BTP PHASE-I PRESENTATION

RELIABILITY DIMENSION IN REVIEW PLATFORM YELP

Submitted By :

Akash Mahalik (150123004)

Tarun Genwa (150123043)

Under the guidance of :

Dr. Ayon Ganguly

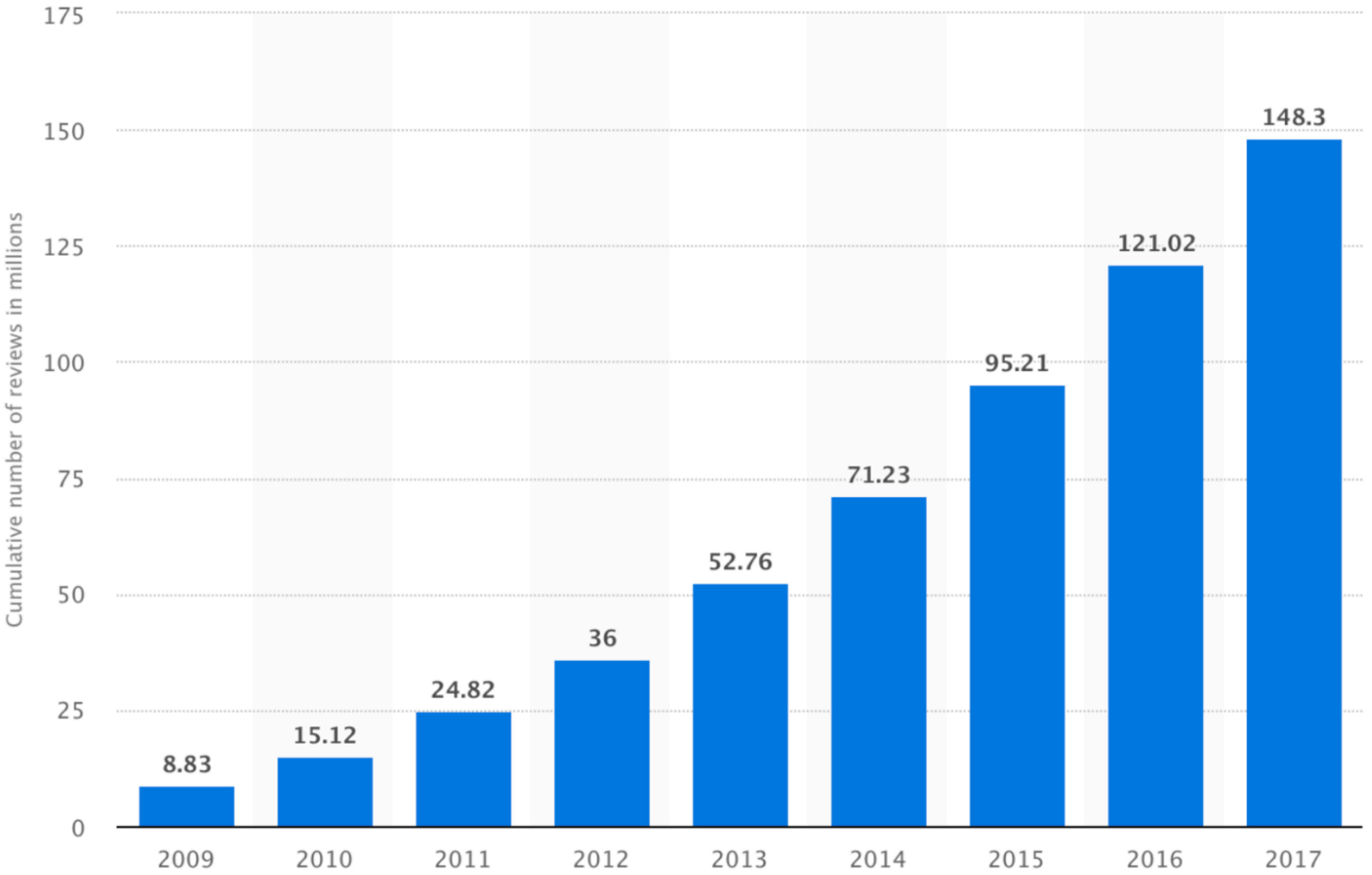
What Is Reliability Analysis of Big Data Systems?

- Reliability is the degree to which an assessment tool produces stable and consistent results.

Choosing Yelp as the Big Data Source

- Yelp is crowd sourced review platform listing hundreds of thousands of local and chain businesses.
- The motivation for choosing Yelp dataset as our big data source was due to the effect that it carries on mass opinion and hence comes out to be a role player in business growth.
- The availability of massive datasets open sourced by Yelp itself for running various dataset challenges.

TOTAL REVIEWS ON YELP



CONSTITUENTS OF RELIABILITY OF RATING MODEL OF BUSINESS

- User Credibility
- Topic Modelling of Review Text
- Geospatial Analysis

Creating a Credibility Score of a User

- Credibility score of a user will help in deciding how relevant a user's review is.
- It would help in taking decisions while filtering fake and biased reviews.
- It also helps in understanding the eliteness levels of different users.

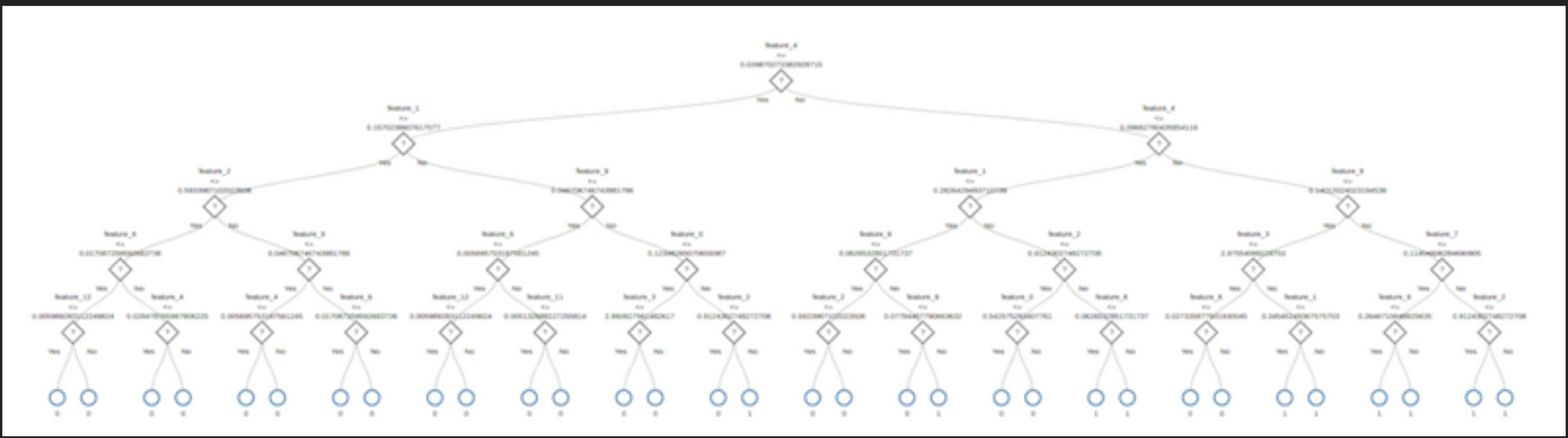
Approach to Credibility ??

- Credibility score is defined by various features in user dataset.
- The dataset already labels a user as elite or not.
- We can leverage machine learning techniques to assign feature importance and decide eliteness of a user.


```
"user_id" : "lzlZwIpuSWXEnNS91wxjHw",
"name" : "Susan",
"review_count" : "1",
"yelping_since": "2015-09-28",
"friends" : "None",
"useful" : "0",
"funny" : "0",
"cool" : "0",
"fans" : "0",
"elite" : "None"
"average_stars" : "2"
"compliment_hot" : "0",
"compliment_more" : "0",
"compliment_profile" : "0",
"compliment_cute" : "0",
"compliment_list" : "0",
"compliment_note" : "0",
"compliment_plain" : "0",
"compliment_cool" : "0",
"compliment_funny" : "0",
"compliment_writer" : "0",
"compliment_photos" : "0"
```

Random Forest Classifier

- Random Forest Classifier is basically an ensemble machine learning which combines weaker predictive models into stronger one by constructing a multitude of decision trees and generalising the model.
- We use Random Forest to obtain importance of the different features in the user dataset and thus calculate the eliteness score.



Calculating Eliteness Score

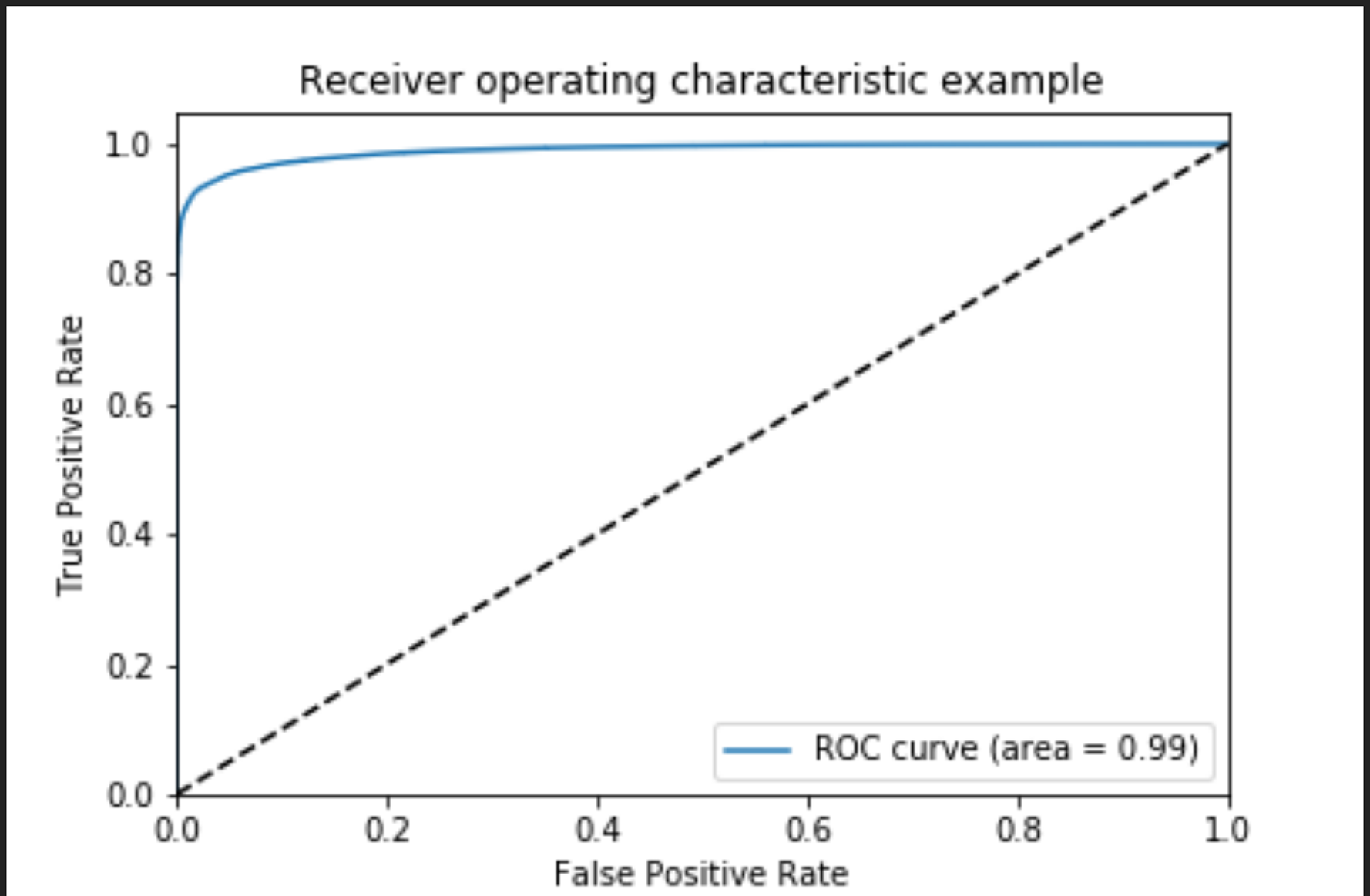
- Given a user it's eliteness score will be dot product of corresponding feature value with its importance.

Features	Values
compliment_writer	0.242482
review_count	0.195236
compliment_cool	0.186909
fans	0.184319
compliment_funny	0.071866
compliment_hot	0.049423
compliment_note	0.037891
cool	0.019294
no_friends	0.004900

Features	Values
average_stars	0.004348
compliment_photos	0.001833
compliment_plain	0.000947
compliment_cute	0.000281
funny	0.000159
compliment_list	0.000057
useful	0.000049
compliment_profile	0.000005

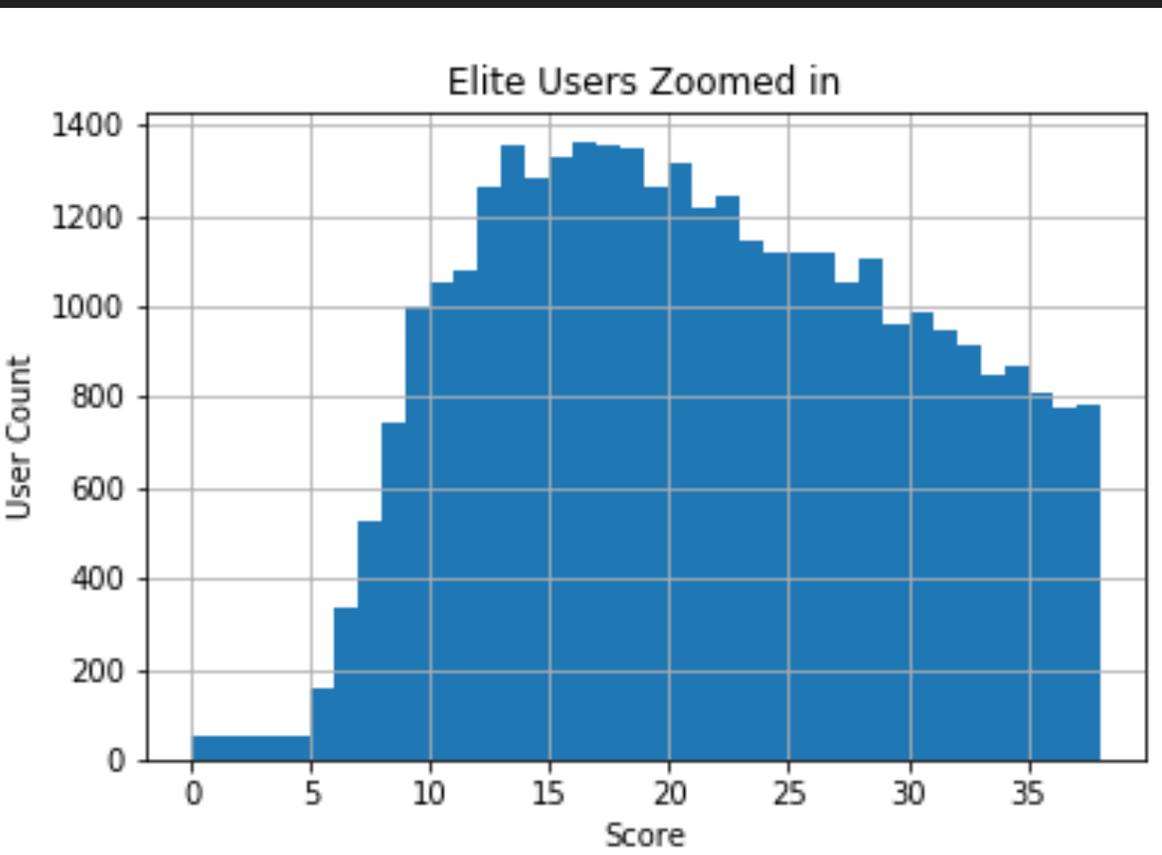
Reliability of Feature Importance

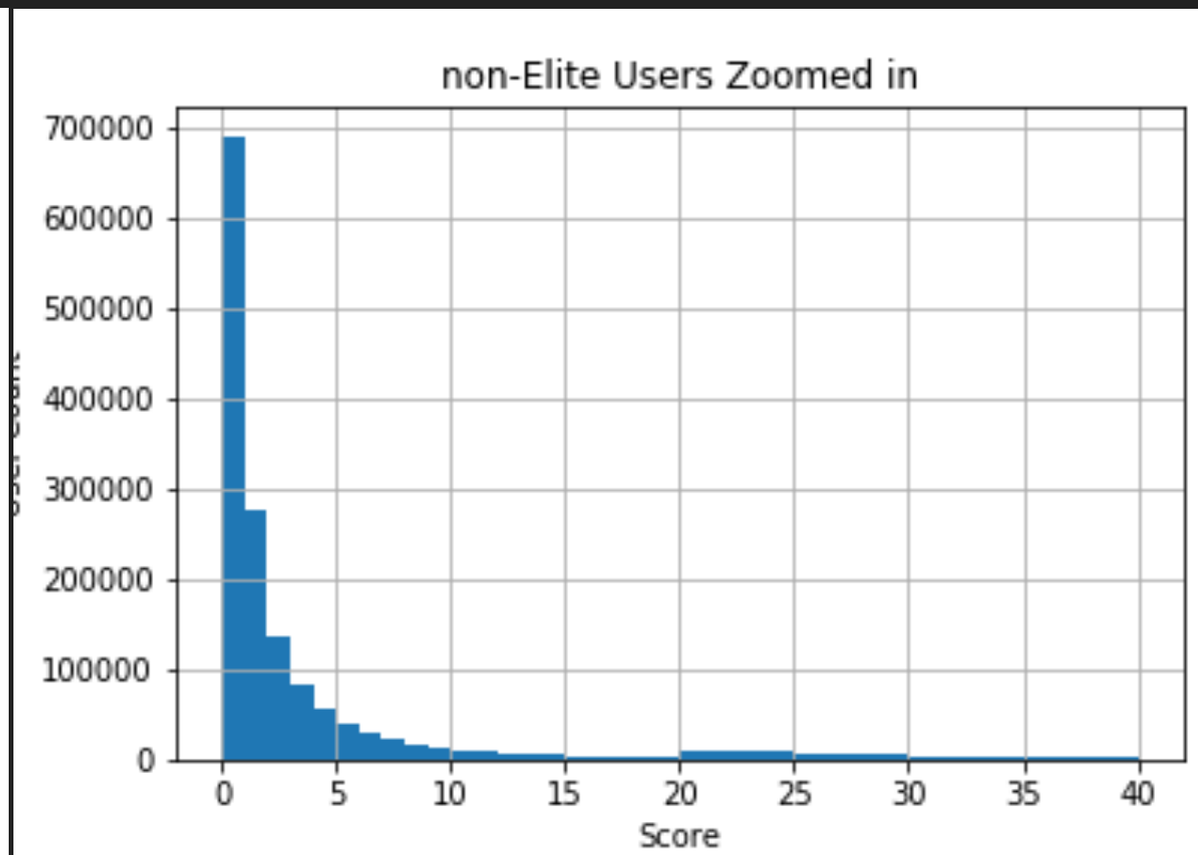
- The random forest model we trained has an accuracy of 98.21% and thus can be seen from the ROC curve.



RESULTS

- Only 4.62% of the total users have been labelled “elite” in the dataset.
- It can be clearly seen from the spike in non-Elite score distribution that majority of the users have low score.





TOPIC MODELLING