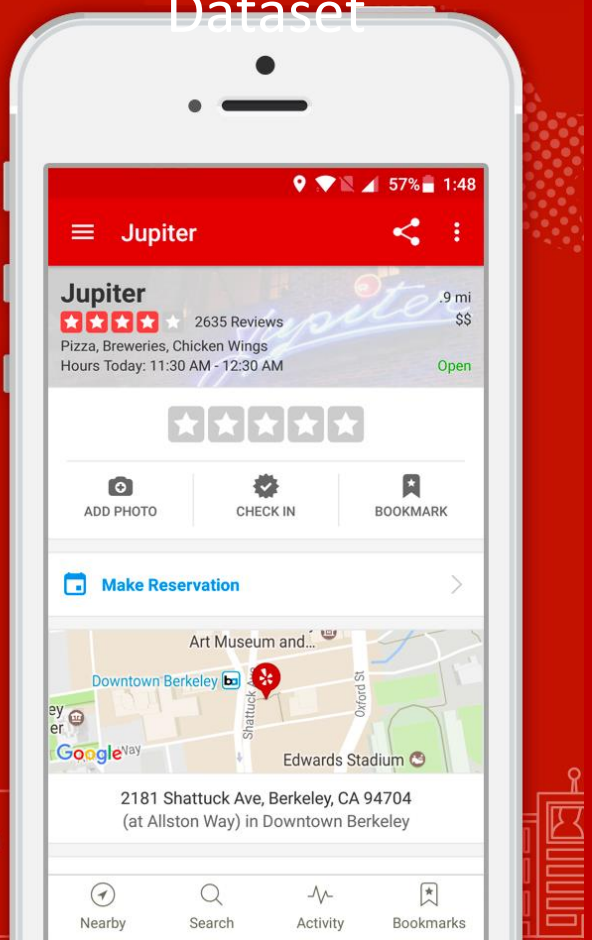
242 - Data Modelling

Project Presentation

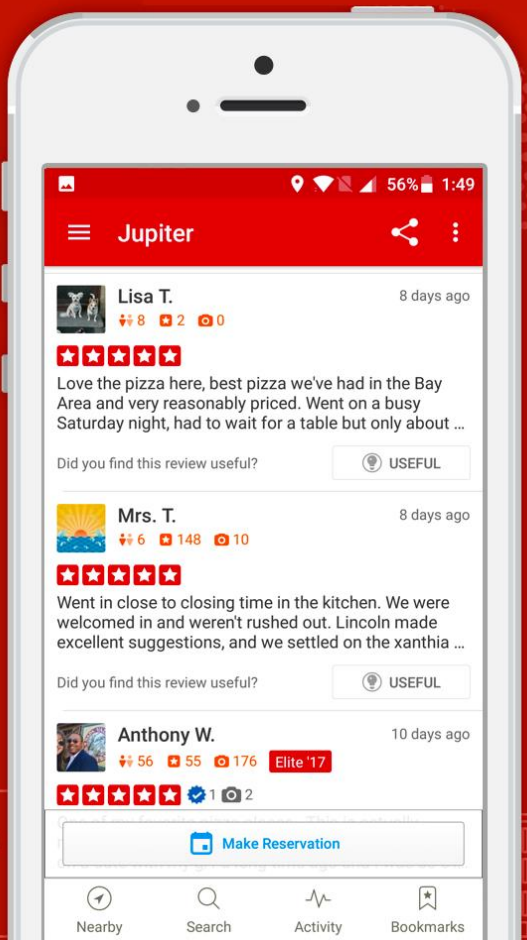
Yelp Dataset

NORTHWEST
MISSOURI STATE UNIVERSITY

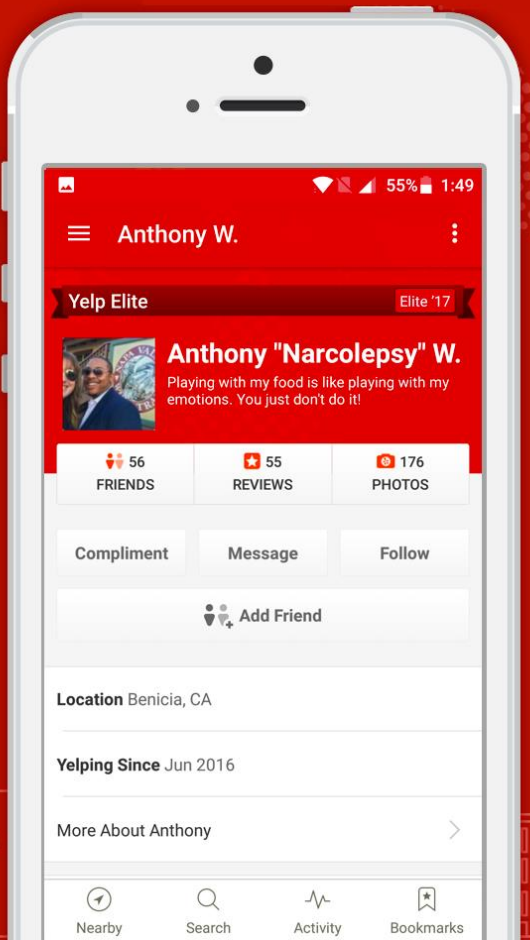
Business Dataset



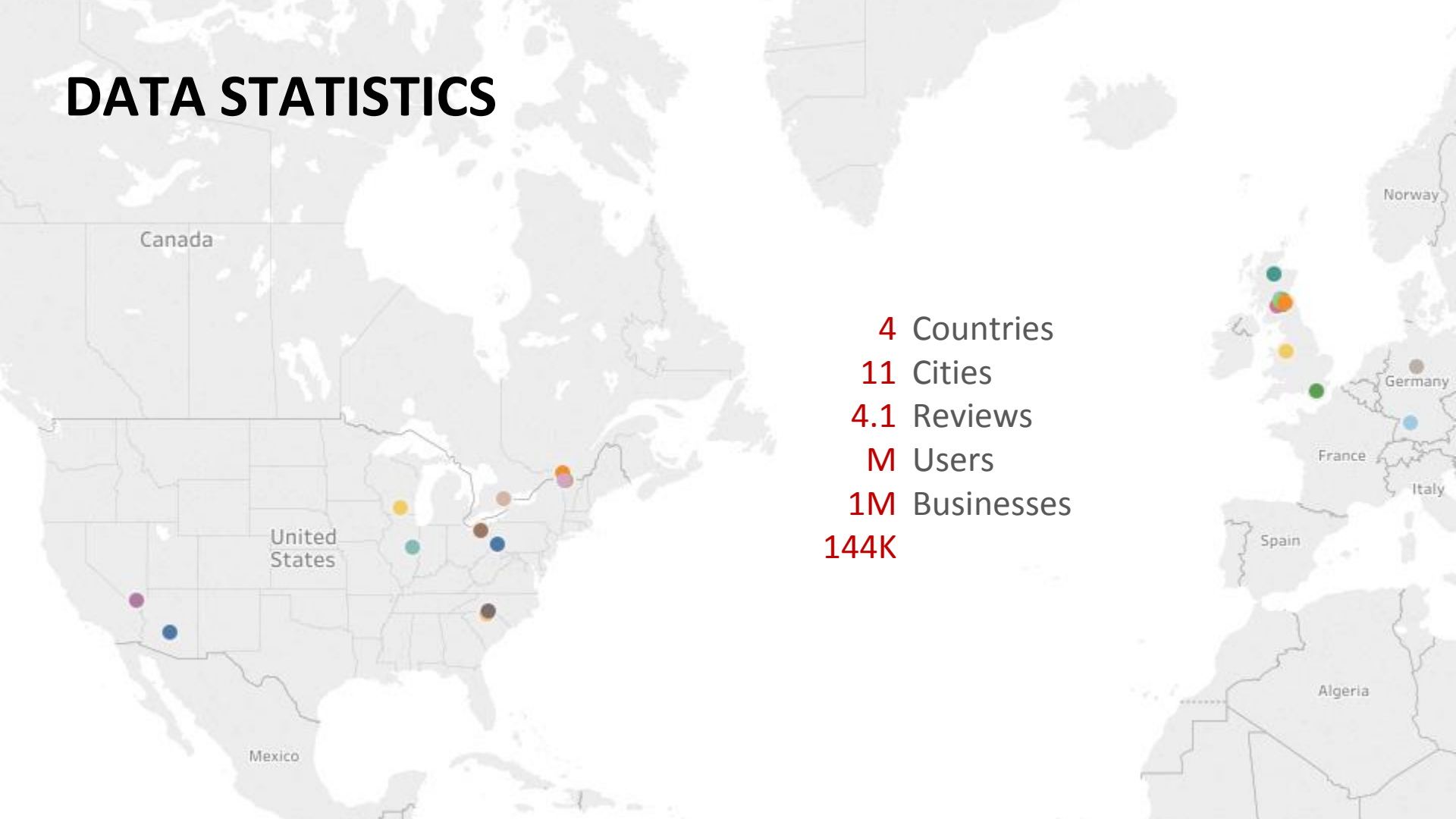
Review Dataset



User Dataset



DATA STATISTICS

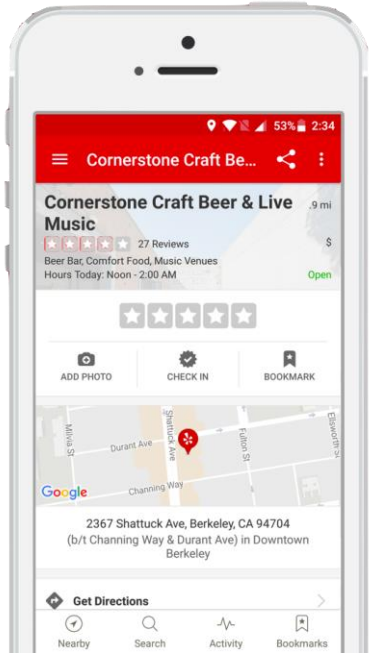


4 Countries
11 Cities
4.1 Reviews
M Users
1M Businesses
144K

PROJECT OBJECTIVE - BUSINESS SOLUTION

NEW BUSINESS

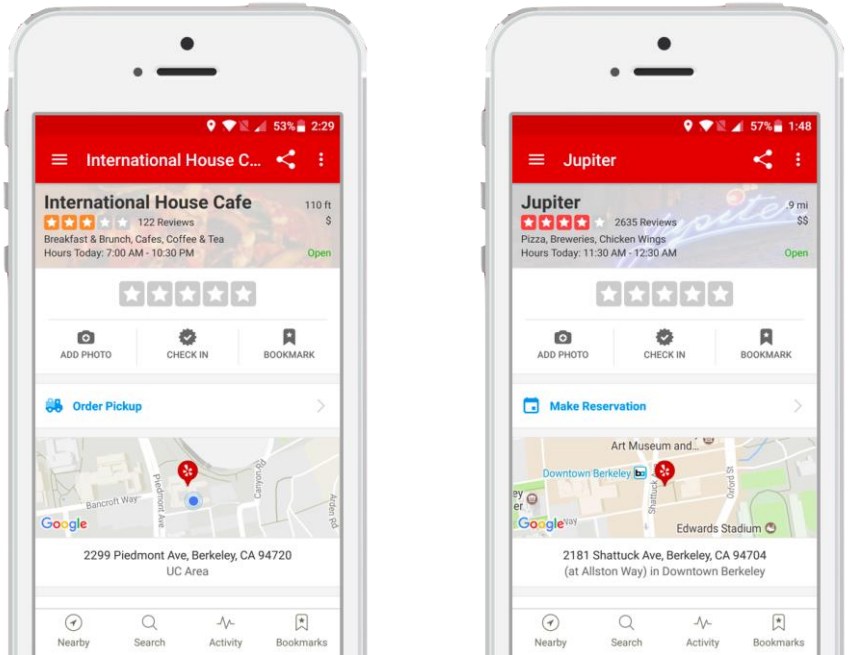
" DETERMINE THE RATING OF A BUSINESS "



PROJECT OBJECTIVE - BUSINESS SOLUTION

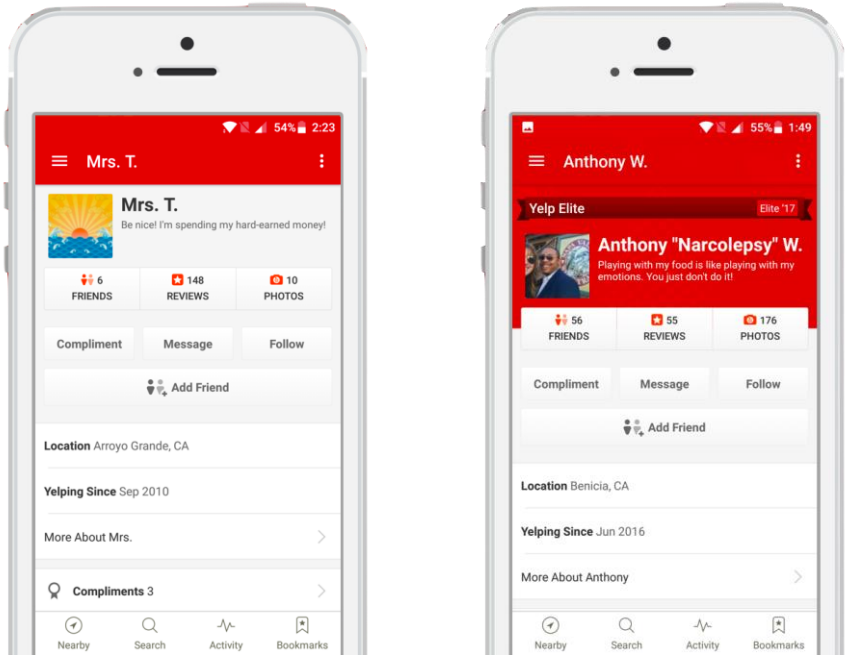
EXISTING BUSINESS

" IMPROVE THE RATING OF A BUSINESS "



PROJECT OBJECTIVE - USER SOLUTION

“ DETERMINE WHEN A CUSTOMER BECOMES AN ‘ELITE’ CUSTOMER ”



NEW BUSINESS : " DETERMINE THE RATING OF A BUSINESS "

Why is this relevant? - People want to have a good business model

Data:

(a) Business Attributes - 80+ business attributes separated out
Examples - Wifi, Pets allowed, Parking, etc.

(b) Business Categories - 1000+ unique categories
Examples - Restaurant, Shopping, Nightlife, etc.

(c) Business Hours - Derived opening hours, closing hours, open hours etc.

Data Processing: Availability of Data

Business Attributes

Feature	Availability
business_id	100
is_open	100
latitude	100
longitude	100
name	100
review_count	100
stars	100
state	100
type	100
garage	64.23801988
validated	63.63554334
BusinessAcceptsCreditCards	61.79687934
GoodForKids	42.81262147
RestaurantsTakeOut	42.77097562
OutdoorSeating	40.77336332
live	40.39785663
neighborhood	40.19448609
RestaurantsGoodForGroups	40.14104059
RestaurantsDelivery	38.15245155
WiFi	34.81523127
NoiseLevel	34.43556
casual	34.407102
RestaurantsTableService	34.08226442
romantic	33.39024932
BikeParking	32.65867066
Caters	27.94297296
ByAppointmentOnly	26.77341885
divey	25.60733522
DogsAllowed	14.96751624
GoodForDancing	13.61124993
HappyHour	13.55155755
CoatCheck	13.31348215

Business Categories

Feature	Availability
address	97.06119163
attributes	88.26281304
Restaurants	33.65331223
Shopping	15.59359209
Food	14.70722972
Beauty & Spas	9.516769393
Home Services	7.802348826
Nightlife	7.304680993
Health & Medical	7.271364318
Bars	6.307263035
Automotive	5.937309123
Local Services	5.64509412
Event Planning & Services	5.014159587
Active Life	4.665722694
Fashion	4.042423233
American (Traditional)	3.687045366
Fast Food	3.644011328
Pizza	3.629435282
Sandwiches	3.623188406
Coffee & Tea	3.539202621
Hair Salons	3.371925149
Hotels & Travel	3.371231051
Arts & Entertainment	3.276833805
Italian	2.858293076
Auto Repair	2.847187517
Home & Garden	2.841634738

NEW BUSINESS

Method Implemented:

Classification (CART)

Prediction: The CART model always predicted a rating of 4

Reasons for failure:

(a) The model could not find any relevant relations between the business attributes and the ratings

(b) The model just predicted based on a simple probability

Rating	1	1.5	2	2.5	3	3.5	**4**	4.5	5
Probability	0.008	0.018	0.050	0.093	0.168	0.220	**0.237**	0.136	0.070

NEW BUSINESS

Cannot find a causal relationship

Real world scenario:
What matters is **CUSTOMER EXPERIENCE** !!

EXISTING BUSINESS : " IMPROVE THE RATING OF A BUSINESS "

Why is this relevant?

- Businesses want to improve their ratings, and in turn understand their customers
- Yelp can consult businesses and make money!

Data:

(a) Not much relation with business attributes

(b) Reviews:

  (i) 'Bag of words' to come up with top influencers

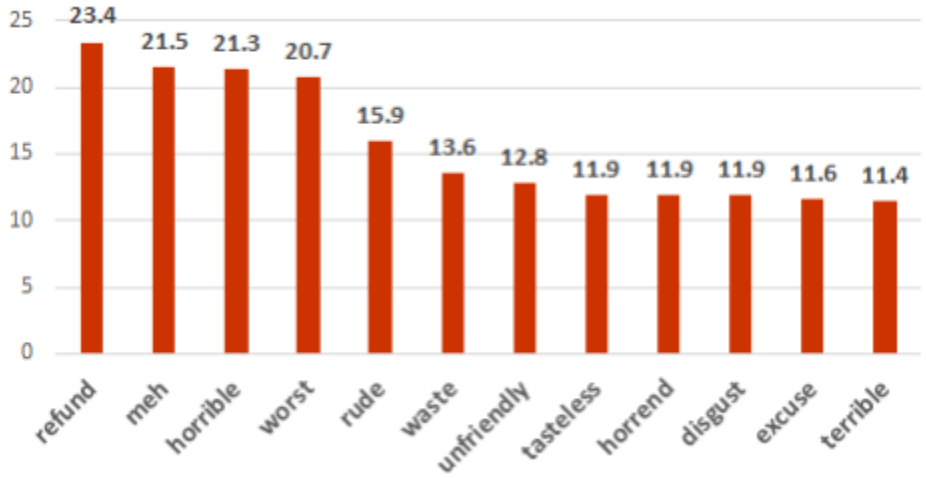
  (ii) Assigning sentiment to reviews (1 if rating > 3.5 ; 0 if rating < 3.5)

Data Processing: Word Cloud

SOME COMMON OPERATIONS:

- Upper case to lower case
- Removed punctuations and any other symbols
- Removed stop words
- Stemming, lemmatization

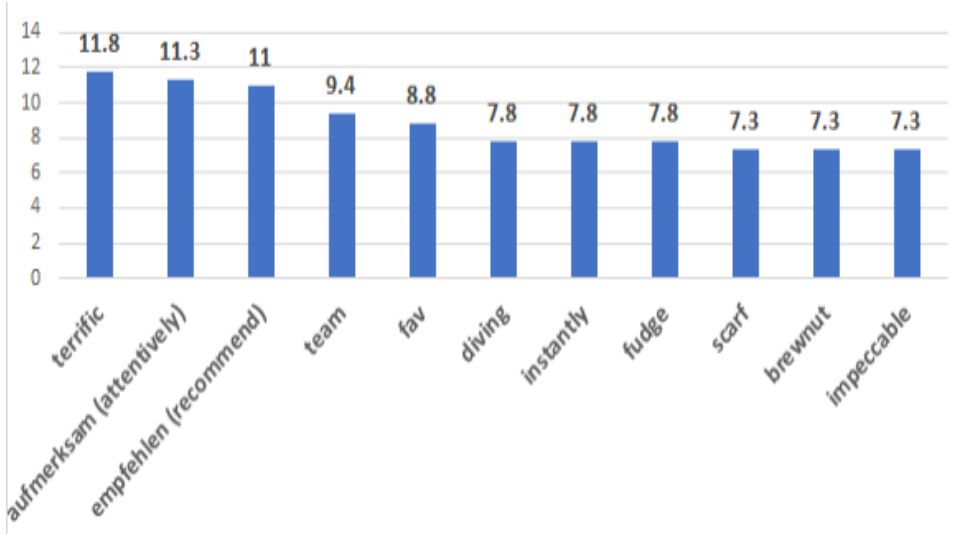
Data Processing: Most common words



Reviews with rating < 3.5

Most common words: Refund, ‘Meh’, Horrible, Worst, Rude, Unfriendly

Data Processing: Most common words



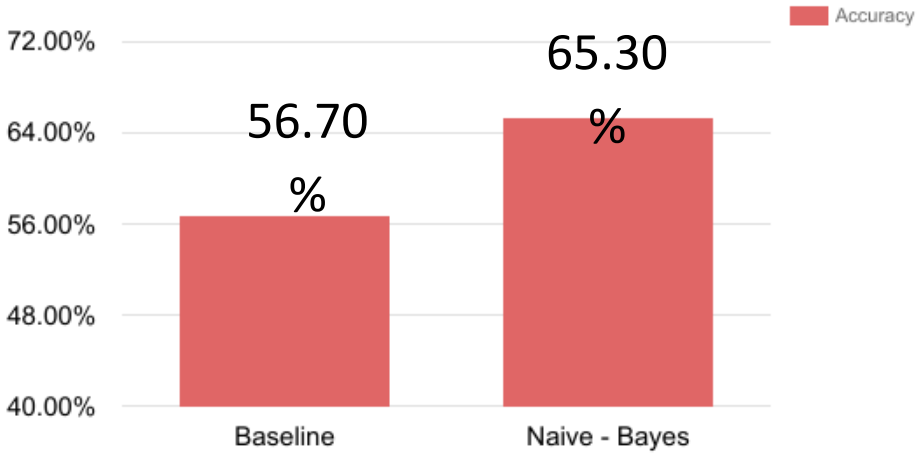
Reviews with rating > 3.5

Most common words: Terrific, ‘Fav’, Recommend, Impeccable, Attentive

EXISTING BUSINESS

Method Implemented:

Classification



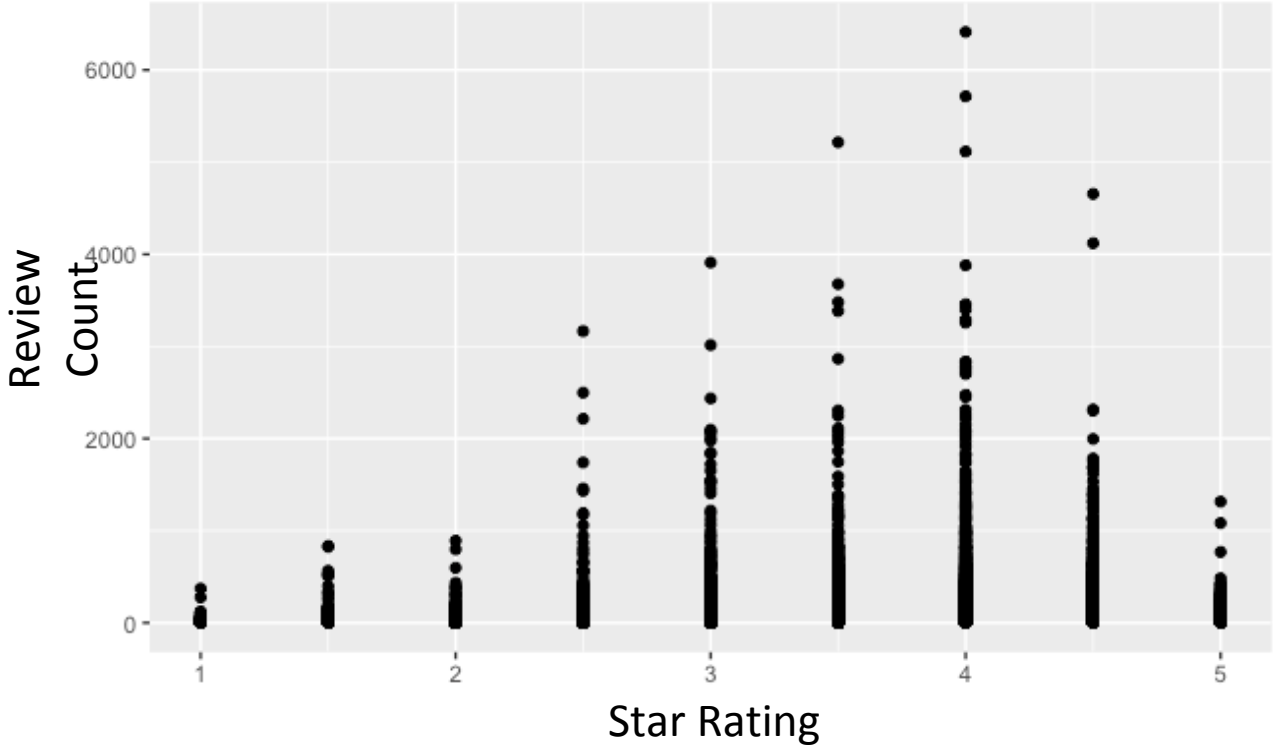
EXISTING BUSINESS

Reviews form a significant feature for predicting ratings of a business

Real world scenario:
What matters is **CUSTOMER EXPERIENCE** !!

INSIGHTS

As the number of reviews go up, the ratings tend to stabilize.



USER SOLUTION : " DETERMINE 'ELITE' CUSTOMERS "

Why is this relevant? - As Yelp puts it " It's neat to be elite! "

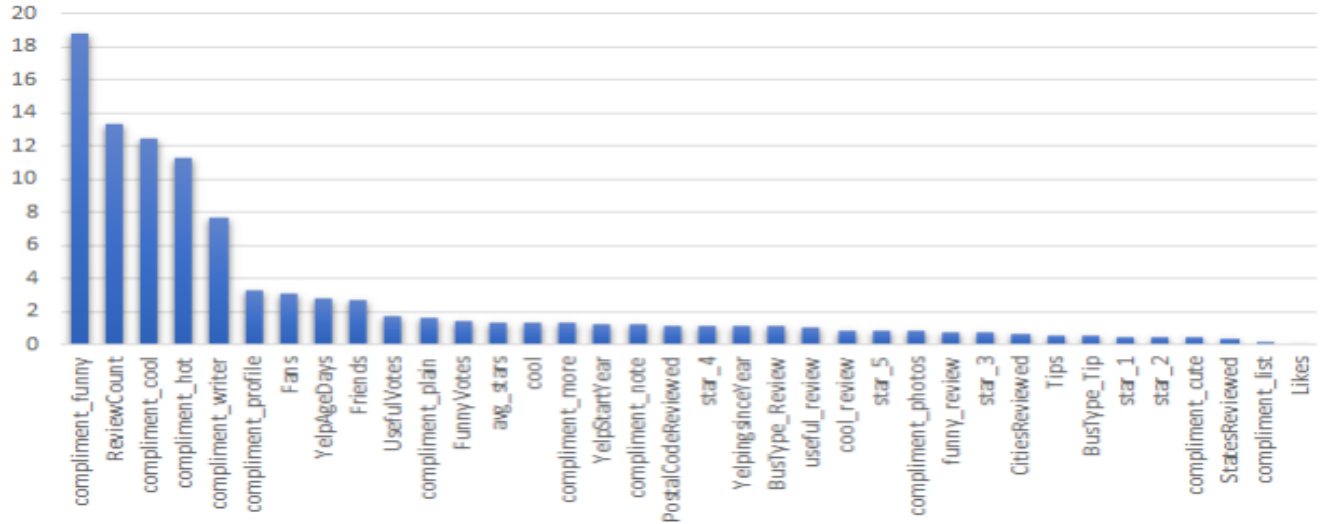
But generally speaking

- Users want the benefits of being elite
- Businesses want to target the people who are elite or are about to be elite
- Yelp wants to maintain the quality of elite users

Data:

(a) User Attributes - Friends, Fans, Reviews given, Photographs, etc.
(b) Compliments Received - Hot, Cool, Funny, etc.
(c) Votes given - Funny, Useful, etc.

Data Processing: Importance Analysis

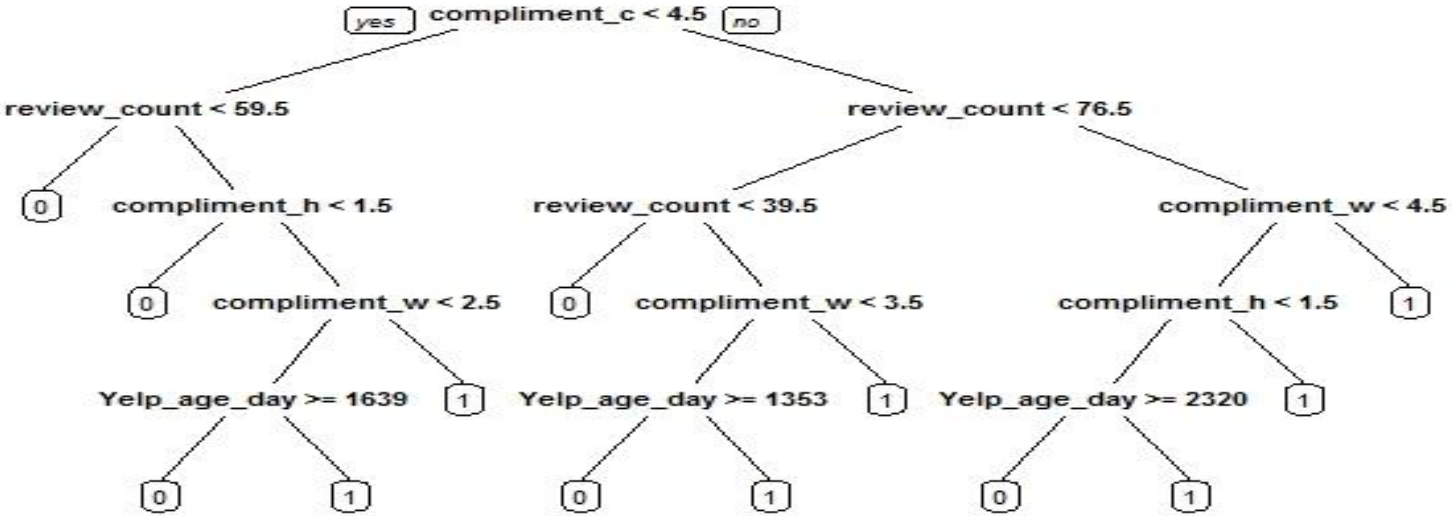


Most important features: Compliments, Review counts, Fans, Friends

USER SOLUTION

Method Implemented:

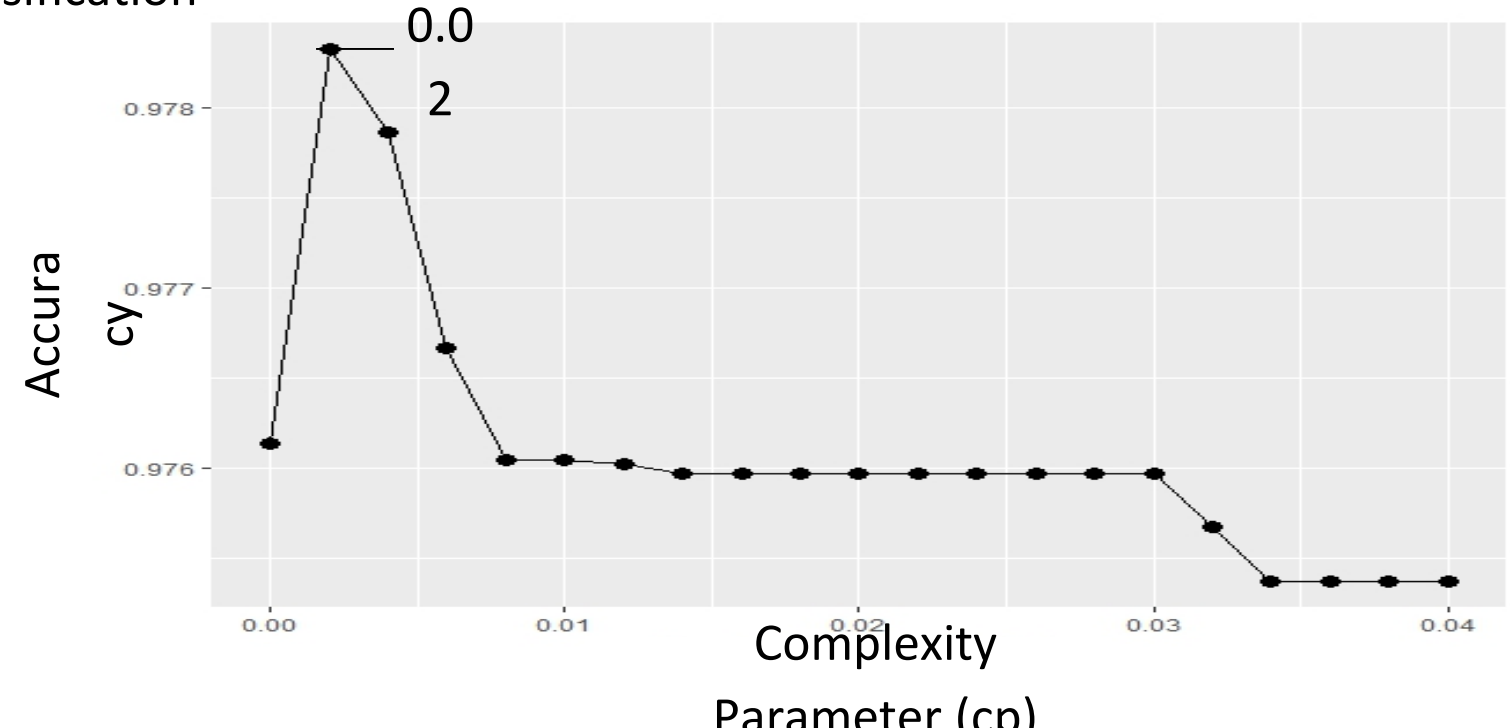
Classification



USER SOLUTION

Method Implemented:

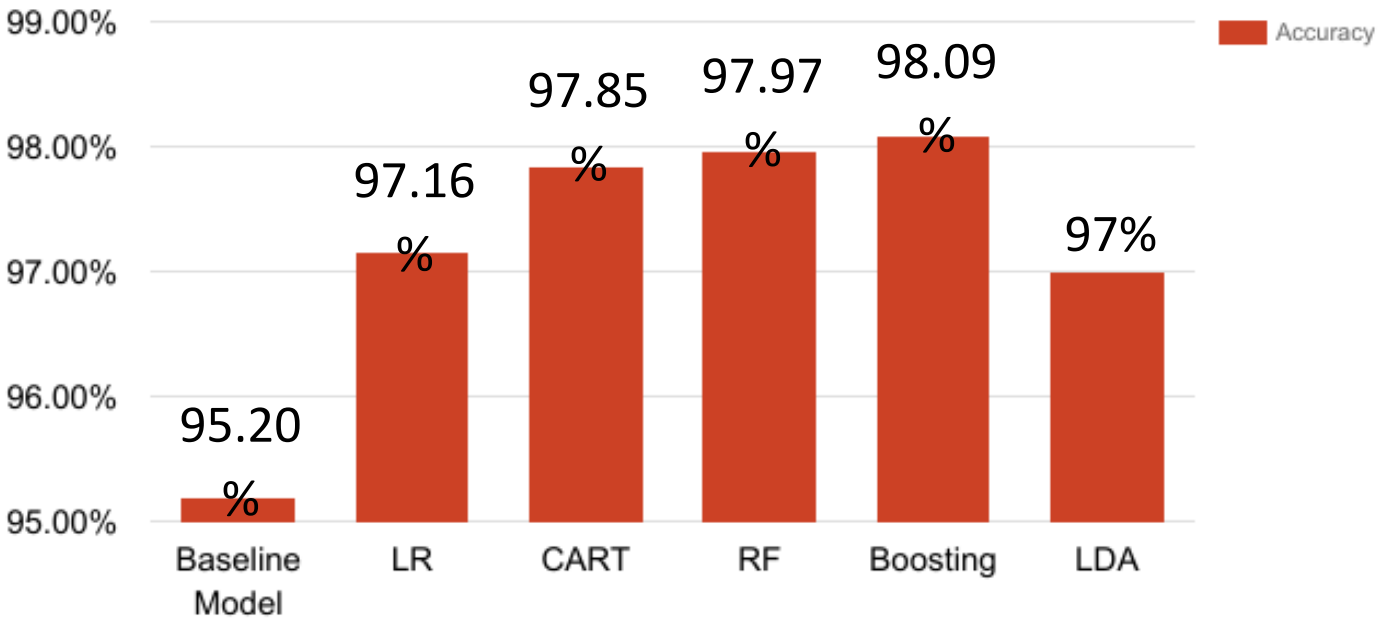
Classification



USER SOLUTION

Method Implemented:

Classification



USER SOLUTION

Model Statistics

Models	Accuracy	TPR	TNR
Baseline	95.20%	0%	95.2%
Logistic Regression	97.16%	58.67%	99.09%
CART	97.85%	**72.38%**	99.13%
Random Forest	**97.97%**	71.85%	**99.29%**
Boosting	**98.09%**	**76.93%**	99.16%
LDA	97%	51.61%	**99.3%**

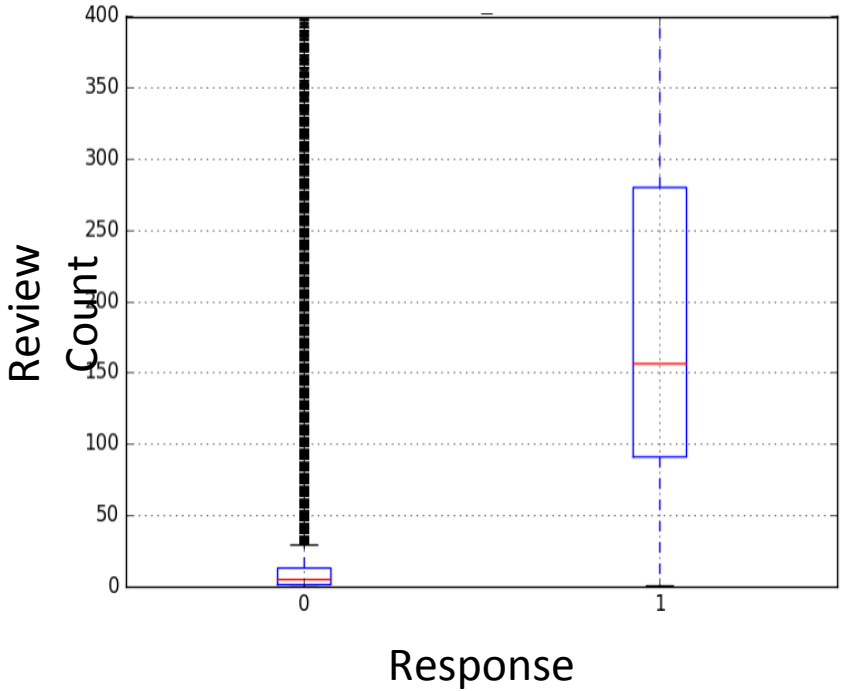
USER SOLUTION

Boosting gives a good prediction

So contact us if you want to become an ‘ELITE’ user.

INSIGHTS

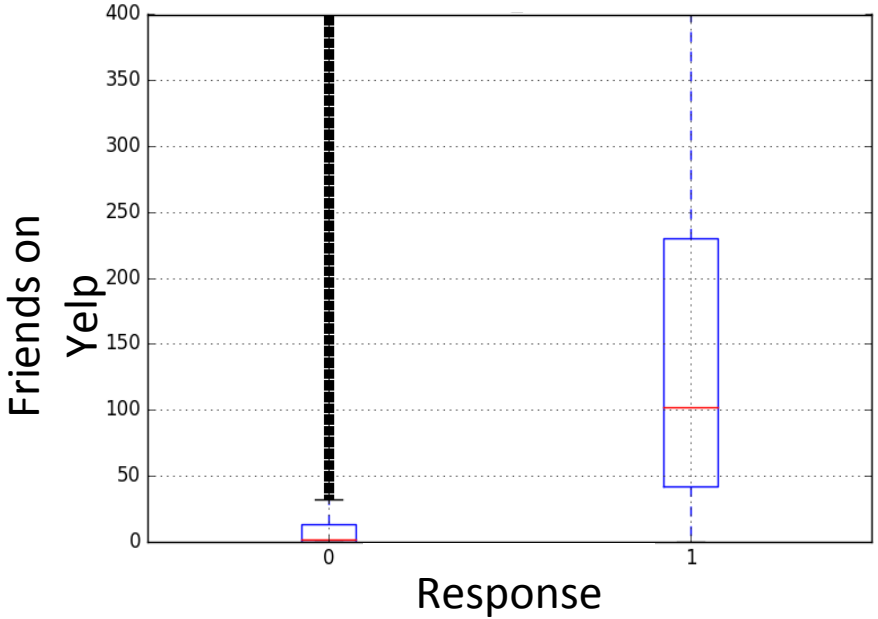
Quality over quantity.



Even though some users have reviews in 1000's but the median for the number of reviews needed to become an 'Elite' customer is 157.

INSIGHTS

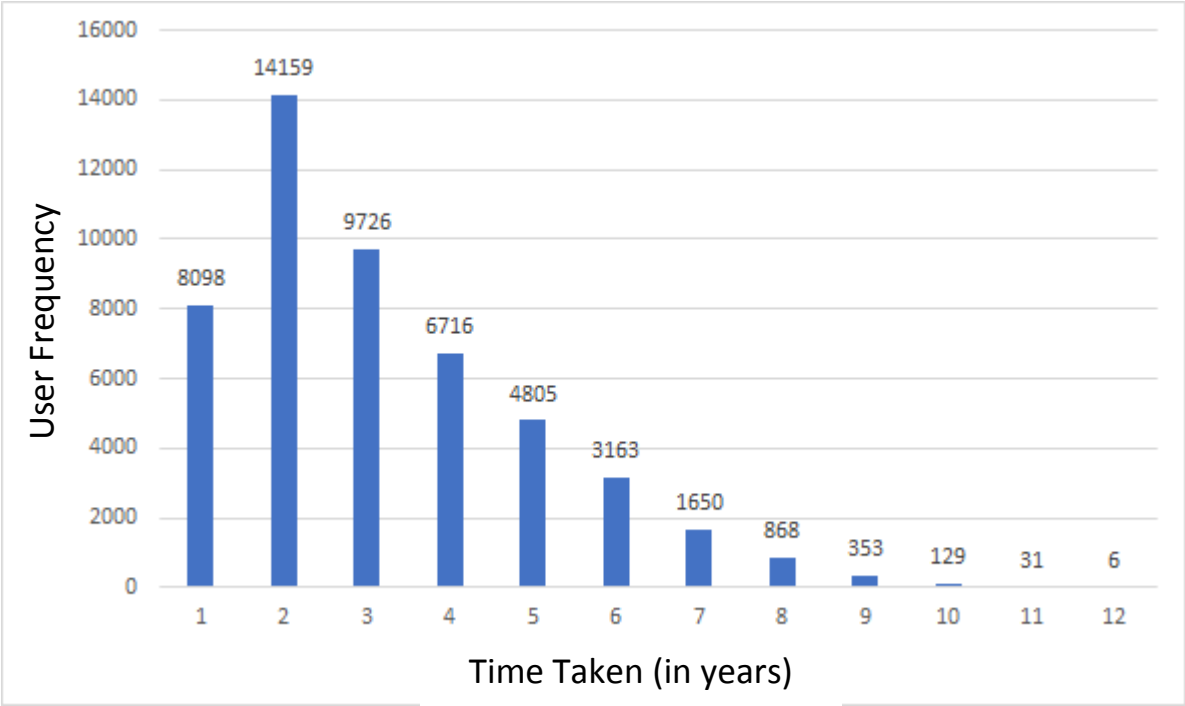
Friends Matte



The Friends Network is highly related to the user being an ‘Elite’

INSIGHTS

Age ≠ Wisdom.



The Yelp Age is not directly proportional to the user being an ‘Elite’

CONCLUSION

We want Yelp to have more insight into businesses and users.

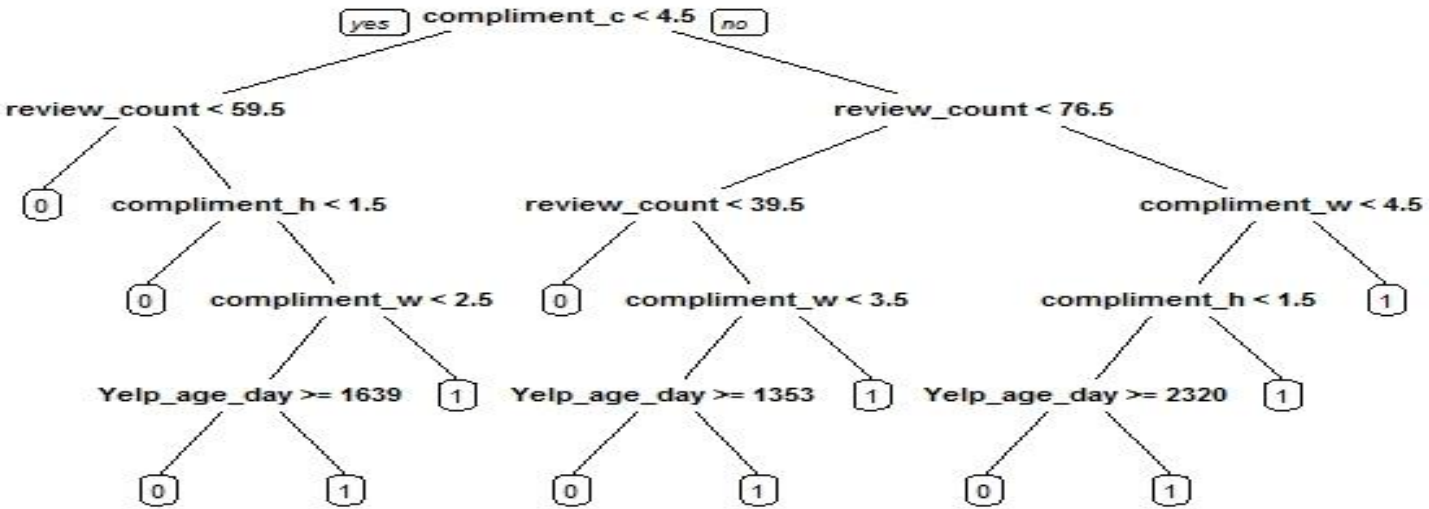
- Business attributes do not have causal relationship with the ratings of a business
- User reviews matter the most for a business rating
- ‘Elite’ user reviews attract more customers

APPENDIX

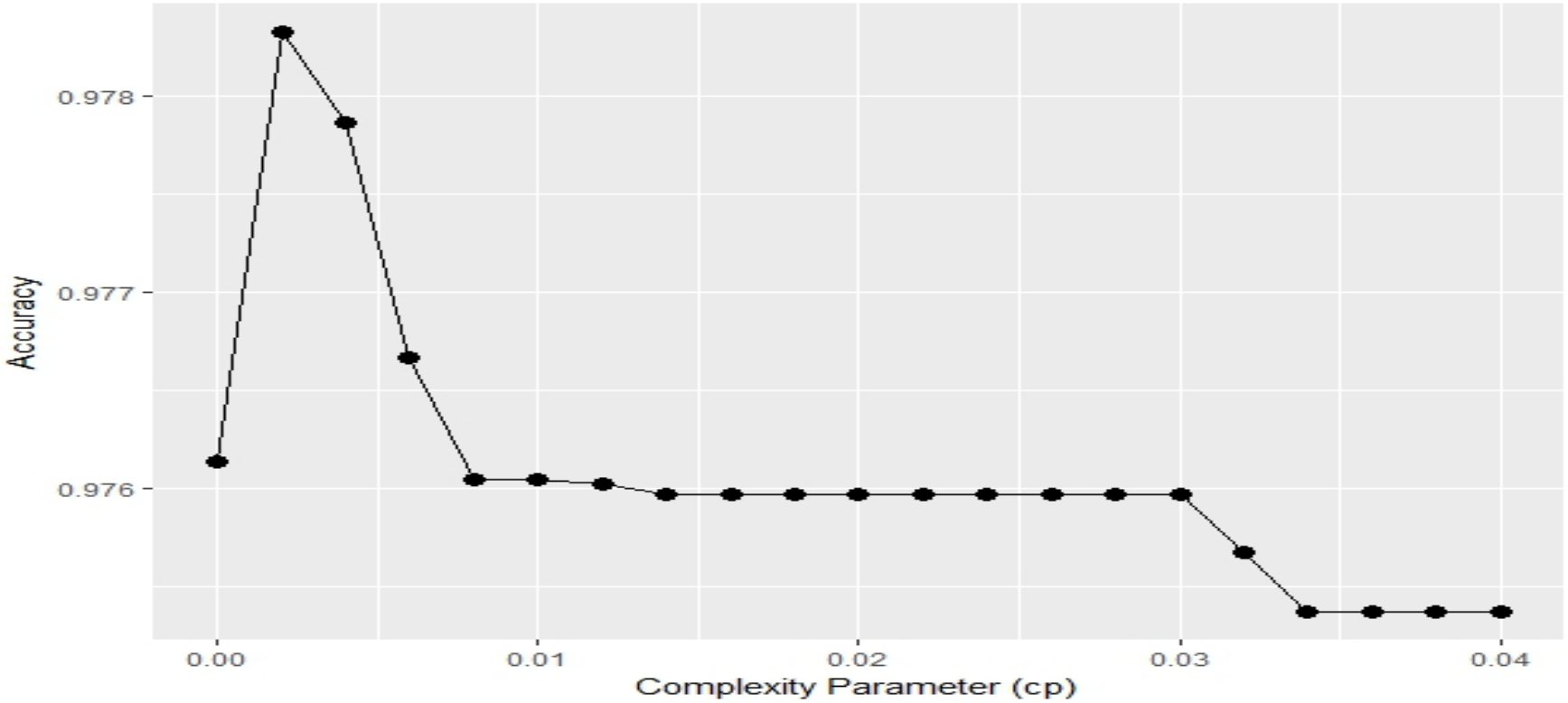
CART model for New Businesses

```
> summary(train.cart)
Call:
rpart(formula = stars_x ~ ., data = X_train, method = "class")
  n= 64440

          CP nsplit rel error xerror xstd
1 0.007880501      0         1      0    0

Node number 1: 64440 observations
  predicted class=4  expected loss=0.7630664  P(node) =1
    class counts:   510  1133  3251  5975 10856 14172 15268  8738  4537
   probabilities: 0.008 0.018 0.050 0.093 0.168 0.220 0.237 0.136 0.070
```

DECISION TREE



CROSS VALIDATION: BEST CP VALUE



LINEAR DISCRIMINANT ANALYSIS vs LOGISTIC REGRESSION