Flipkart Grid 3.0

Problem Statement

Trust and Affluence signal extraction from social media data.

Team Name : GitGud

Institute Name : Kalinga Institute of Industrial Technology, Bhubaneswar

Team Member Details

Team Name	GitGud		
Institute Name	Kalinga Institute of Industrial Technology, Bhubaneswar		
Team Members	1	2	3
Name	Debangshu Bhattacharjee	Aditya Kumar Dalmia	Debmalya Chatterjee
Batch	2023	2023	2023

Table of Contents

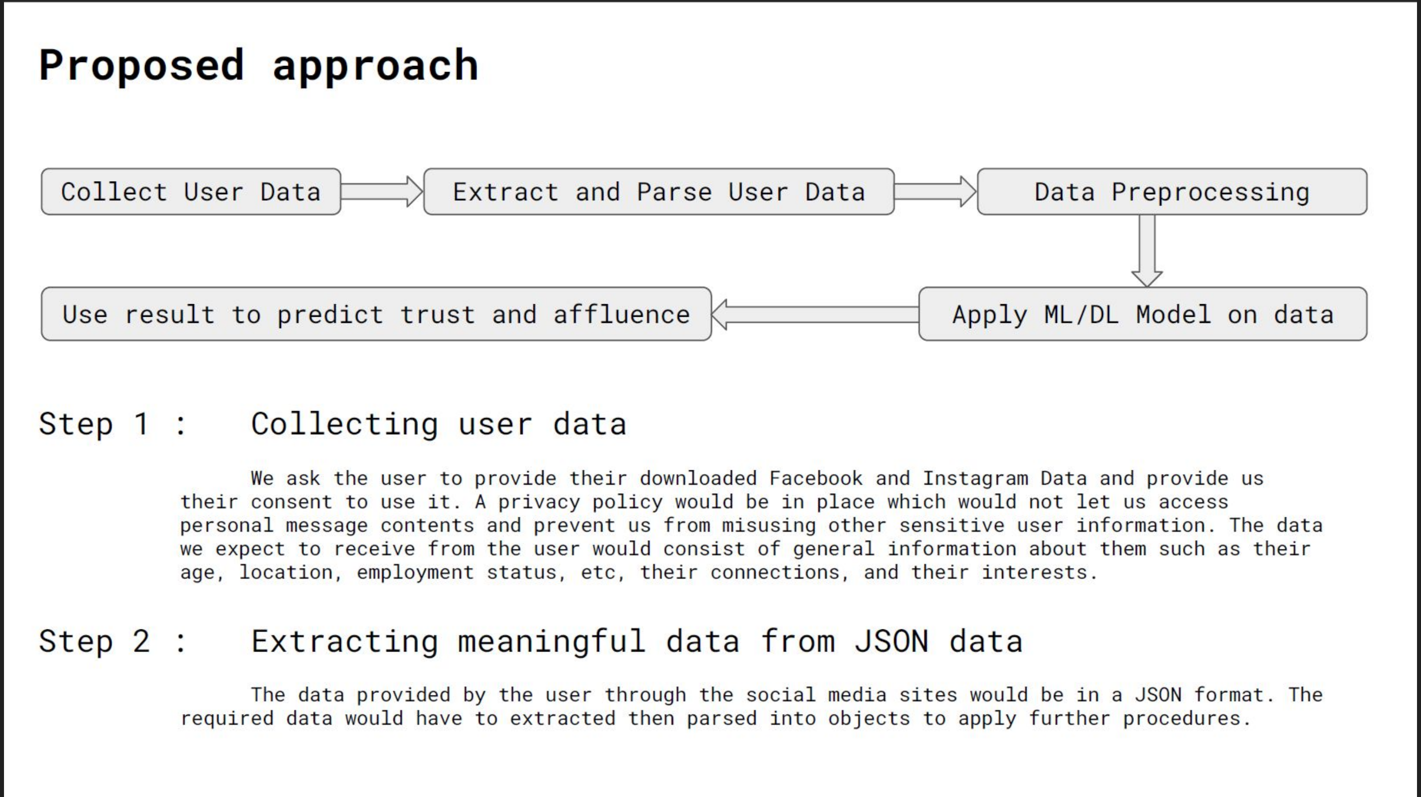
1. Our Approach till Now
2. General Overview
3. Overcoming Limitations
4. Our Approach
5. Code Execution
6. Limitations
7. Improvements
8. Future Scope
9. Conclusion

Table of Contents

1. Our Approach till Now
2. General Overview
3. Overcoming Limitations
4. Our Approach
5. Code Execution
6. Limitations
7. Improvements
8. Future Scope
9. Conclusion

Our Approach till now

Our Approach till now



Proposed approach

Collect User Data → Extract and Parse User Data → Data Preprocessing → Apply ML/DL Model on data → Use result to predict trust and affluence

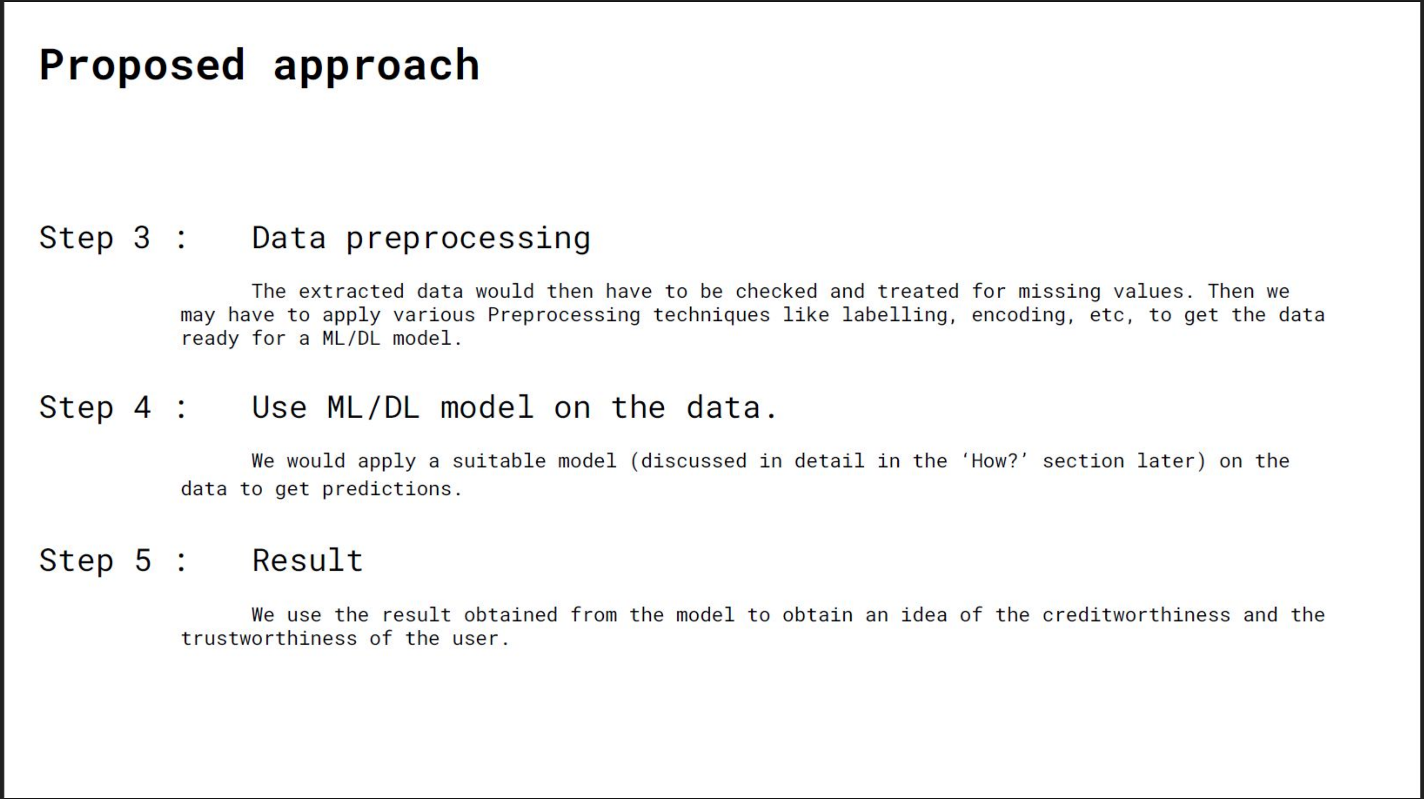
Step 1 : Collecting user data

We ask the user to provide their downloaded Facebook and Instagram Data and provide us their consent to use it. A privacy policy would be in place which would not let us access personal message contents and prevent us from misusing other sensitive user information. The data we expect to receive from the user would consist of general information about them such as their age, location, employment status, etc, their connections, and their interests.

Step 2 : Extracting meaningful data from JSON data

The data provided by the user through the social media sites would be in a JSON format. The required data would have to extracted then parsed into objects to apply further procedures.

Our Approach till now



Proposed approach

Step 3 : Data preprocessing

The extracted data would then have to be checked and treated for missing values. Then we may have to apply various Preprocessing techniques like labelling, encoding, etc, to get the data ready for a ML/DL model.

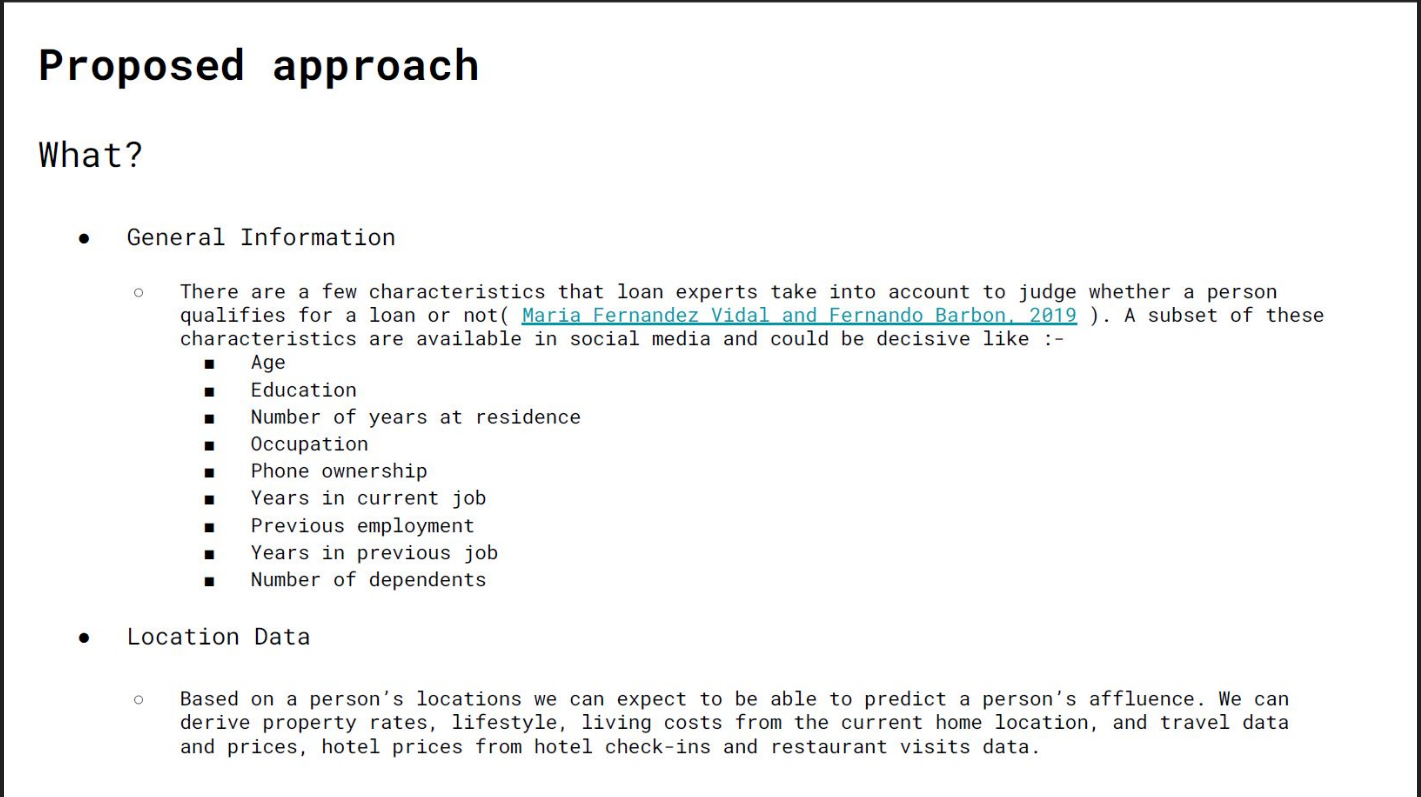
Step 4 : Use ML/DL model on the data.

We would apply a suitable model (discussed in detail in the ‘How?’ section later) on the data to get predictions.

Step 5 : Result

We use the result obtained from the model to obtain an idea of the creditworthiness and the trustworthiness of the user.

Our Approach till now

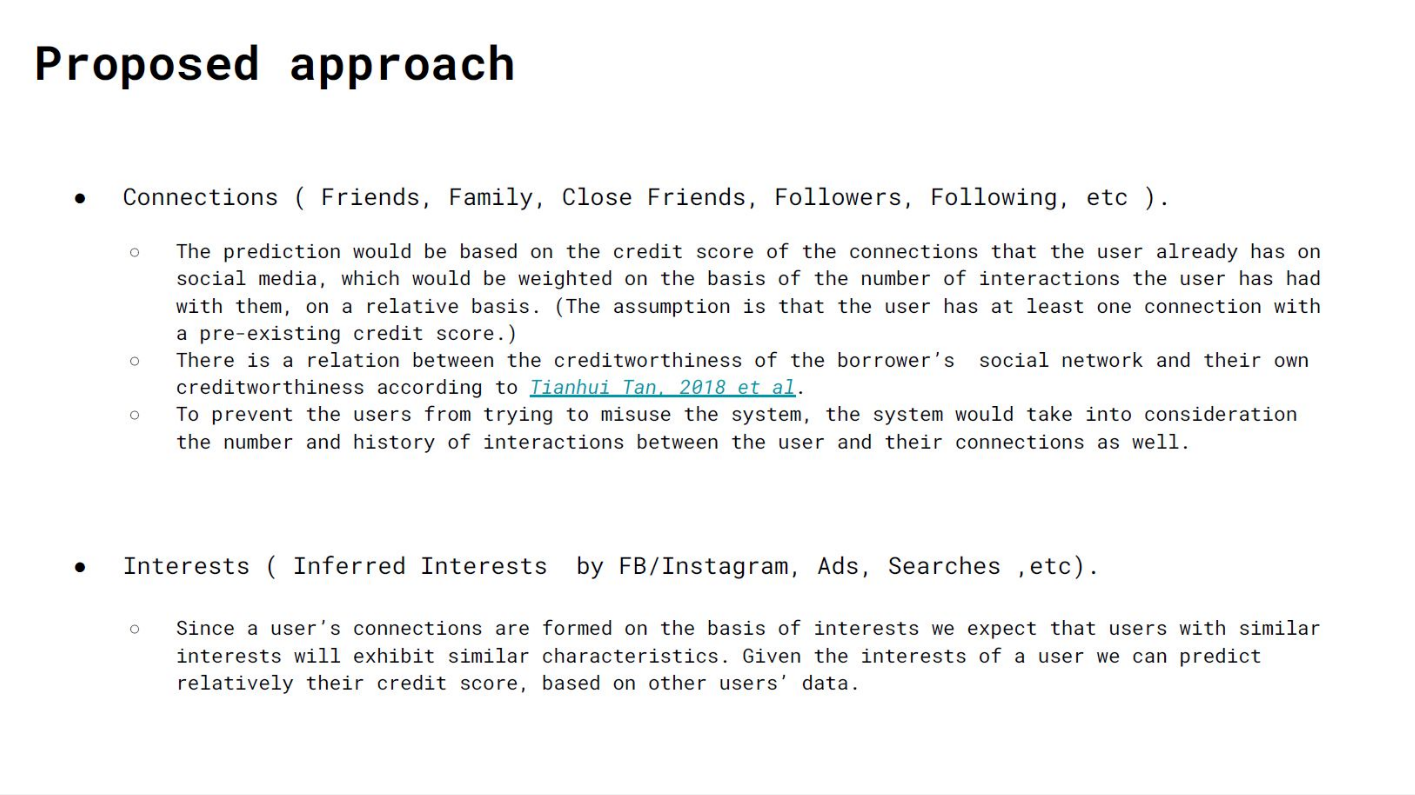


Proposed approach

What?

- General Information
  - There are a few characteristics that loan experts take into account to judge whether a person qualifies for a loan or not(Maria Fernandez Vidal and Fernando Barbon, 2019). A subset of these characteristics are available in social media and could be decisive like :-
    - Age
    - Education
    - Number of years at residence
    - Occupation
    - Phone ownership
    - Years in current job
    - Previous employment
    - Years in previous job
    - Number of dependents
- Location Data
  - Based on a person's locations we can expect to be able to predict a person's affluence. We can derive property rates, lifestyle, living costs from the current home location, and travel data and prices, hotel prices from hotel check-ins and restaurant visits data.

Our Approach till now

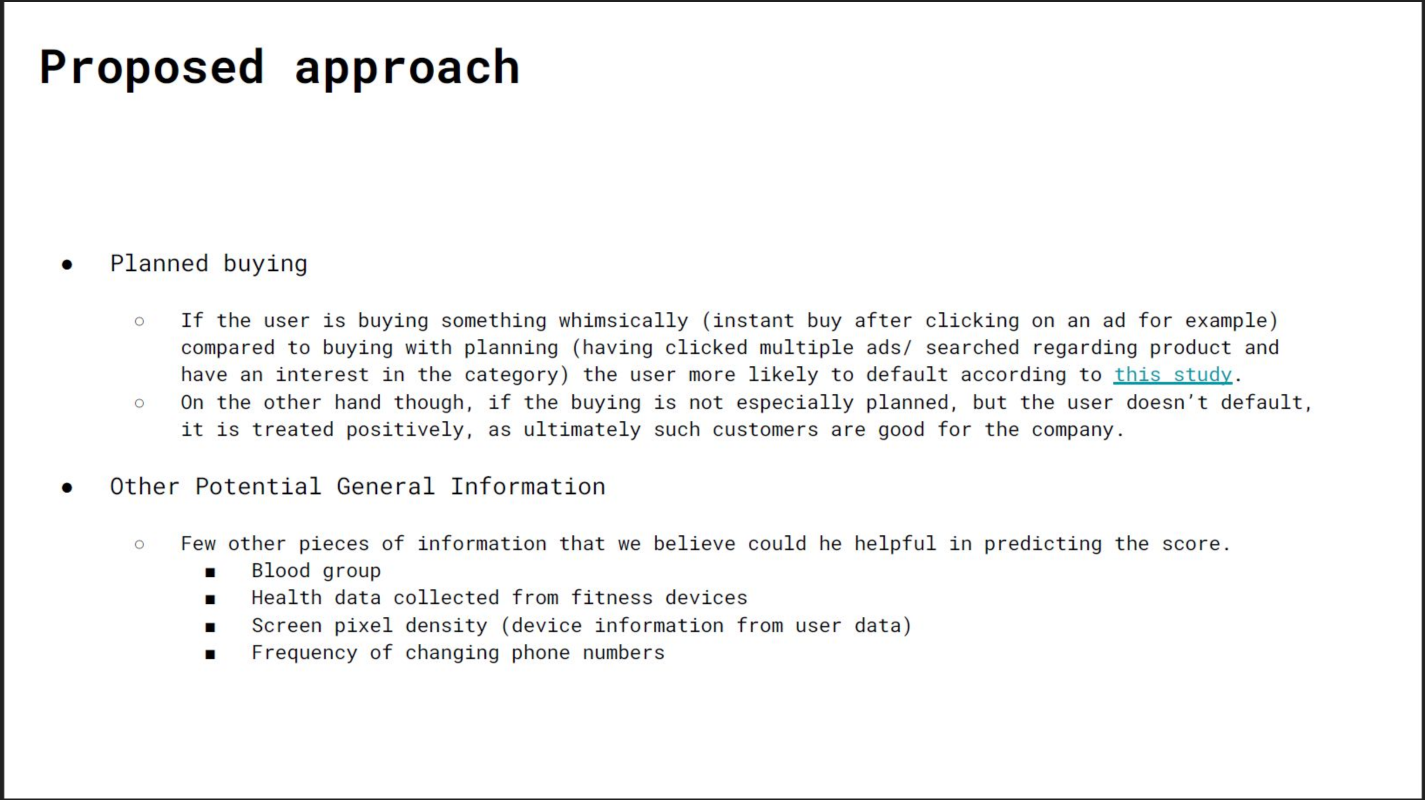


Proposed approach

- Connections (Friends, Family, Close Friends, Followers, Following, etc).
  - The prediction would be based on the credit score of the connections that the user already has on social media, which would be weighted on the basis of the number of interactions the user has had with them, on a relative basis. (The assumption is that the user has at least one connection with a pre-existing credit score.)
  - There is a relation between the creditworthiness of the borrower's social network and their own creditworthiness according to *Tianhui Tan, 2018 et al*.
  - To prevent the users from trying to misuse the system, the system would take into consideration the number and history of interactions between the user and their connections as well.

- Interests (Inferred Interests by FB/Instagram, Ads, Searches ,etc).
  - Since a user's connections are formed on the basis of interests we expect that users with similar interests will exhibit similar characteristics. Given the interests of a user we can predict relatively their credit score, based on other users' data.

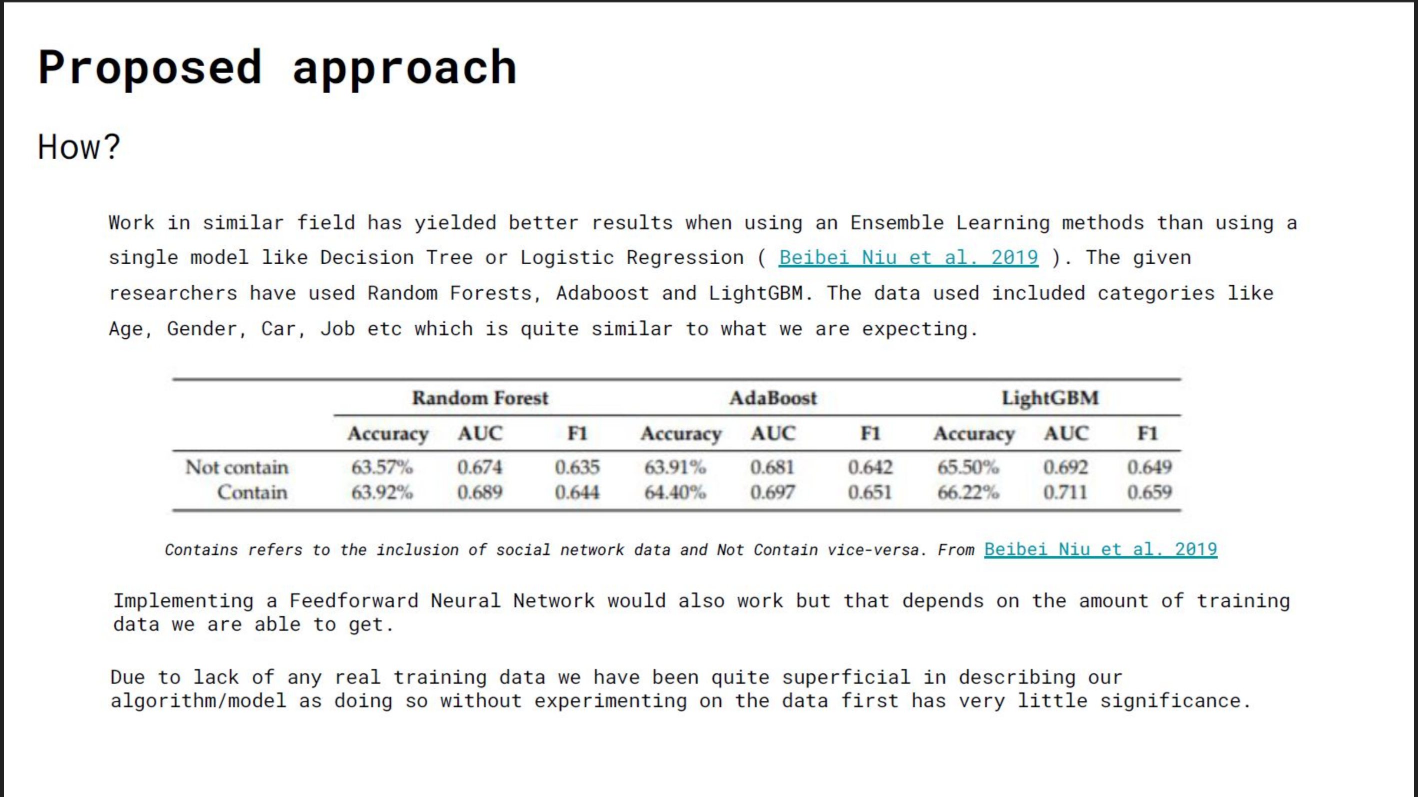
Our Approach till now



Proposed approach

- Planned buying
    - ○ If the user is buying something whimsically (instant buy after clicking on an ad for example) compared to buying with planning (having clicked multiple ads/ searched regarding product and have an interest in the category) the user more likely to default according to this study.
    - ○ On the other hand though, if the buying is not especially planned, but the user doesn't default, it is treated positively, as ultimately such customers are good for the company.
- Other Potential General Information
    - ○ Few other pieces of information that we believe could he helpful in predicting the score.
        - ■ Blood group
        - ■ Health data collected from fitness devices
        - ■ Screen pixel density (device information from user data)
        - ■ Frequency of changing phone numbers

Our Approach till now



Proposed approach

How?

Work in similar field has yielded better results when using an Ensemble Learning methods than using a single model like Decision Tree or Logistic Regression (Beibei Niu et al. 2019). The given researchers have used Random Forests, Adaboost and LightGBM. The data used included categories like Age, Gender, Car, Job etc which is quite similar to what we are expecting.

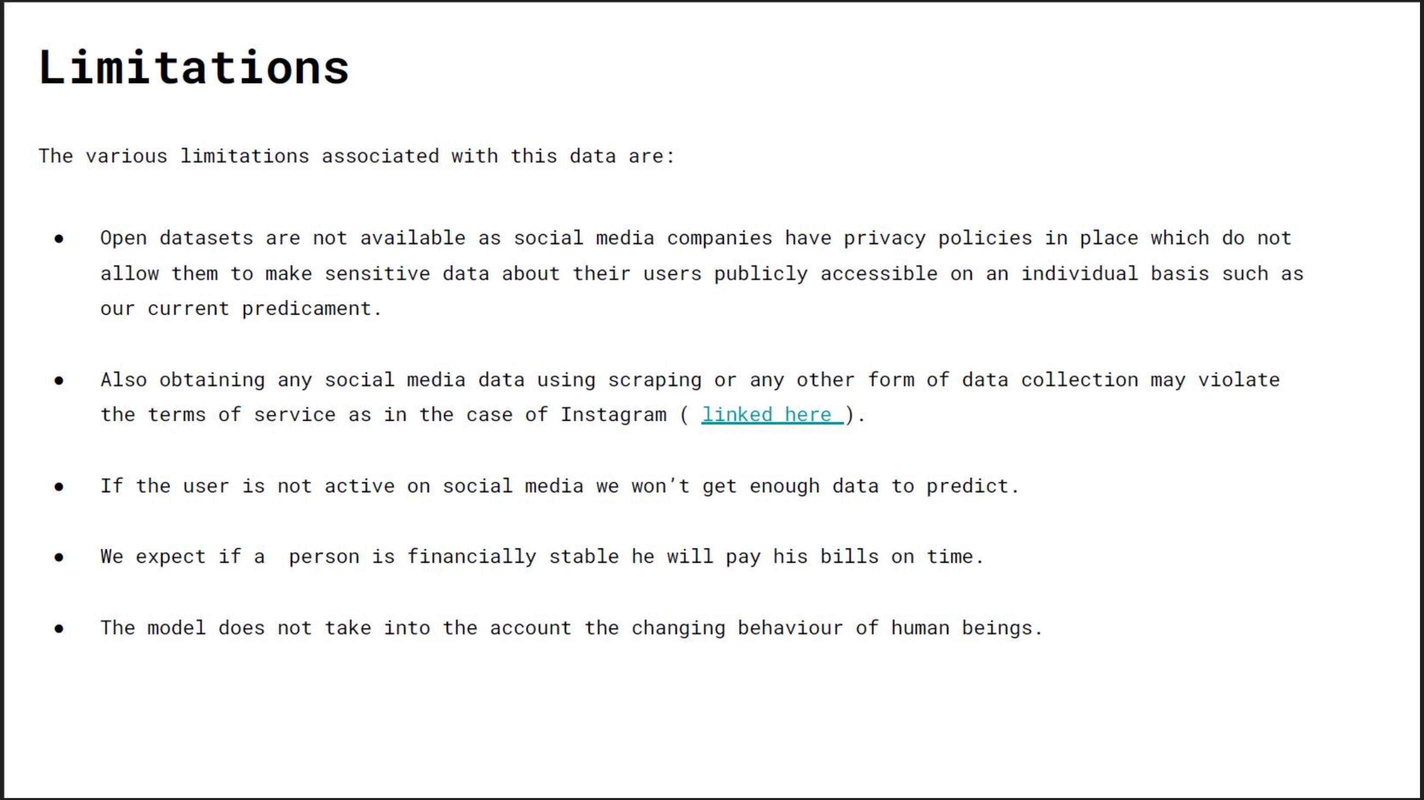
	Random Forest			AdaBoost			LightGBM		
	Accuracy	AUC	F1	Accuracy	AUC	F1	Accuracy	AUC	F1
Not contain	63.57%	0.674	0.635	63.91%	0.681	0.642	65.50%	0.692	0.649
Contain	63.92%	0.689	0.644	64.40%	0.697	0.651	66.22%	0.711	0.659

Contains refers to the inclusion of social network data and Not Contain vice-versa. From Beibei Niu et al. 2019

Implementing a Feedforward Neural Network would also work but that depends on the amount of training data we are able to get.

Due to lack of any real training data we have been quite superficial in describing our algorithm/model as doing so without experimenting on the data first has very little significance.

Our Approach till now



Limitations

The various limitations associated with this data are:

- Open datasets are not available as social media companies have privacy policies in place which do not allow them to make sensitive data about their users publicly accessible on an individual basis such as our current predicament.
- Also obtaining any social media data using scraping or any other form of data collection may violate the terms of service as in the case of Instagram (linked here).
- If the user is not active on social media we won't get enough data to predict.
- We expect if a person is financially stable he will pay his bills on time.
- The model does not take into the account the changing behaviour of human beings.

Table of Contents

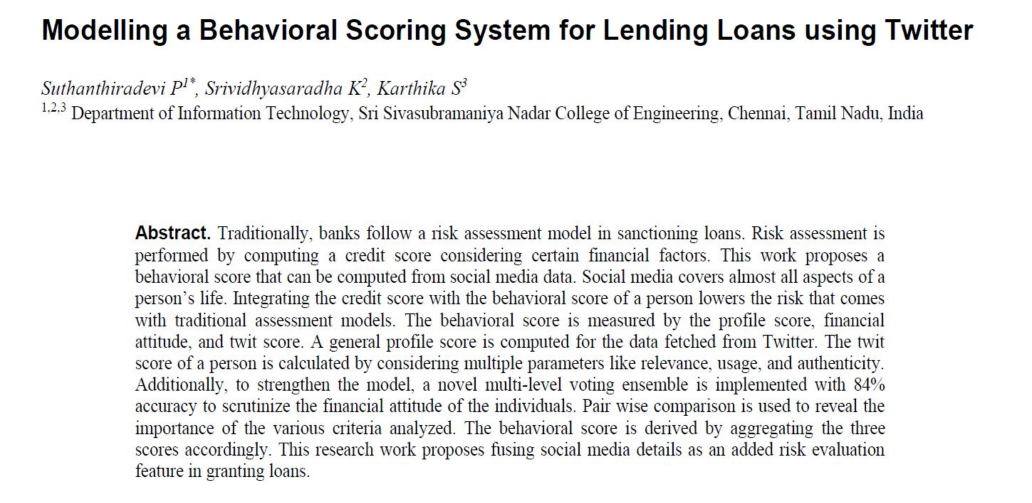
1. Our Approach till Now
2. General Overview
3. Overcoming Limitations
4. Our Approach
5. Code Execution
6. Limitations
7. Improvements
8. Future Scope
9. Conclusion

General Overview

General Overview

Due to lack of enough data to verify any of the methods that we devised, we have taken inspiration and have based most of our approach on this paper published in the ITM Web of Conferences, pertaining to a similar topic, because it had precedent. The paper proposes a method to calculate a social media score (referred to as *Behavioral Score*) derived from a user's Twitter profile and Twitter timeline, which they later propose combining with their traditional credit scores.

Taking inspiration from the paper, we have modelled our approach around finding a behavioral score for a person which acts as a measure of their trust and affluence, helping us assess their credit worthiness.



Modelling a Behavioral Scoring System for Lending Loans using Twitter

Suthanthiradevi P[1], Srividhyasaradha K[2], Karthika S[3]*
[1,2,3] Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India

Abstract. Traditionally, banks follow a risk assessment model in sanctioning loans. Risk assessment is performed by computing a credit score considering certain financial factors. This work proposes a behavioral score that can be computed from social media data. Social media covers almost all aspects of a person's life. Integrating the credit score with the behavioral score of a person lowers the risk that comes with traditional assessment models. The behavioral score is measured by the profile score, financial attitude, and twit score. A general profile score is computed for the data fetched from Twitter. The twit score of a person is calculated by considering multiple parameters like relevance, usage, and authenticity. Additionally, to strengthen the model, a novel multi-level voting ensemble is implemented with 84% accuracy to scrutinize the financial attitude of the individuals. Pair wise comparison is used to reveal the importance of the various criteria analyzed. The behavioral score is derived by aggregating the three scores accordingly. This research work proposes fusing social media details as an added risk evaluation feature in granting loans.

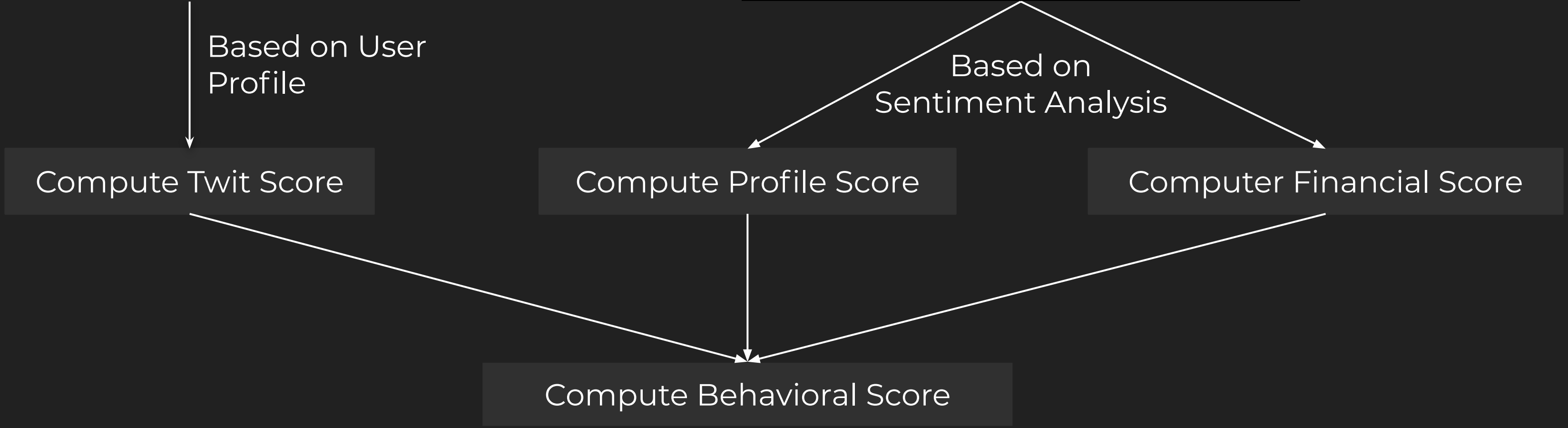
General Overview

We have attempted to derive - *Twit Score, Profile Score and Financial Score* for a user, which were aggregated to get our final measure (details on the scores and the approach are in later slides.)

Elon Musk @elonmusk · Jun 6

Don’t kill what you hate,
Save what you love

43.2K 70.2K 473.5K



Based on User Profile

Based on Sentiment Analysis

Compute Twit Score

Compute Profile Score

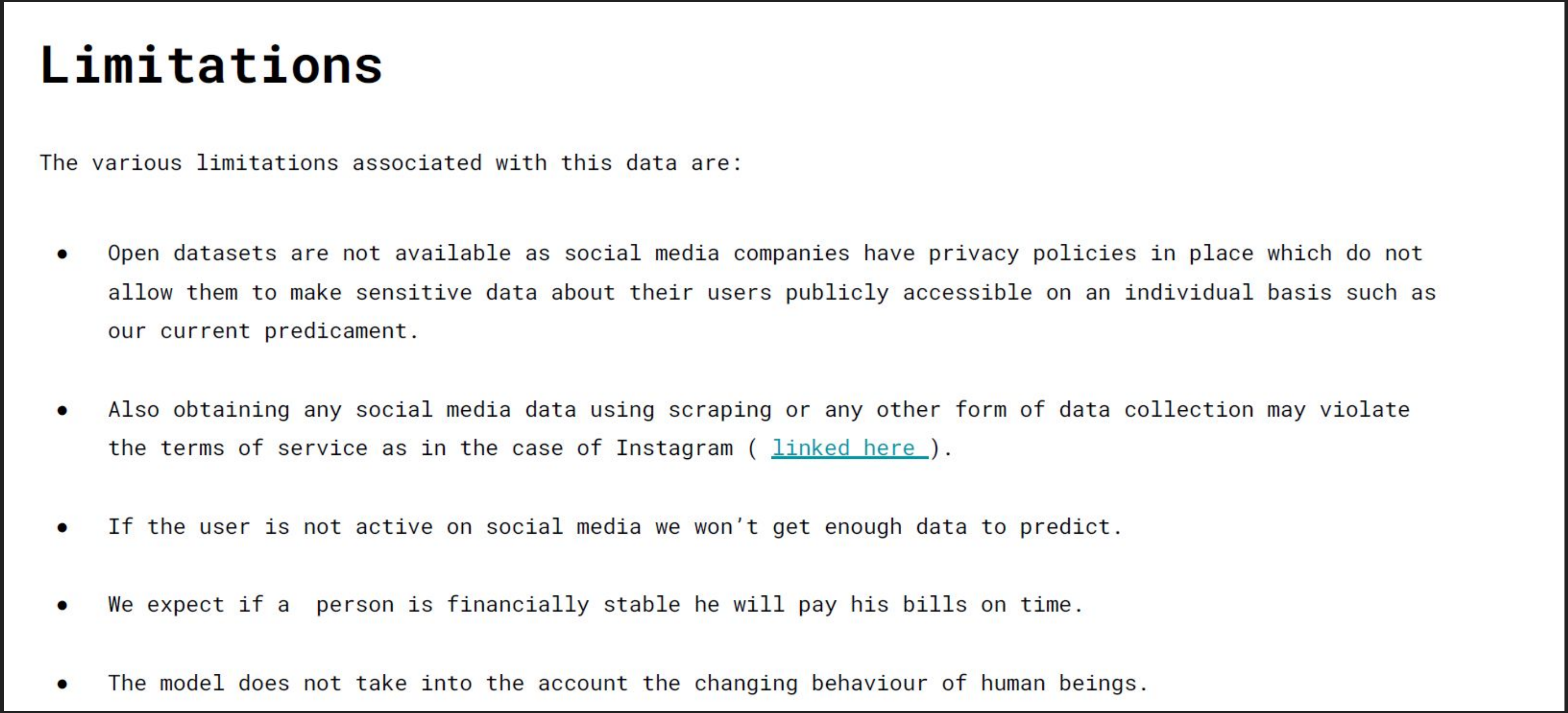
Computer Financial Score

Compute Behavioral Score

Table of Contents

1. Our Approach till Now
2. General Overview
3. Overcoming Limitations
4. Our Approach
5. Code Execution
6. Limitations
7. Improvements
8. Future Scope
9. Conclusion

Overcoming Limitations

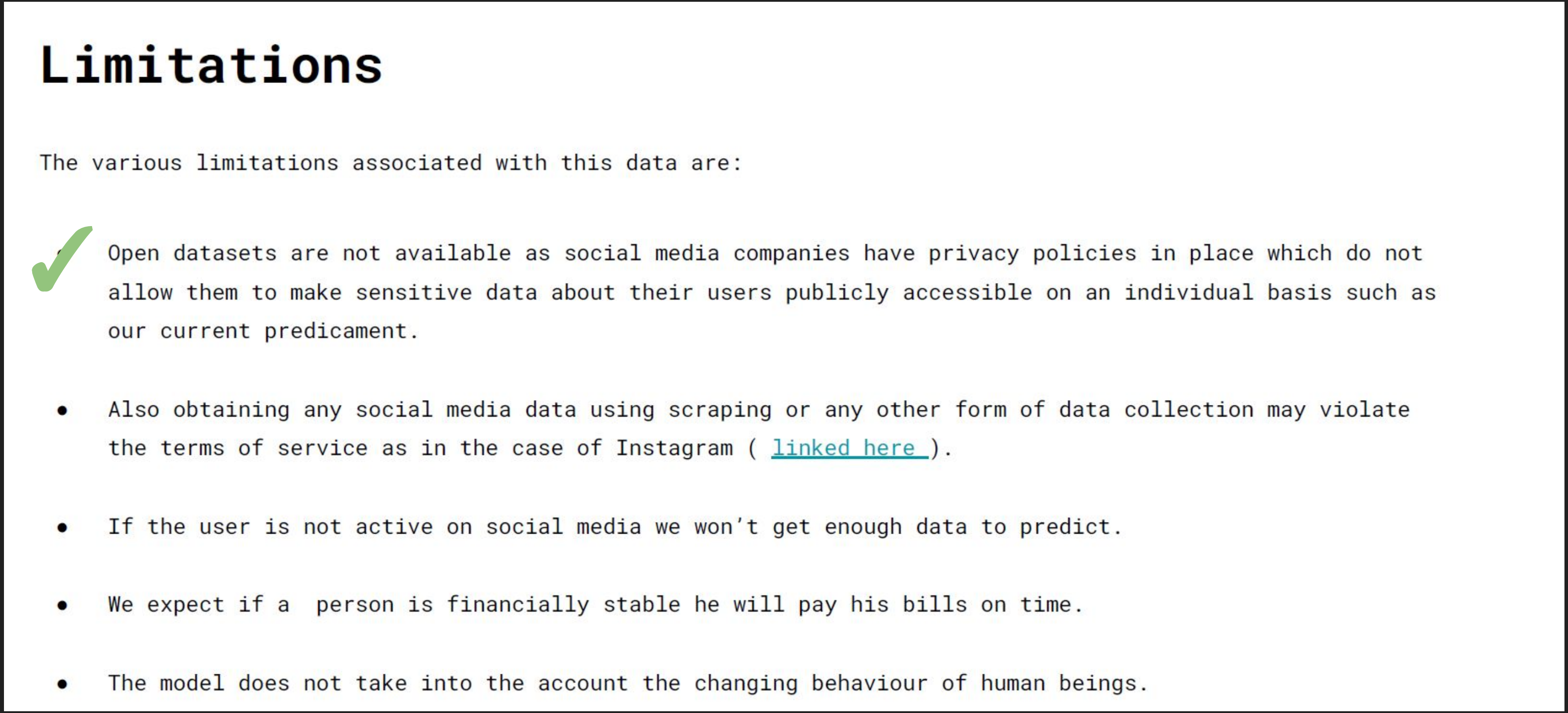


Limitations

The various limitations associated with this data are:

- Open datasets are not available as social media companies have privacy policies in place which do not allow them to make sensitive data about their users publicly accessible on an individual basis such as our current predicament.
- Also obtaining any social media data using scraping or any other form of data collection may violate the terms of service as in the case of Instagram (linked here).
- If the user is not active on social media we won't get enough data to predict.
- We expect if a person is financially stable he will pay his bills on time.
- The model does not take into the account the changing behaviour of human beings.

Overcoming Limitations

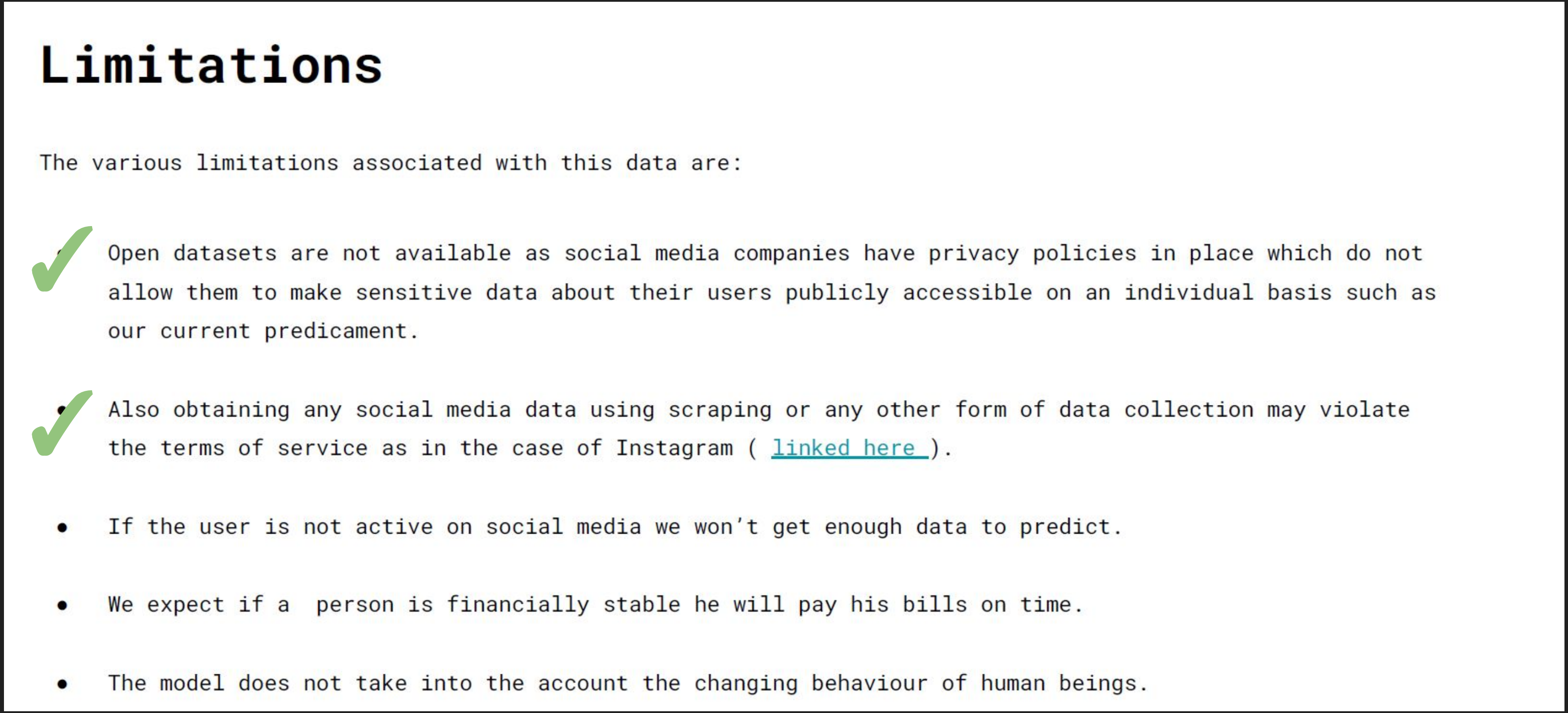


Limitations

The various limitations associated with this data are:

- ✓ Open datasets are not available as social media companies have privacy policies in place which do not allow them to make sensitive data about their users publicly accessible on an individual basis such as our current predicament.
- Also obtaining any social media data using scraping or any other form of data collection may violate the terms of service as in the case of Instagram (linked here).
- If the user is not active on social media we won't get enough data to predict.
- We expect if a person is financially stable he will pay his bills on time.
- The model does not take into the account the changing behaviour of human beings.

Overcoming Limitations

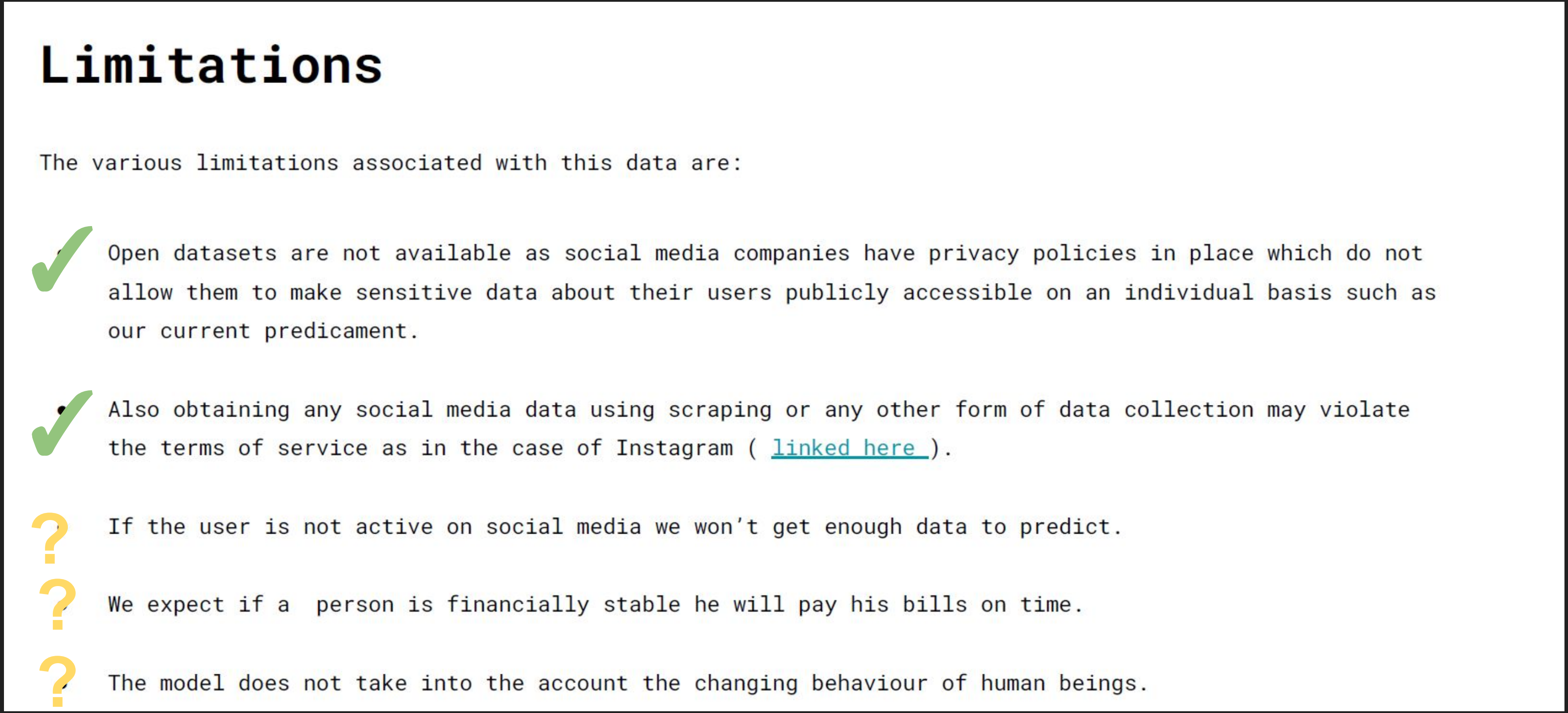


Limitations

The various limitations associated with this data are:

- ✓ Open datasets are not available as social media companies have privacy policies in place which do not allow them to make sensitive data about their users publicly accessible on an individual basis such as our current predicament.
- ✓ Also obtaining any social media data using scraping or any other form of data collection may violate the terms of service as in the case of Instagram (linked here).
- If the user is not active on social media we won't get enough data to predict.
- We expect if a person is financially stable he will pay his bills on time.
- The model does not take into the account the changing behaviour of human beings.

Overcoming Limitations



Limitations

The various limitations associated with this data are:

- ✓ Open datasets are not available as social media companies have privacy policies in place which do not allow them to make sensitive data about their users publicly accessible on an individual basis such as our current predicament.
- ✓ Also obtaining any social media data using scraping or any other form of data collection may violate the terms of service as in the case of Instagram (linked here).
- ? If the user is not active on social media we won’t get enough data to predict.
- ? We expect if a person is financially stable he will pay his bills on time.
- ? The model does not take into the account the changing behaviour of human beings.

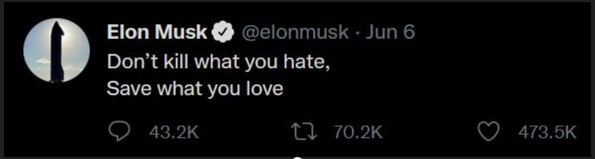
Table of Contents

1. Our Approach till Now
2. General Overview
3. Overcoming Limitations
4. Our Approach
5. Code Execution
6. Limitations
7. Improvements
8. Future Scope
9. Conclusion

Our Approach

For evaluating a user:

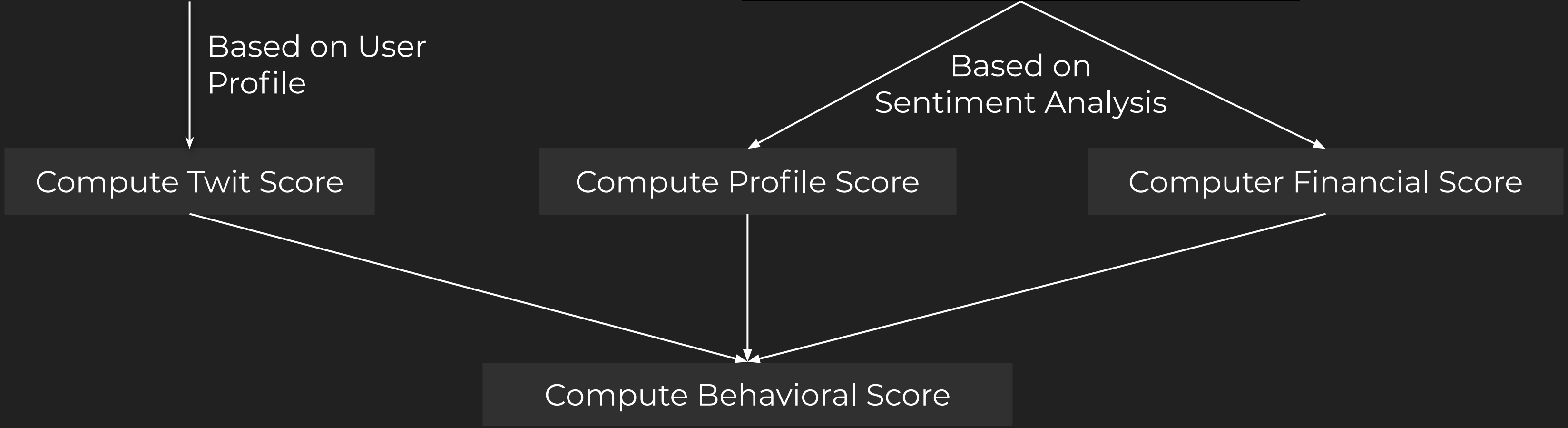
Behavioral Score = Twit Score + Profile Score + Financial Score



Elon Musk @elonmusk · Jun 6

Don’t kill what you hate,
Save what you love

43.2K 70.2K 473.5K



Based on User Profile

Based on Sentiment Analysis

Compute Twit Score

Compute Profile Score

Computer Financial Score

Compute Behavioral Score

Our Approach

The user data is collected using Tweepy, a Python library for accessing the Twitter API.

We collected data for 34 users of varying backgrounds. We extracted the features from the data, applied feature engineering on them, and scaled all the numerical values.

Using these features we then calculate the scores.

Using the previously trained model and this normalization factors we can predict the scores for a new user / users.

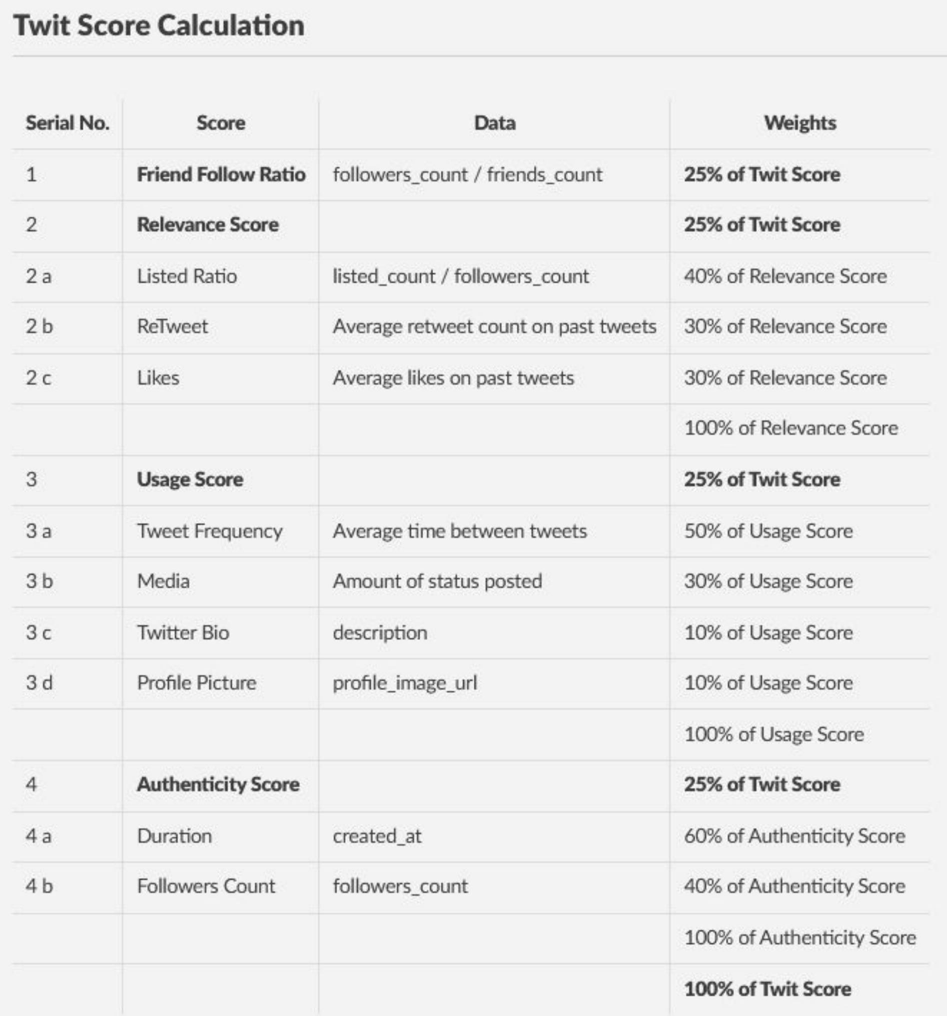
Our Approach

Scoring Mechanism

Scores	Definition
Twit Score	The user profile of the person forms the basis of computing the twit score. It is an attempt to rate the quality of Twitter user by various metrics available through the API. A Twitter user with a relatively low twit score is more likely to be a sign of a spam account or a less safe user.
Profile Score	The 200 most recent tweets of the user form the basis of this score. They are pre-processed by tokenization, lemmatization, stop-word removal, etc. Subsequently, sentiment analysis is performed and the percentage of positive tweets is chosen to be the profile score.
Financial Score	The financial tweets of the user form the basis of this score. The financial tweets are identified by checking against a corpus of such terms. To classify the financial tweets as positive or negative, sentiment analysis is performed on them. The financial score is the average of the percentage of financial tweets and the percentage of positive ones among them.

Our Approach

Weightages of different factors taken into consideration for the calculation of Twit Score.



Twit Score Calculation

Serial No.	Score	Data	Weights
1	**Friend Follow Ratio**	followers_count / friends_count	**25% of Twit Score**
2	**Relevance Score**		**25% of Twit Score**
2 a	Listed Ratio	listed_count / followers_count	40% of Relevance Score
2 b	ReTweet	Average retweet count on past tweets	30% of Relevance Score
2 c	Likes	Average likes on past tweets	30% of Relevance Score
			100% of Relevance Score
3	**Usage Score**		**25% of Twit Score**
3 a	Tweet Frequency	Average time between tweets	50% of Usage Score
3 b	Media	Amount of status posted	30% of Usage Score
3 c	Twitter Bio	description	10% of Usage Score
3 d	Profile Picture	profile_image_url	10% of Usage Score
			100% of Usage Score
4	**Authenticity Score**		**25% of Twit Score**
4 a	Duration	created_at	60% of Authenticity Score
4 b	Followers Count	followers_count	40% of Authenticity Score
			100% of Authenticity Score
			100% of Twit Score

Our Approach

For performing Sentiment Analysis we trained a custom model using the Sentiment140 dataset consisting of labeled tweets.

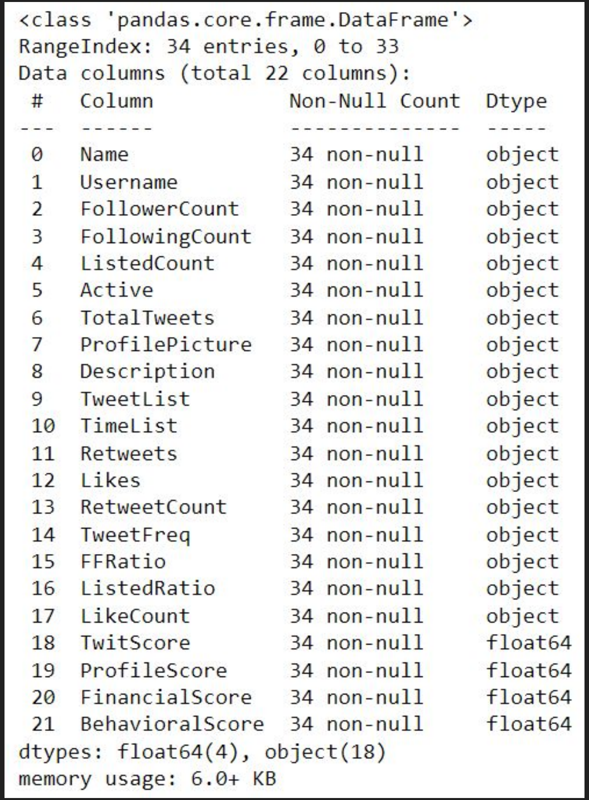
We tested Logistic Regression, Random Forests, Gradient Boosting, Adaptive Boosting, and Voting Classifier with all the same models as well, by feeding Count Vectorized Data followed by TFIDF Vectorized data.

We concluded that Logistic Regression with Count Vectorized data would be the best choice. The code for running the different models we tried, along with their results as confusion matrices is provided in the submitted code (Trials/SentimentAnalysisTrials.ipynb).

For finding whether a tweet is a financial tweet or not, we compiled a list of financial terms (source: Investopedia, etc.) and checked the tweets for common and advanced words related to finance.

Our Approach

The final dataset looks something like this after all the features have been put in it, and the scores have been evaluated.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34 entries, 0 to 33
Data columns (total 22 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Name             34 non-null     object 
 1   Username         34 non-null     object 
 2   FollowerCount    34 non-null     object 
 3   FollowingCount   34 non-null     object 
 4   ListedCount      34 non-null     object 
 5   Active           34 non-null     object 
 6   TotalTweets      34 non-null     object 
 7   ProfilePicture   34 non-null     object 
 8   Description      34 non-null     object 
 9   TweetList        34 non-null     object 
 10  TimeList         34 non-null     object 
 11  Retweets         34 non-null     object 
 12  Likes            34 non-null     object 
 13  RetweetCount     34 non-null     object 
 14  TweetFreq        34 non-null     object 
 15  FFRatio          34 non-null     object 
 16  ListedRatio      34 non-null     object 
 17  LikeCount        34 non-null     object 
 18  TwitScore        34 non-null     float64
 19  ProfileScore     34 non-null     float64
 20  FinancialScore   34 non-null     float64
 21  BehavioralScore  34 non-null     float64
dtypes: float64(4), object(18)
memory usage: 6.0+ KB
```

Table of Contents

1. Our Approach till Now
2. General Overview
3. Overcoming Limitations
4. Our Approach
5. Code Execution
6. Limitations
7. Improvements
8. Future Scope
9. Conclusion

Code Execution

Table of Contents

1. Our Approach till Now
2. General Overview
3. Overcoming Limitations
4. Our Approach
5. Code Execution
6. Limitations
7. Improvements
8. Future Scope
9. Conclusion

Limitations

In our current solution, we have taken relative values of the elements used for calculating the the Twit Score. The accuracy is limited by the number and variety of people, because it is relative.

Verification of the accuracy of correlation between the behavioral score and the actual credit score has not been done by us as it was not possible to get real credit scores for people on Twitter.

Twitter users are younger, more highly educated and have higher incomes than average adults overall (source: Pew Research). It can be said that our solution is slightly skewed towards such more affluent kind of people. There could potentially be a lot of low income people whose credit worthiness if we could determine and provide them credit, it would actually lead to development at the grassroots level.

Table of Contents

1. Our Approach till Now
2. General Overview
3. Overcoming Limitations
4. Our Approach
5. Code Execution
6. Limitations
7. Improvements
8. Future Scope
9. Conclusion

Improvements

Expanding to Instagram and Facebook

Instagram and Facebook have a more tightly knit network of users, compared to twitter where everyone can just follow anyone they wish to. They are also less skewed towards the more affluent people compared to Twitter.

Instagram and Facebook have similar concepts of posts, comments, activity, tags, mentions, etc. With minor modification to our current code, we can process that information too with relative ease. Other useful information such as educational qualifications, employment history, number of dependants can also be extracted, along with an estimation of their income levels based off geotags on posts, etc, leading to more accurate and well rounded overall score

Table of Contents

1. Our Approach till Now
2. General Overview
3. Overcoming Limitations
4. Our Approach
5. Code Execution
6. Limitations
7. Improvements
8. Future Scope
9. Conclusion

Future Scope

Applying network analysis using people with pre-existing credit scores in a person’s network, which we had previously proposed, as an additional element to the solution would yield even better results.

Creating clusters of people who already have credit scores, according to their interests data extracted from Instagram and Facebook, and other financial and personal data, and assigning new similar people to those clusters could result in us being able to determine scores of people in a more accurate and easier way as well.

The aforementioned clustering can have alternative use-cases like **recommender systems and social commerce applications** as well.

Analyzing a person’s transactions history, with their input on it, and using it in tandem with the social media score can help establish a tighter relationship between their social media behaviour and their financial attitudes.

Table of Contents

1. Our Approach till Now
2. General Overview
3. Overcoming Limitations
4. Our Approach
5. Code Execution
6. Limitations
7. Improvements
8. Future Scope
9. Conclusion

Conclusion

From running analysis on 34 samples, it can be deduced that a person with behavioral score around 150 or above has a good level of trust and affluence.

Our solution is proof of our concept that data extracted from social media can contribute in predicting a person's credit worthiness.

We believe that we are aware of our shortcomings; and with enough time in our hands, and proper resources, such as a means of collecting user social media data, the improvements and future scope of our solution is not at all a pipe dream, and is something we could totally achieve.

Thank You