
Actual	1	0
1	200	79
0	61	170

Table 3: Random Forest Classifier – 79.2%

Classification Report	*Precision*	*Recall*	*F1-Score*	*Support*
False	0.81	0.81	0.81	279
True	0.77	0.77	0.77	231
Average/Total	0.79	0.79	0.79	510

Confusion Matrix	*Predicted*	
Actual	1	0
1	227	52
0	54	177

Based on the model evaluation metrics above, Random Forest and Logistic Regression have better performance than the Decision Tree Classifier. In order to see if the prediction accuracy can be improved, Recursive Feature Elimination (REF) is also applied to the Logistic Regression model, but the performance is lower than original model, from 77.6% to 76.8%. For the Random Forest Classifier, I also tried to improve the prediction accuracy by applying feature scaling to standardize the range of independent features. Consequently, the accuracy did improve from 79.2% to 80%. Therefore, the Classification Model being selected for TED main dataset is Random Forest Classifier, which could also identify what are the important features with feature importance metrics.

4.5.2 Transcript Classification model:

There are two algorithms being used for this Classification model and the model evaluation metrics includes accuracy score, classification report, confusion matrix, and visualization of AUC curve in Figure 20. In order to compare two models, I applied Cross-Validation technique with K-Fold Cross Validation (k=10) [8, 19], and the mean accuracy of Multinomial Naïve Bayes (65%) is still greater than Logistic Regression (60%), therefore, the selected model for transcript dataset shall be Multinomial Naïve Bayes.

Table 4: Logistic Regression - 60.5%

Classification Report	*Precision*	*Recall*	*F1-Score*	*Support*
False	0.64	0.56	0.60	259
True	0.58	0.65	0.61	235
Average/Total	0.61	0.61	0.60	494

Confusion Matrix	*Predicted*	
Actual	1	0
1	146	113
0	82	153

Table 5: Multinomial Naïve Bayes – 67.8%

Classification Report	*Precision*	*Recall*	*F1-Score*	*Support*
False	0.71	0.66	0.68	259
True	0.65	0.70	0.67	235
Average/Total	0.68	0.68	0.68	494

Confusion Matrix	*Predicted*	
Actual	1	0
1	170	89
0	70	165

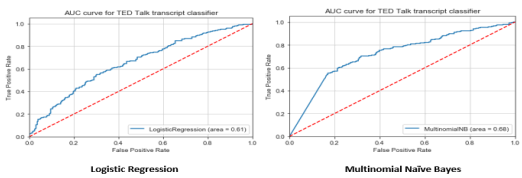

Figure 20: AUC Curve for Classifiers

5. CONCLUSION & LSSONS LEARNED

5.1 Conclusion:

- ***What makes a TED Talk popular?***

Figure 21 below is a bar plot based on the feature importance from Random Forest Classifier. I also used the RFE to select the top 5 features, which is same as Random Forest Classifier, *comments*, *languages*, *ratings*, *published weekday*, and *number of tags.*

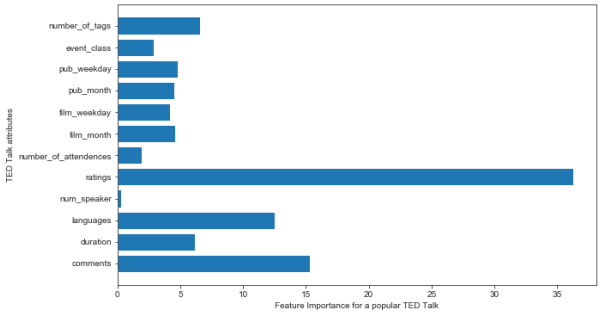

Figure 21: Feature Importance of Random Forest Classifier

- ***What are the most popular topics in TED Talks?***

Since the tags column represent the topic based on its sample data, there is a Word Cloud generated based on this column, which represents all the popular words that appear more frequently, which is Figure 22. In addition, Figure 23 below is a histogram generated after formatting tags to a list that counts the occurrences of each topic.

Figure 22: Most Popular Topics via Word Cloud

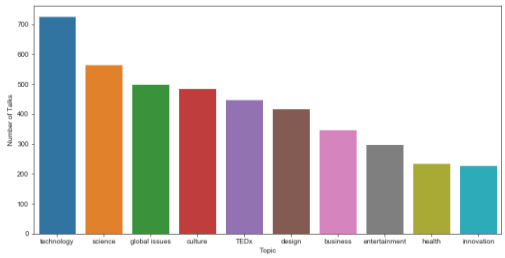

Figure 23: TOP 10 popular topics

- ***Could classification model being used to predict the popularity of a TED Talk based on its transcript?***

Based on the Model Selection section, the Multinomial Naïve Bayes Classifier has a 68% prediction accuracy for predicting the popularity of TED Talk transcript. So, the answer for this question would be YES.

5.2 Lessons Learned:

While building the Classification Model for TED main dataset, I tried to use Recursive Feature Elimination (REF) to improve the prediction accuracy score, but the prediction became lower. The lessons learned from here is that applying feature selection algorithm after manually selection may reduce the prediction accuracy instead of improving.

6. DEPLOY

GitHub Repository that contains following items can be found at

https://github.com/Jacob13209/CS522_Predict_TED_Talk_Popularity

- Dataset from Kaggle
- Jupyter Notebook
- Project Paper (Word and PDF)
- Presentation Slides (PPT and PDF)
- Poster (Power Point and PDF)

Project Introduction Video:

https://www.youtube.com/watch?v=SLNXuF-Izgo

7. FUTURE WORK

This project has met its original objectives that it could identify those key features that contribute to the popularity of a TED talk and it could also predict the popularity based on transcript, but the features being used to train the classification model is limited and the prediction accuracy is lower than expected.

Future work will be carried out to understand why the prediction accuracy is lower and do further pre-tuning and post-tuning to enhance the classification model. Further, understanding the reality of all features and take full advantage of all available features are recommended. While training the classification model, such as title, description, and tags are not being used for this project, but those features are actually important contributors for TED Talk popularity. Consequently, future work would also be carried out to apply Topic Modeling algorithms to classify the topic class to a categorical value, the proposed algorithm is *Latent Dirichlet allocation* (LDA) [17].

8. ACKNOWLEDGEMENTS

This project is a course project for Data Mining Class at Hood College during the Fall Semester of 2018. Firstly, I would like to thank Dr. Liu, Xinlian for giving us the fantastic opportunity choosing project topic by ourselves and the great guidance given through the semester. In addition, I would like to give thanks to classmates and friends for providing suggestions on this project. Finally, I would like to express my deepest thanks and sincere appreciation to my family, girlfriend, and colleagues for their love, understanding, and support.

9. REFERENCES

[1] Wikipedia. 2018. Wikipedia: TED (conference). Retrieved from https://www.wikipedia.org.

[2] TED. 2018. TED Talks. Retrieved from https://www.ted.com/talks.

[3] Charlie Rose. 2015. CBS: TED Talks. (April 2015). Retrieved October 20, 2018 from

https://www.cbsnews.com/news/ted-talks-60-minutes-charlie-rose/

[4] Ralph Jacobson. 2013. IBM. (April 2013) Retrieved October 20, 2018 from https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/

[5] Vasant Dhar. 2013. Data science and prediction. Commun. ACM 56, 12 (December 2013), 64-73. DOI: https://doi.org/10.1145/2500499

[6] Richard Socher, Yoshua Bengio, and Christopher D. Manning. 2012. Deep learning for NLP (without magic). In Tutorial Abstracts of ACL 2012 (ACL '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 5-5.

[7] Rounak Banik. 2017. TED Talk Dataset. (2017) Retrieved October 20, 2018 from https://www.kaggle.com/rounakbanik/ted-talks/home

[8] Usman Malik. 2018. Stack Abuse: Cross Validation and Grid Search for Model Selection in Python. (2018) Retrieved December 15, 2018 from https://stackabuse.com/cross-validation-and-grid-search-for-model-selection-in-python

[9] Sarah Hong, Yuyi Liu, and Zhao Xiao. 2016. Visual Analysis of TED Talk Topic Trends. In Proceedings of the 9th International Symposium on Visual Information Communication and Interaction (VINCI '16). ACM, New York, NY, USA, 150-151. DOI: https://doi.org/10.1145/2968220.2972225

[10] J. Oh, I. Lee, Y. Seonwoo, S. Sung, I. Kwon and J. Lee, "TED Talk Recommender Using Speech Transcripts," 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, 2018, pp. 598- 600.

doi:10.1109/ASONAM.2018.8508644

[11] Wikipedia. 2018. Wikipedia: Exploratory data analysis. Retrieved December 15 from https://www.wikipedia.org.

[12] Komorowski, Matthieu & Marshall, Dominic & Salciccioli, Justin & Crutain, Yves. 2016. Exploratory Data Analysis.

[13] Anaconda. 2018. Anaconda: missingno. Retrieved December 15 from

https://anaconda.org/conda-forge/missingno

[14] Adi Bronshtein. 2018. Towards Data Science: Train/Test Spilt and Cross Validation in Python. (2018) Retrieved December 15 from

https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

[15] J. R. Zhang, J. Sherwin, J. Dmochowski, P. Sajda, and J. R. Kender, “Correlating speaker gestures in political debates with au- dience engagement measured via eeg,” in Proc. 22nd ACM Int’l Conf. on Multimedia. 2654909: ACM, pp. 387–396.

[16] M. Karg, A. A. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic, “Body movements for affective expression: A survey of automatic recognition and generation,” Affective Computing, IEEE Trans., vol. 4, no. 4, pp. 341–359, 2013.

[17] Machine Learning Plus. 2018. Topic Modeling with Gensim (Python). Retrieved December 16, 2018 from

https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

[18] Cullen, A., & Harte, N. (2017). Thin slicing to predict viewer impressions of TED Talks. In International Conference on Auditory-Visual Speech Processing. pp. 58–63.

[19] Decruyenaere, A., Decruyenaere, P., Peeters, P., Vermassen, F., Dhaene, T., & Couckuyt, I.2015. Prediction of delayed graft function after kidney transplantation: comparison between logistic regression and machine learning methods. BMC medical informatics and decision making, 15, 83. doi:10.1186/s12911-015-0206-y

[20] PAK, M , GUNAL, S . 2017. THE IMPACT OF TEXT REPRESENTATION AND PREPROCESSING ON AUTHOR IDENTIFICATION. Anadolu Üniversitesi Bilim Ve Teknoloji Dergisi A - Uygulamalı Bilimler ve Mühendislik, 18 (1), 218-224. DOI: 10.18038/aubtda.270276

[21] L. Jiang, Z. Cai & D. Wang. 2010. Improving Naive Bayes for Classification, International Journal of Computers and Applications, 32:3, 328-332, DOI: 10.2316/Journal.202.2010.3.202-2747

[22] Gary Stein, Bing Chen, Annie S. Wu, and Kien A. Hua. 2005. Decision tree classifier for network intrusion detection with GA-based feature selection. In Proceedings of the 43rd annual Southeast regional conference - Volume 2 (ACM-SE 43), Vol. 2. ACM, New York, NY, USA, 136-141. DOI=http://dx.doi.org/10.1145/1167253.1167288

[23] Couronné, Raphael & Probst, Philipp & Boulesteix, Anne-Laure. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. BMC Bioinformatics. 19. 10.1186/s12859-018-2264-5.

[24] Jaideep Vaidya, Basit Shafiq, Anirban Basu, and Yuan Hong. 2013. Differentially Private Naive Bayes Classification. In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01 (WI-IAT '13), Vol. 1. IEEE Computer Society, Washington, DC, USA, 571-576. DOI=http://dx.doi.org/10.1109/WI-IAT.2013.80