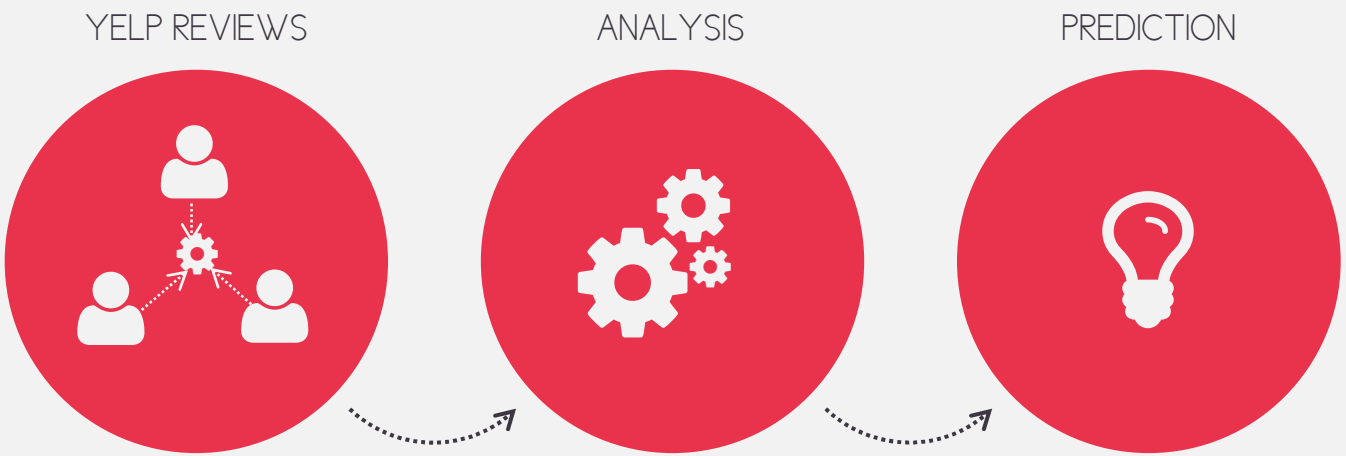
YELP

Customer Ratings Prediction

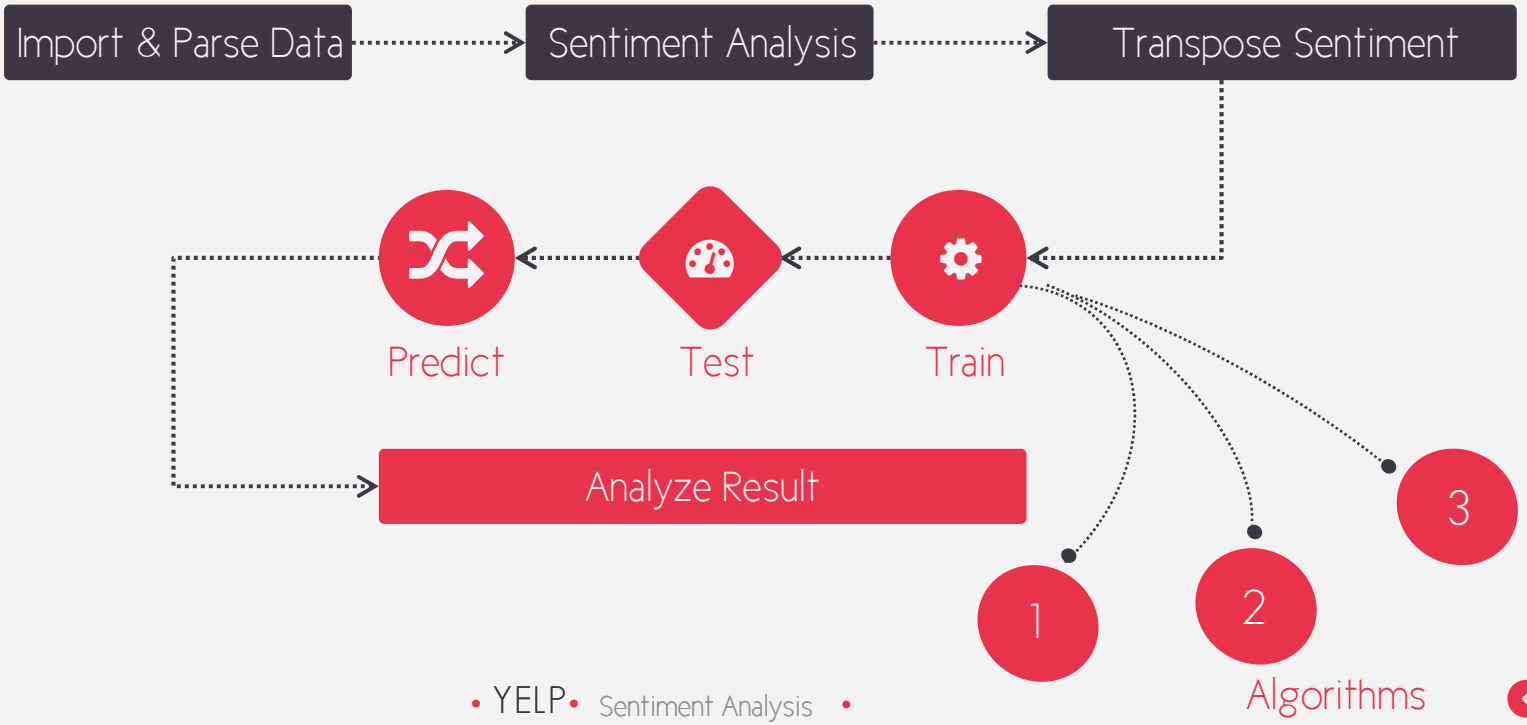
By
Robert D. Driesch
Todd Miller
Parastoo Karacic
Shraddha Mandhale
Suraj Jois
Princewill Eneh

Project Overview

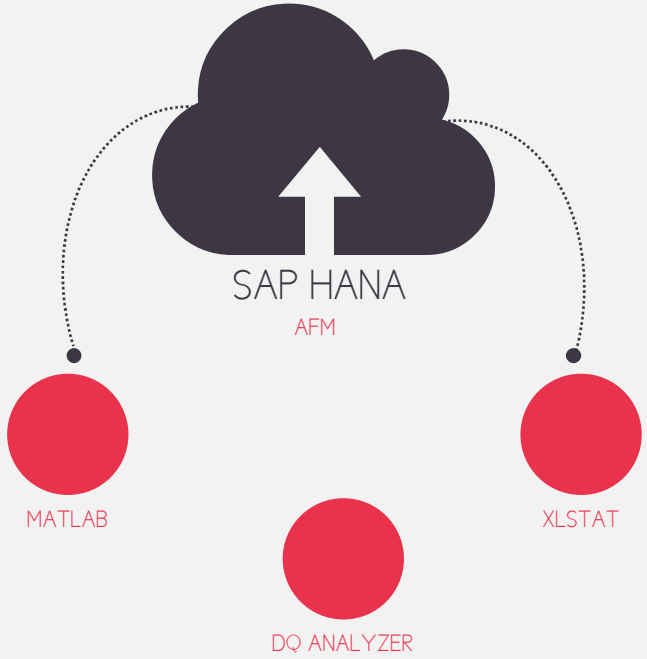
Objective : To build a model to help predict the star ratings a business is likely to receive based upon previous customer written reviews while taking into account differences observed within the data corpus for different geographies.



Workflow


Import & Parse Data
Sentiment Analysis
Transpose Sentiment
Predict
Test
Train
Analyze Result
1
2
3
Algorithms

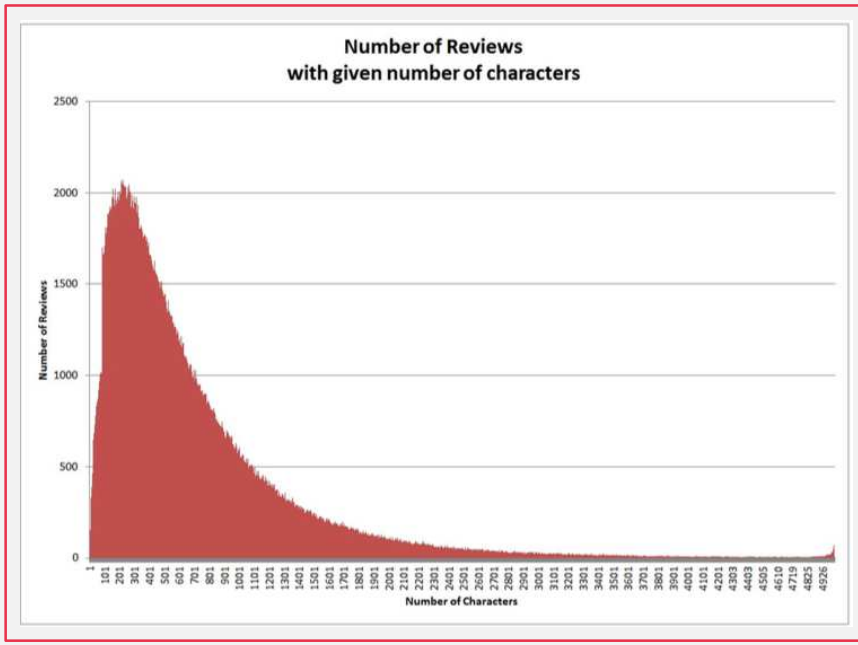
Tools



• YELP • Sentiment Analysis •

Preliminary Analysis

Data Profiling

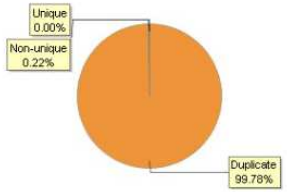


Basic | Frequency | Quantiles | Groups

Basic Analyses

Expression: REVIEW_DATE
Data type: DATETIME
Domain:
Rows: 1,569,264

Counts

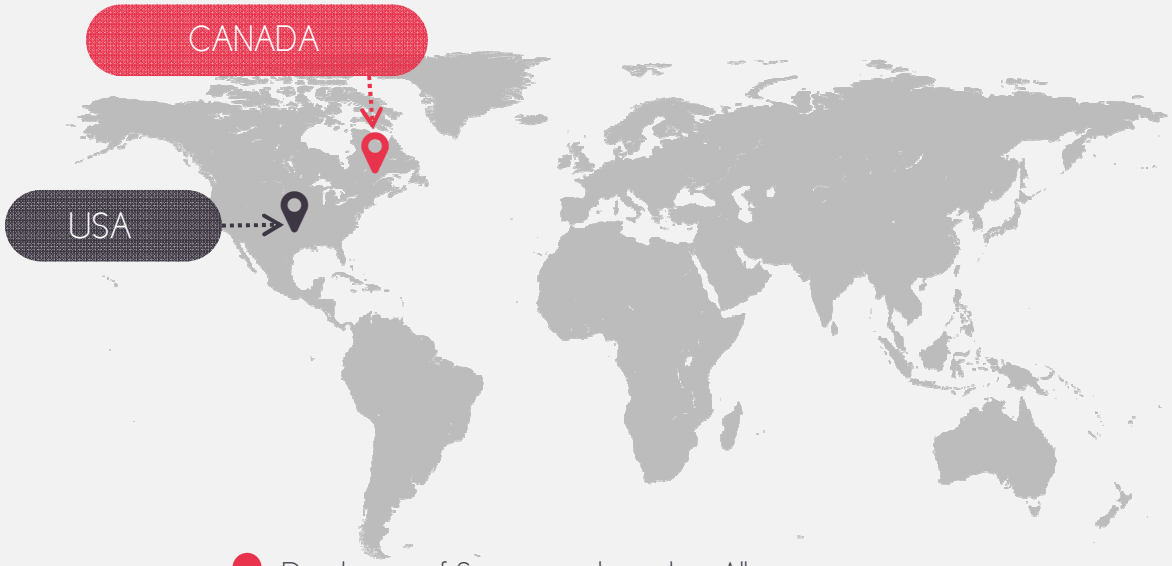


Type	Count	%
Null	0	0.00%
Non-null	1,569,264	100.00%
Duplicate	1,565,819	99.78%
Distinct	3,445	0.22%
Non-unique	3,388	0.22%
Unique	57	0.00%

Statistics

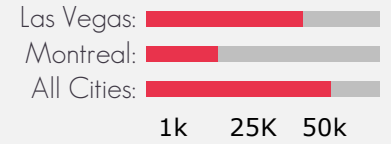
Type	Value	Frequency
Minimum value	2004-10-12 00:...	2
Median value	2013-03-04 00:...	956
Maximum value	2015-01-08 00:...	1,834

Type	Value
Average	2012-09-23 18:15:15.000
Variance	439,404.79
Standard devia...	662.88



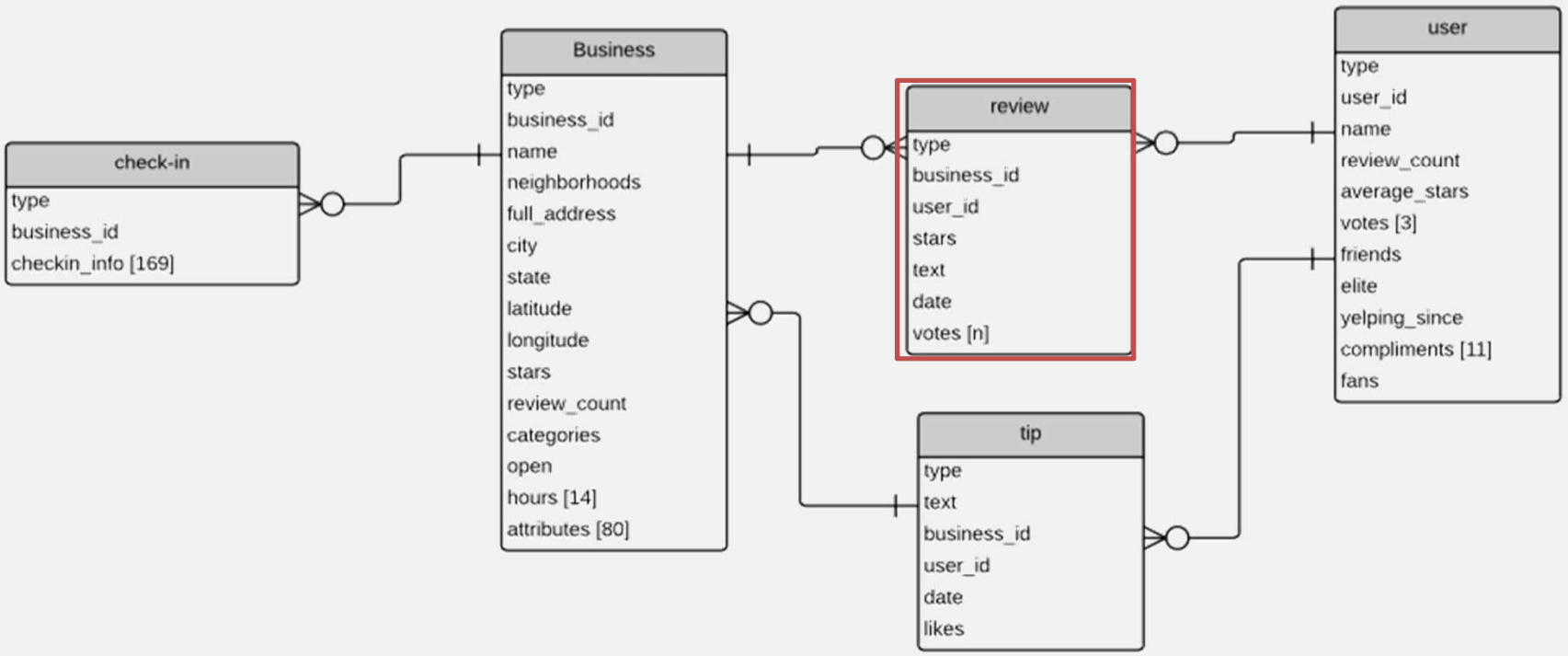
Distribution of Star ratings based on All reviews

CLUSTER	ONE_STAR	TWO_STAR	THREE_STAR	FOUR_STAR	FIVE_STAR
URBANA-CHAMPAIGN	0.07%	0.08%	0.12%	0.23%	0.27%
KARLSRUHE	0.01%	0.02%	0.03%	0.05%	0.06%
EDINBURGH	0.05%	0.11%	0.31%	0.64%	0.41%
LAS VEGAS	4.62%	3.97%	6.37%	12.51%	15.83%
MONTREAL	0.20%	0.25%	0.51%	1.16%	1.04%
WATERLOO	0.02%	0.02%	0.03%	0.06%	0.05%
CHARLOTTE	0.54%	0.58%	1.04%	2.14%	1.96%
PHOENIX	4.13%	3.27%	4.69%	10.68%	14.91%
MADISON	0.22%	0.27%	0.42%	0.91%	0.96%
PITTSBURGH	0.32%	0.40%	0.67%	1.37%	1.45%



Subset for Training and Test Data we used

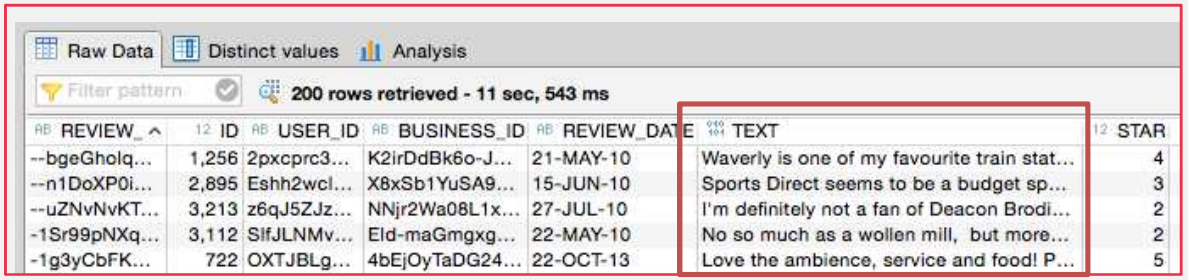
Yelp ERD


check-in
type
business_id
checkin_info [169]
Business
type
business_id
name
neighborhoods
full_address
city
state
latitude
longitude
stars
review_count
categories
open
hours [14]
attributes [80]
review
type
business_id
user_id
stars
text
date
votes [n]
tip
type
text
business_id
user_id
date
likes
user
type
user_id
name
review_count
average_stars
votes [3]
friends
elite
yelping_since
compliments [11]
fans

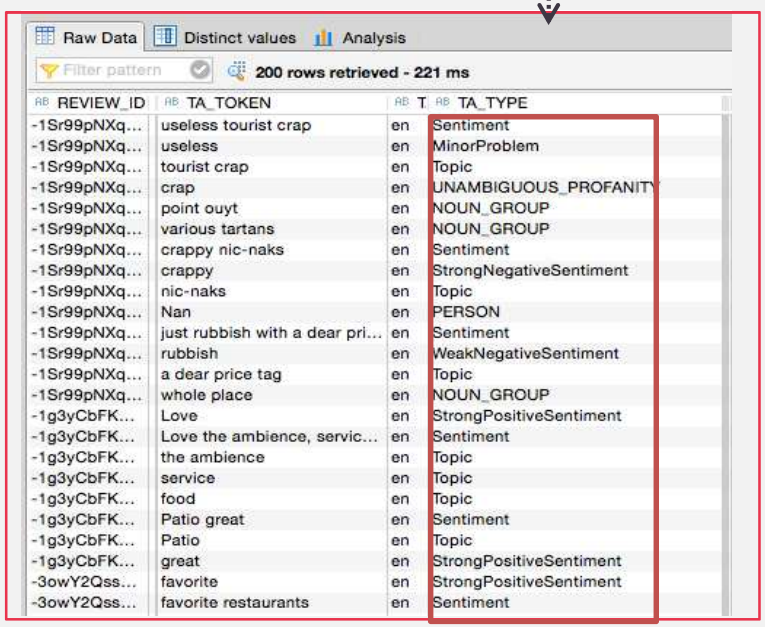
Yelp Review Table

Review _ id	Id	User_ id	Business _id	Review data	Text	Stars
ae3udsjd	1	ejfjdsfs	widsjfqak	21-MAY-11	Awesome Place to Eat	5
djeualdda	2	fswowew	sjfajqwhaa	1-JAN-10	Good Service	4
eienfjsjs	3	ssdsdsfa	qjqdooma	21-JUN-11	I like it	3
eeeerwr4	4	sdqdqdq	ifnafuqdqd	8-FEB-14	Not good	2
ddfoejd	5	seqeqr	dadkanda	17-AUG-13	Worst place ever	1

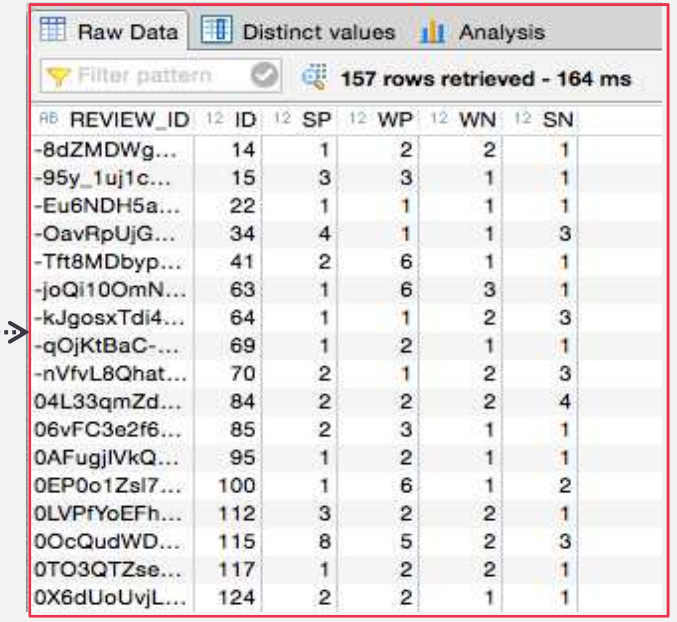
Sentiment Analysis



Original Data



Sentiment Analysis



Transposed And Counted

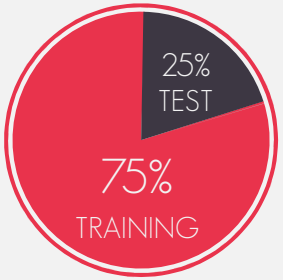
Training & Test Dataset

Raw Data | Distinct values | Analysis

Filter pattern | 157 rows retrieved - 464 ms

REVIEW_ID	ID	SP	WP	WN	SN	STARS
-8dZMDWg...	14	1	2	2	1	3
-95y_1uj1c...	15	3	3	1	1	4
-Eu6NDH5a...	22	1	1	1	1	1
-OavRpUjG...	34	4	1	1	3	1
-Tft8MDbyp...	41	2	6	1	1	2
-joQi10OmN...	63	1	6	3	1	1
-kJgosxTdi4...	64	1	1	2	3	1
-qOjKtBaC-...	69	1	2	1	1	1
-nVfvL8Qhat...	70	2	1	2	3	1
04L33qmZd...	84	2	2	2	4	2
06vFC3e2f6...	85	2	3	1	1	3
0AFugjlVkQ...	95	1	2	1	1	3
0EP0o1Zsl7...	100	1	6	1	2	3
0LVPfYoEFh...	112	3	2	2	1	2
0OcQudWD...	115	8	5	2	3	4

Training Data

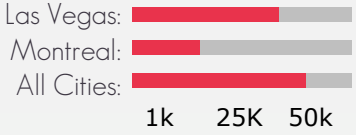


Raw Data | Distinct values | Analysis

Filter pattern | 158 rows retrieved - 295 ms

REVIEW_ID	ID	SP	WP	WN	SN
9h5Teg0uH...	4	1	2	1	1
9krwyqu9C...	12	4	5	1	1
9sQACA8O...	20	2	1	4	2
9tOsnshKxp...	24	3	1	1	1
A-jvMI_ltaE...	30	2	4	1	2
A4NXPRrltfs...	35	1	2	2	3
ABIMbqWNj...	48	13	13	1	2
AStVogtaH...	59	3	5	1	2
Amr2qoQ9y...	77	4	8	1	1
B-5wgF7km...	92	4	2	1	1
BQKv3f3gU...	126	1	1	2	2
BQcc2ZXn2...	128	1	1	2	2
BS5c9CNQ...	130	2	1	1	3
BYTA-A7DD...	134	1	3	1	1
Bpp8unXZm...	148	1	2	1	1
BqMZbVgF-...	150	1	4	2	3
9qYgkbSljm...	16	0	1	1	1

Test Data



Subset for Training and Test Data we used

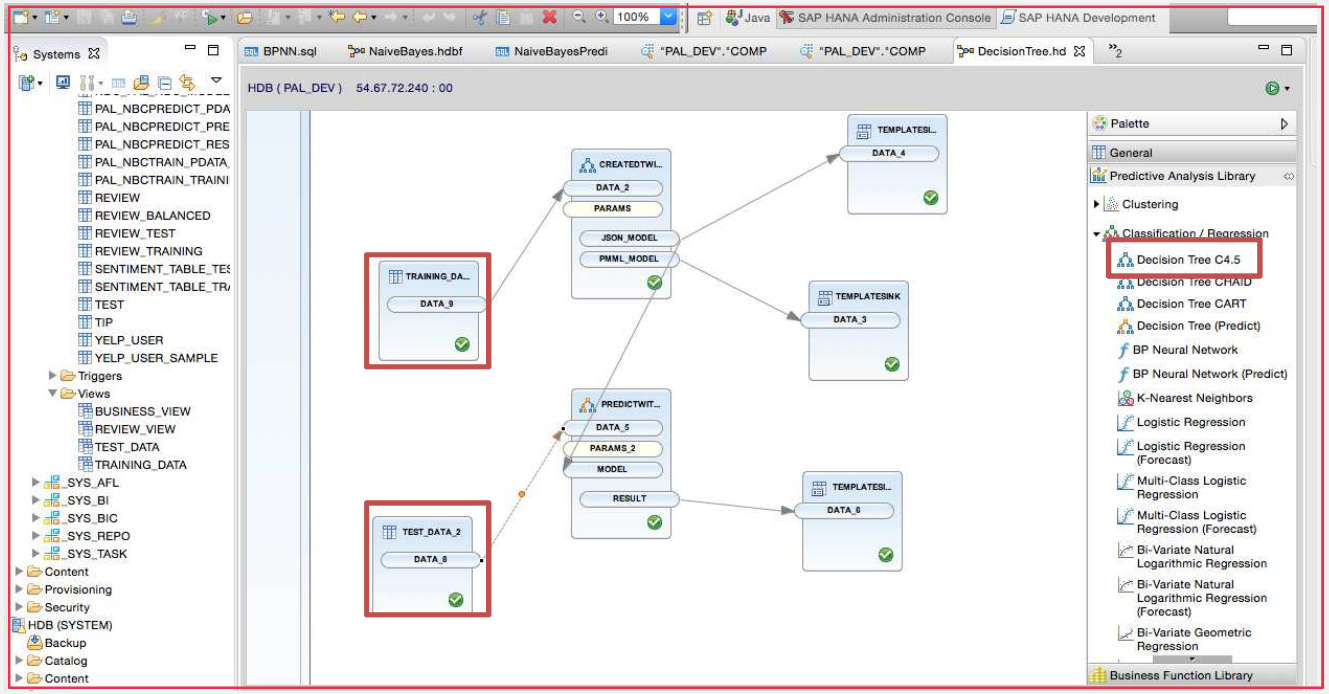
Algorithms

NAÏVE BAYES

NEURAL NETWORK

DECISION TREES

Application Function Modeler (AFM)

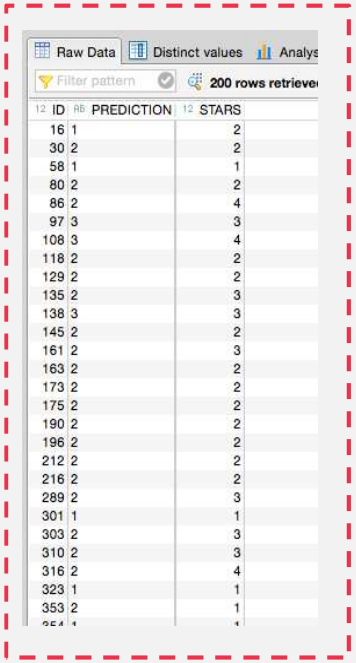


Separate AFM models for each Algorithm.

Result Analysis

Confusion Matrix , Roc Curve & Percent Off Diagonal

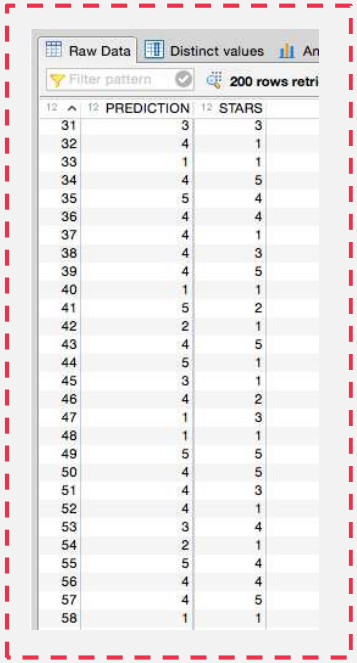
Results


Raw Data | Distinct values | Analys

Filter pattern | 200 rows retrieve

ID	PREDICTION	STARS
16	1	2
30	2	2
58	1	1
80	2	2
86	2	4
97	3	3
108	3	4
118	2	2
129	2	2
135	2	3
138	3	3
145	2	2
161	2	3
163	2	2
173	2	2
175	2	2
190	2	2
196	2	2
212	2	2
216	2	2
289	2	3
301	1	1
303	2	3
310	2	3
316	2	4
323	1	1
353	2	1

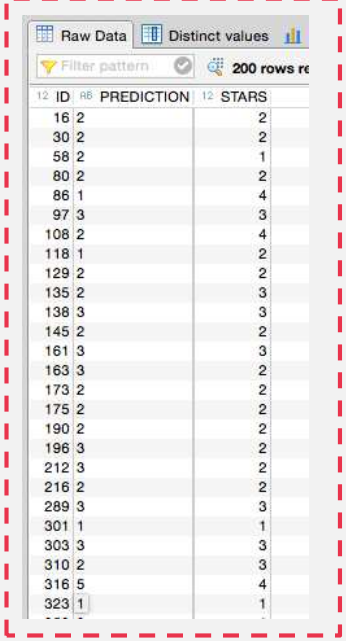
VS


Raw Data | Distinct values | An

Filter pattern | 200 rows retri

	PREDICTION	STARS
31	3	3
32	4	1
33	1	1
34	4	5
35	5	4
36	4	4
37	4	1
38	4	3
39	4	5
40	1	1
41	5	2
42	2	1
43	4	5
44	5	1
45	3	1
46	4	2
47	1	3
48	1	1
49	5	5
50	4	5
51	4	3
52	4	1
53	3	4
54	2	1
55	5	4
56	4	4
57	4	5
58	1	1

VS


Raw Data | Distinct values

Filter pattern | 200 rows re

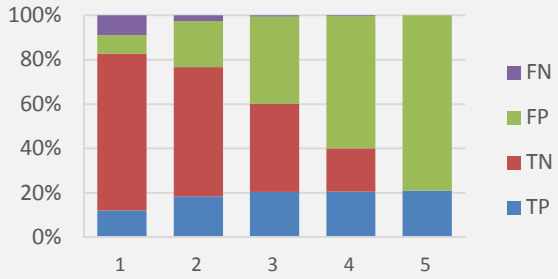
ID	PREDICTION	STARS
16	2	2
30	2	2
58	2	1
80	2	2
86	1	4
97	3	3
108	2	4
118	1	2
129	2	2
135	2	3
138	3	3
145	2	2
161	3	3
163	3	2
173	2	2
175	2	2
190	2	2
196	3	2
212	3	2
216	2	2
289	3	3
301	1	1
303	3	3
310	2	3
316	5	4
323	1	1

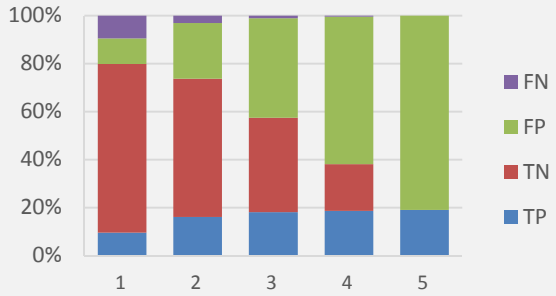
Naïve Bayes

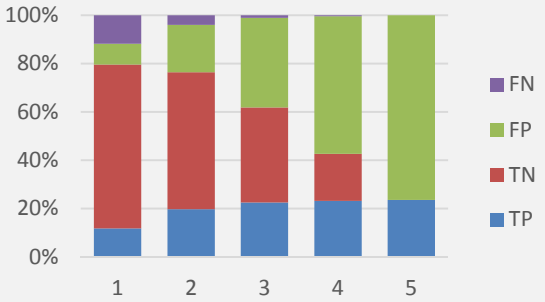
Neural Network

Decision Tree

Results All Cities







Naïve Bayes

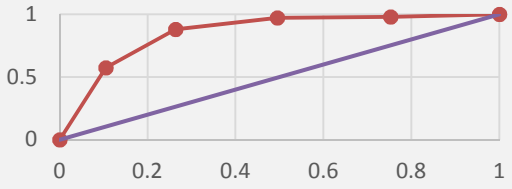
VS

Neural Network

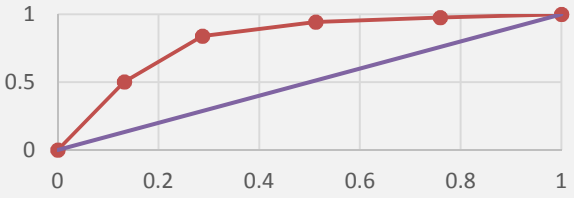
VS

Decision Tree

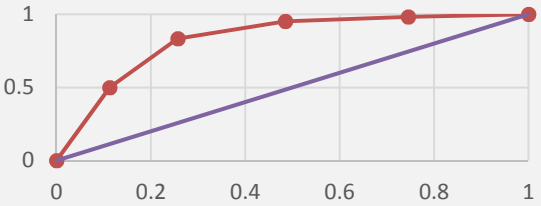
AUC=0.856



AUC=0.813

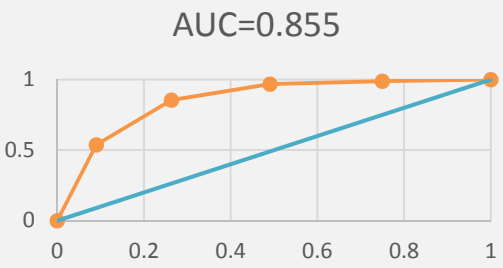


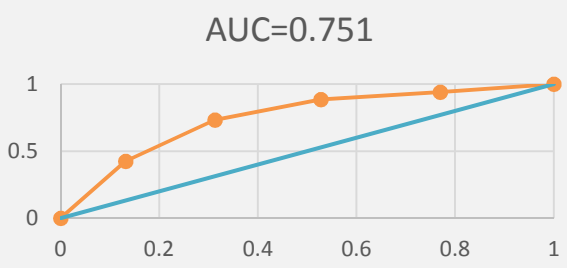
AUC=0.833

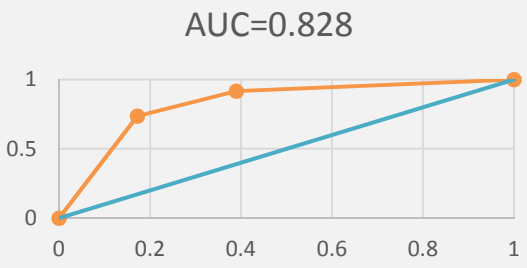



LAS VEGAS
MONTREAL
ALL CITIES
100%
80%
60%
40%
20%
0%
1
2
3
4
5
FN
FP
TN
TP
Confusion Matrix
Naïve Bayes
VS
Neural Network
VS
Decision Tree

LAS VEGAS

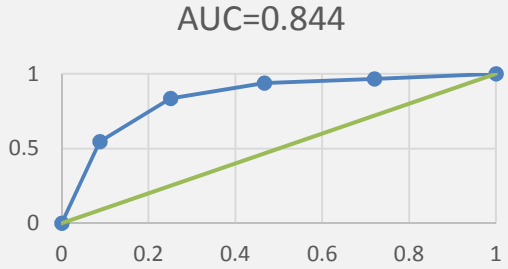


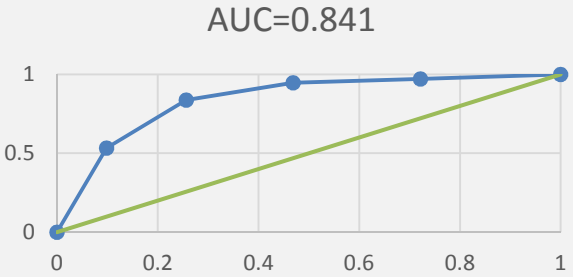


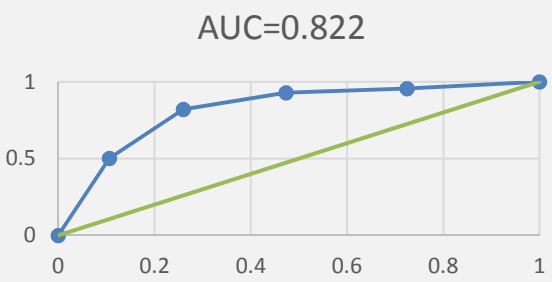


ROC Curve

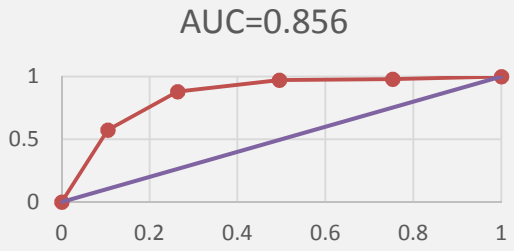
MONTREAL

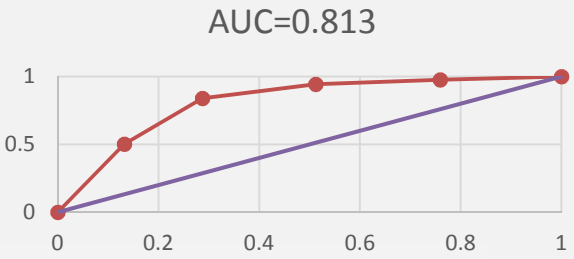


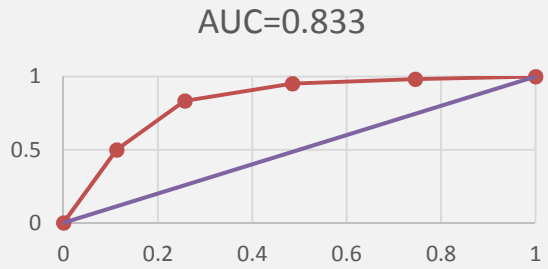




ALL CITIES







Naïve Bayes

VS

Neural Network

VS

Decision Tree

Average AUC

All Naïve Bayes

All Neural Network

All Decision Tree

All Cities and All Samples Tested.

Percent Off Diagonal

UNBUCKETED

Predicted \ Actual	1	2	3	4	5
1	12.04%	6.41%	1.94%	0.18%	0.42%
2	5.12%	6.32%	5.03%	2.08%	0.92%
3	1.29%	3.14%	5.95%	3.09%	0.97%
4	0.51%	1.66%	3.55%	6.09%	4.57%
5	1.38%	1.43%	3.74%	9.13%	13.05%

Pct Off Diagonal: 56.55%

On Diagonal	Total	Off Diagonal
43.45%	100.00%	56.55%

Predicted \ Actual	1	2	3	4	5
1	9.64%	6.46%	1.98%	0.60%	0.46%
2	2.91%	4.06%	3.55%	2.35%	0.88%
3	5.30%	4.43%	8.21%	4.57%	3.04%
4	1.01%	1.71%	2.68%	3.41%	3.74%
5	1.48%	2.31%	3.78%	9.64%	11.81%

Pct Off Diagonal: 62.87%

On Diagonal	Total	Off Diagonal
37.13%	100.00%	62.87%

Predicted \ Actual	1	2	3	4	5
1	11.76%	7.89%	2.81%	0.69%	0.42%
2	4.20%	4.01%	3.83%	2.49%	1.06%
3	2.44%	4.43%	6.00%	3.32%	1.38%
4	0.60%	1.06%	3.78%	4.06%	4.15%
5	1.34%	1.57%	3.78%	10.01%	12.92%

Pct Off Diagonal: 61.25%

On Diagonal	Total	Off Diagonal
38.75%	100.00%	61.25%

BUCKETED

Predicted \ Actual	1 & 2	3	4 & 5
1 & 2	29.89%	6.96%	3.60%
3	4.43%	5.95%	4.06%
4 & 5	4.98%	7.29%	32.84%

Pct Off Diagonal: 31.32%

On Diagonal	Total	Off Diagonal
68.68%	100%	31.32%

All Naïve Bayes

Predicted \ Actual	1 & 2	3	4 & 5
1 & 2	23.06%	5.54%	4.29%
3	9.73%	8.21%	7.61%
4 & 5	6.50%	6.46%	28.60%

Pct Off Diagonal: 40.13%

On Diagonal	Total	Off Diagonal
59.87%	100%	40.13%

All Neural Network

Predicted \ Actual	1 & 2	3	4 & 5
1 & 2	27.86%	6.64%	4.66%
3	6.87%	6.00%	4.70%
4 & 5	4.57%	7.56%	31.13%

Pct Off Diagonal: 35.01%

On Diagonal	Total	Off Diagonal
64.99%	100%	35.01%

All Decision Tree

Challenges

- Pre-processing the data before importing into SAP.
- Handling the Skew of the data
- Tool learning curve (AFM, Python, XLSTAT)
- Focusing on a specific problem.

Conclusion

- Naïve Bayes performed best because it allows each attribute to contribute equally towards the final decision.
- Predication accuracy for different geographic location were consistent because we used the same configuration file /dictionary to perform our text analysis.
- Binning the predictions 1-2, 3 ,4-5 gave us the best result.

Thanks and Questions