Machine Learning Analysis of Hotel Reviews: Insights and Predictions

Aalto-yliopisto

2024

Overview of the Project

- Customer feedback can significantly influence **business outcomes**

- Analyzing reviews is a crucial part of understanding customer **satisfaction and improving services**

- This analysis is based on the dataset of **TripAdvisor** hotel reviews

- Aim of this analysis is to deliver actionable insights through **machine learning**

- Understanding these aspects can help hotels tailor their strategies to **enhance customer experience and manage their online reputation**

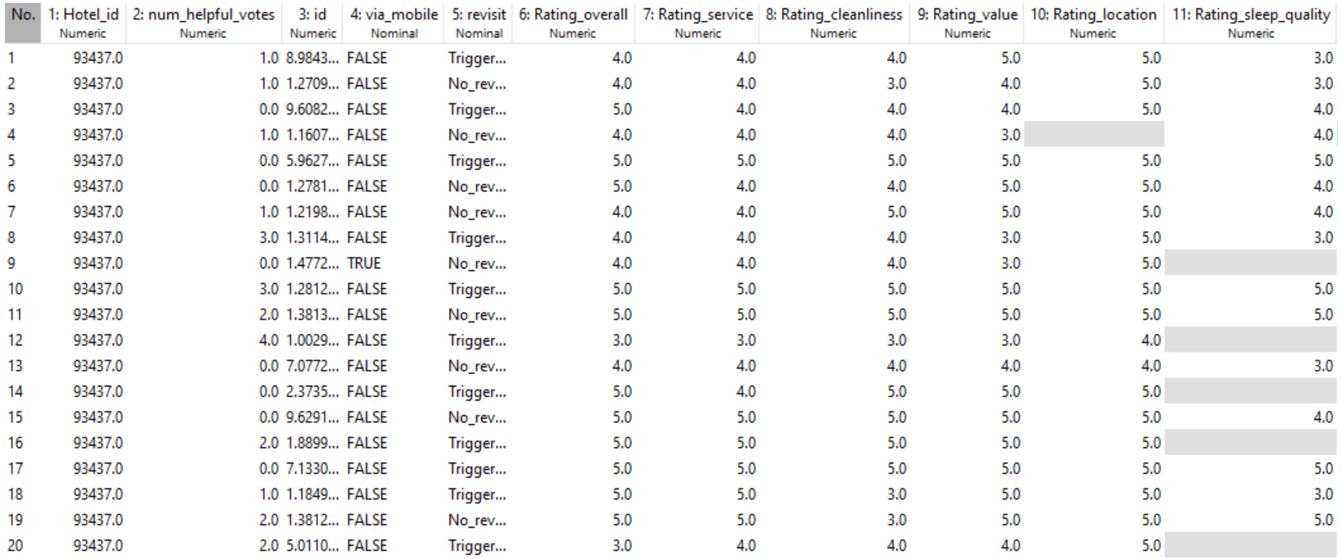
Objective of the Project

- Objective:
  - Develop a model that will quantify the effects of reviewers with one function presenting the effect as a function of the number of cities he visited and another from the helpfulness of his reviews. The scope of this work is to identify those reviewers who had visited more than 15 cities and had their reviews voted as helpful more than 100 times by users. From this analysis, the company can single out the experienced reviewer who significantly influences the market, thus allowing targeting marketing strategies and personalized engagements with such customers

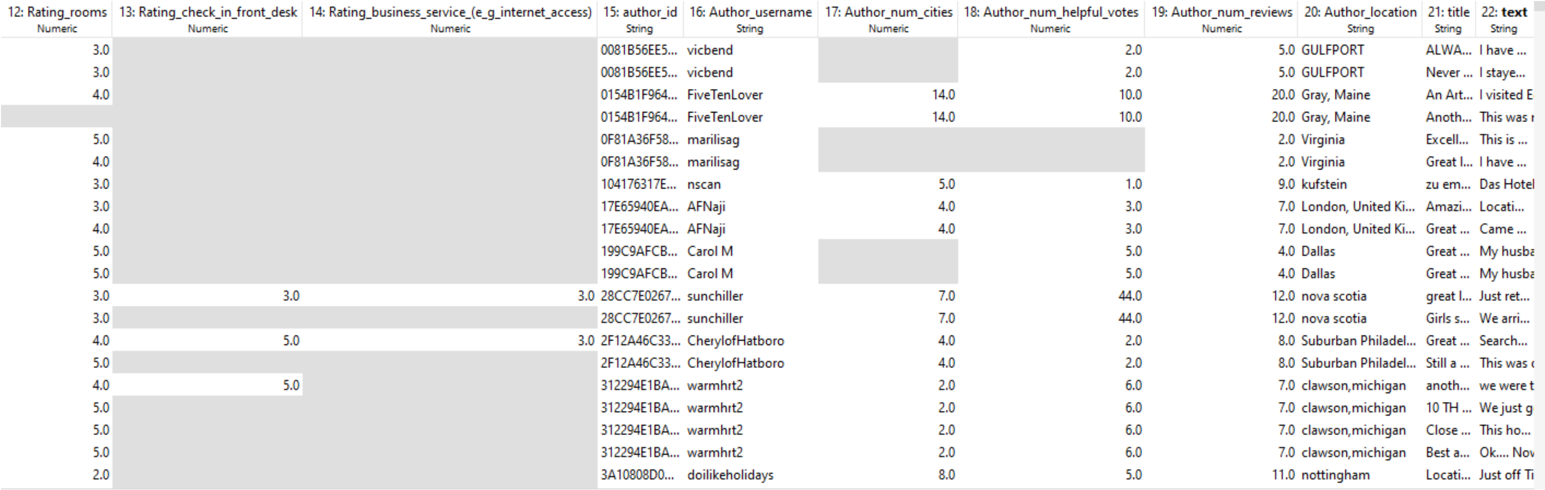
- This analysis aims to provide answers for tasks related to **Business Scenario 2**:
  - 1. Utilize all the variables avalailable how influential the review author is in real life ("Author_num_cities")
  - 2. Utilize all the variables available to influential the review author is in the online environment ("Author_num_helpful_votes")

Overview of the Original Dataset

- Dataset contains:
  - 22 attributes
  - 3118 instances



No.	1: Hotel_id Numeric	2: num_helpful_votes Numeric	3: id Numeric	4: via_mobile Nominal	5: revisit Nominal	6: Rating_overall Numeric	7: Rating_service Numeric	8: Rating_cleanliness Numeric	9: Rating_value Numeric	10: Rating_location Numeric	11: Rating_sleep_quality Numeric
1	93437.0	1.0	8.9843...	FALSE	Trigger...	4.0	4.0	4.0	5.0	5.0	3.0
2	93437.0	1.0	1.2709...	FALSE	No_rev...	4.0	4.0	3.0	4.0	5.0	3.0
3	93437.0	0.0	9.6082...	FALSE	Trigger...	5.0	4.0	4.0	4.0	5.0	4.0
4	93437.0	1.0	1.1607...	FALSE	No_rev...	4.0	4.0	4.0	3.0		4.0
5	93437.0	0.0	5.9627...	FALSE	Trigger...	5.0	5.0	5.0	5.0	5.0	5.0
6	93437.0	0.0	1.2781...	FALSE	No_rev...	5.0	4.0	4.0	5.0	5.0	4.0
7	93437.0	1.0	1.2198...	FALSE	No_rev...	4.0	4.0	5.0	5.0	5.0	4.0
8	93437.0	3.0	1.3114...	FALSE	Trigger...	4.0	4.0	4.0	3.0	5.0	3.0
9	93437.0	0.0	1.4772...	TRUE	No_rev...	4.0	4.0	4.0	3.0	5.0	
10	93437.0	3.0	1.2812...	FALSE	Trigger...	5.0	5.0	5.0	5.0	5.0	5.0
11	93437.0	2.0	1.3813...	FALSE	No_rev...	5.0	5.0	5.0	5.0	5.0	5.0
12	93437.0	4.0	1.0029...	FALSE	Trigger...	3.0	3.0	3.0	3.0	4.0	
13	93437.0	0.0	7.0772...	FALSE	No_rev...	4.0	4.0	4.0	4.0	4.0	3.0
14	93437.0	0.0	2.3735...	FALSE	Trigger...	5.0	4.0	5.0	5.0	5.0	
15	93437.0	0.0	9.6291...	FALSE	No_rev...	5.0	5.0	5.0	5.0	5.0	4.0
16	93437.0	2.0	1.8899...	FALSE	Trigger...	5.0	5.0	5.0	5.0	5.0	
17	93437.0	0.0	7.1330...	FALSE	Trigger...	5.0	5.0	5.0	5.0	5.0	5.0
18	93437.0	1.0	1.1849...	FALSE	Trigger...	5.0	5.0	3.0	5.0	5.0	3.0
19	93437.0	2.0	1.3812...	FALSE	No_rev...	5.0	5.0	3.0	5.0	5.0	5.0
20	93437.0	2.0	5.0110...	FALSE	Trigger...	3.0	4.0	4.0	4.0	5.0	

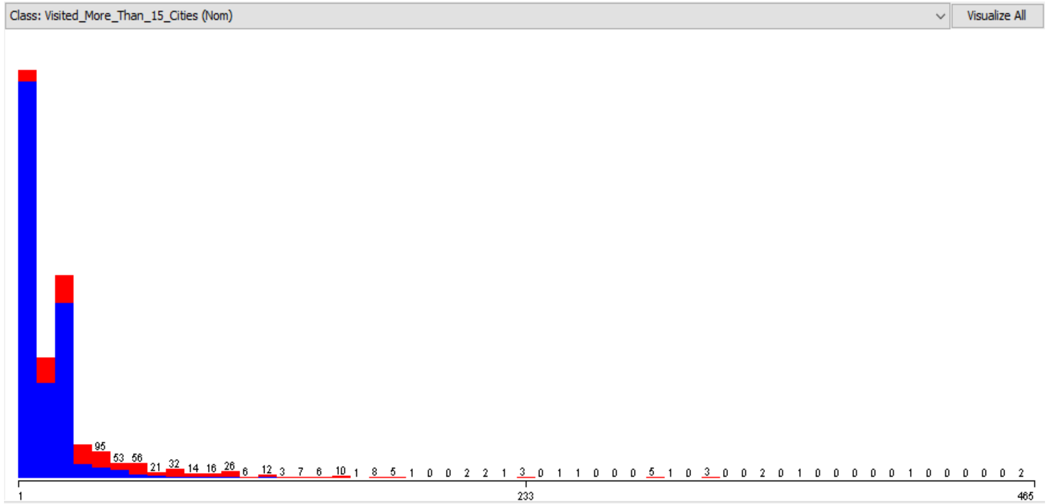


12: Rating_rooms Numeric	13: Rating_check_in_front_desk Numeric	14: Rating_business_service_(e_g_internet_access) Numeric	15: author_id String	16: Author_username String	17: Author_num_cities Numeric	18: Author_num_helpful_votes Numeric	19: Author_num_reviews Numeric	20: Author_location String	21: title String	22: **text** String
3.0			0081B56EE5...	vicbend		2.0	5.0	GULFPORT	ALWA...	I have ...
3.0			0081B56EE5...	vicbend		2.0	5.0	GULFPORT	Never ...	I staye...
4.0			0154B1F964...	FiveTenLover	14.0	10.0	20.0	Gray, Maine	An Art...	I visited E
			0154B1F964...	FiveTenLover	14.0	10.0	20.0	Gray, Maine	Anoth...	This was r
5.0			0F81A36F58...	marilisag			2.0	Virginia	Excell...	This is ...
4.0			0F81A36F58...	marilisag			2.0	Virginia	Great l...	I have ...
3.0			104176317E...	nscan	5.0	1.0	9.0	kufstein	zu em...	Das Hotel
3.0			17E65940EA...	AFNaji	4.0	3.0	7.0	London, United Ki...	Amazi...	Locati...
4.0			17E65940EA...	AFNaji	4.0	3.0	7.0	London, United Ki...	Great ...	Came ...
5.0			199C9AFCB...	Carol M		5.0	4.0	Dallas	Great ...	My husba
5.0			199C9AFCB...	Carol M		5.0	4.0	Dallas	Great ...	My husba
3.0	3.0	3.0	28CC7E0267...	sunchiller	7.0	44.0	12.0	nova scotia	great l...	Just ret...
3.0			28CC7E0267...	sunchiller	7.0	44.0	12.0	nova scotia	Girls s...	We arri...
4.0	5.0	3.0	2F12A46C33...	CherylofHatboro	4.0	2.0	8.0	Suburban Philadel...	Great ...	Search...
5.0			2F12A46C33...	CherylofHatboro	4.0	2.0	8.0	Suburban Philadel...	Still a ...	This was
4.0	5.0		312294E1BA...	warmhrt2	2.0	6.0	7.0	clawson,michigan	anoth...	we were t
5.0			312294E1BA...	warmhrt2	2.0	6.0	7.0	clawson,michigan	10 TH ...	We just g
5.0			312294E1BA...	warmhrt2	2.0	6.0	7.0	clawson,michigan	Close ...	This ho...
5.0			312294E1BA...	warmhrt2	2.0	6.0	7.0	clawson,michigan	Best a...	Ok.... Nov
2.0			3A10808D0...	doilikeholidays	8.0	5.0	11.0	nottingham	Locati...	Just off Ti

- Attributes contain:
  - 2 nominal attributes
  - 3 strings attributes
  - 18 numerical attributes

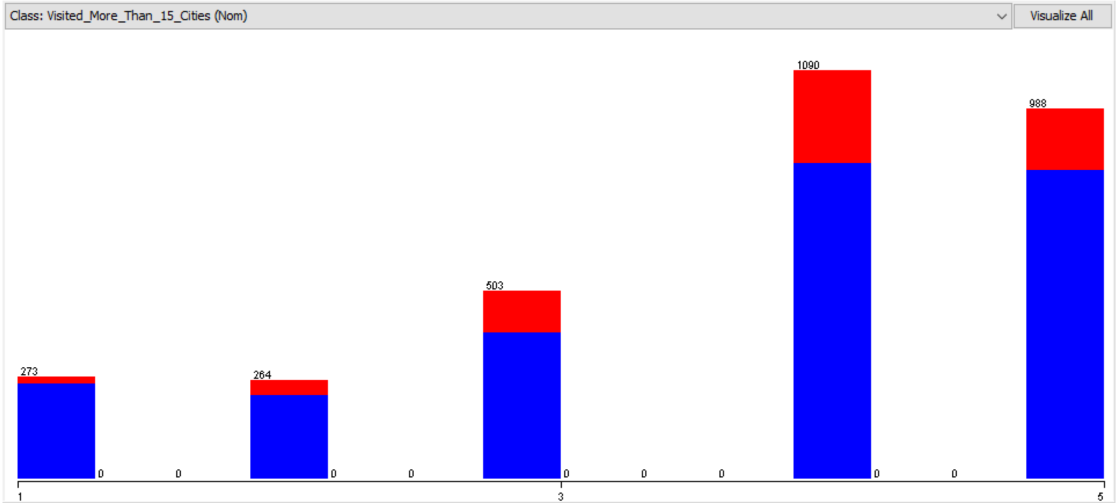
Overview of the Key Attributes

Author_num_helpful_votes



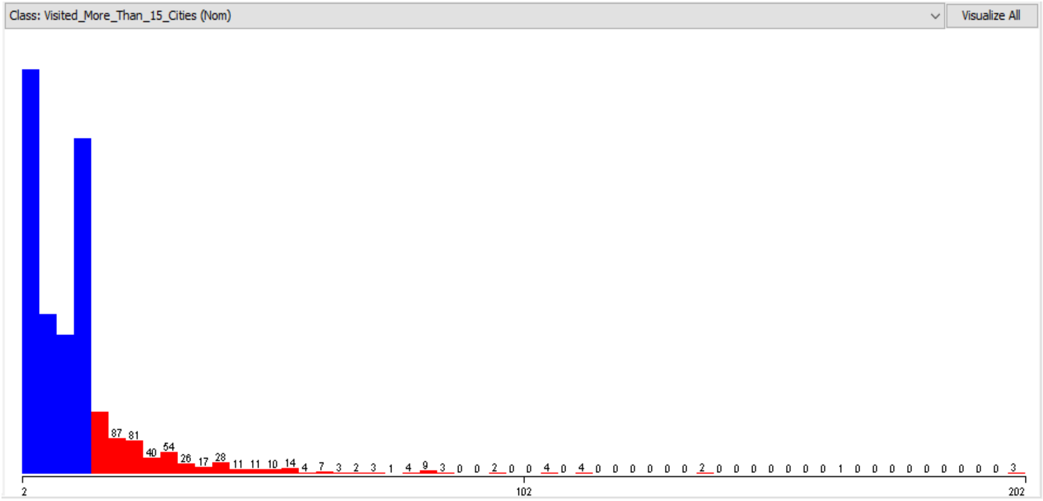
Class: Visited_More_Than_15_Cities (Nom)

Overall_rating



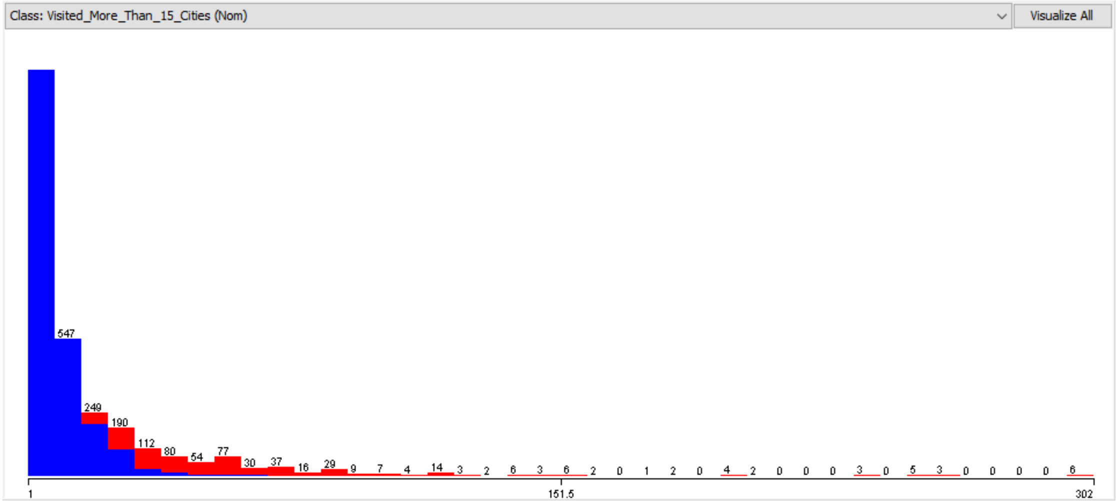
Class: Visited_More_Than_15_Cities (Nom)

Author_num_cities



Class: Visited_More_Than_15_Cities (Nom)

Author_num_reviews



Class: Visited_More_Than_15_Cities (Nom)

Data Preprocessing: Dealing with missing values

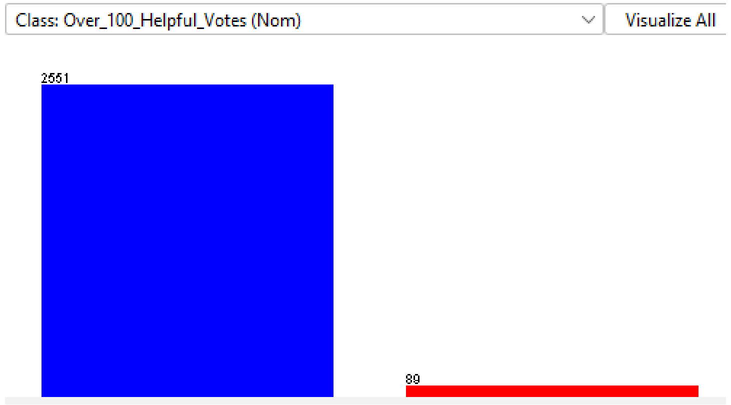
- There was a **large** amount of missing variables:
  - 9% from rating_service
  - 8% from rating_cleanliness
  - 9% from rating_value
  - 17% from rating_location
  - 39% from rating_sleep_quality
  - 14% from rating_rooms
  - 87% from rating_check_in_front_desk
  - 93% from rating_business_service_(e_g_internet_access)
  - 21% from author_num_cities
  - 15% from author_num_helpful_votes
  - 6% from author_location
- Because vast majority of variables were missing from attributes "**rating_business_service**" and "**rating_check_in_front_desk**" they were completely deleted
- Numerical missing variables were replaced with median variable. Filter "**ReplaceMissingValues**" was used
- All the string variables were removed.

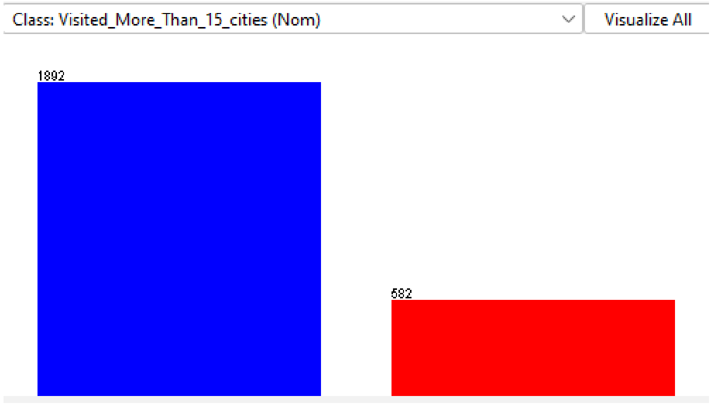
Feature Engineering and Class Attribute Selection

- New attributes were created to help with the analysis
  - Nominal attributes were created to reflect whether an author has visited more than 15 cities or received more than 100 helpful votes. "**AddExpression**" filter was used to create two new attributes:
    - "**Visited_More_Than_15_cities**" to indicate whether the author had visited over 15 cities or not
    - "**Over_100_Helpful_Votes**" to indicate whether the author had received over 100 helpful votes or not
    - These steps were done separataly to the same dataset

- Only 582 authors had visited over 15 cities
- Only 89 authors had received over 100 helpful votes

- These new attributes were set as class separately to proceed with the analysis





Attribute Evaluation

- Attributes "**Visited_More_Than_15_cities**" and "**Over_100_Helpful_Votes**" were made to be nominal variables to identify influencial factors through "**InfoGainAttributeEval**" attribute evaluator where search method was "**Ranked**"

- Attributes were nominally categorized as "**1**" and "**0**"
  - For "**Visited_More_Than_15_cities**" instances with over 15 as a value were consider "**1**"
  - For "**Over_100_Helpful_Votes**" instances with over 100 as a value were considered "**1**"

Attribute Evaluation (cont'd)

- For ”Visited_More_Than_15_cities” the most important predictors were: ”author_num_reviews”, ”author_num_helpful_votes”
  - ”author_num_reviews” suggests that authors who have written more reviews are more likely to have visited more cities
  - ”author_num_helpful_votes” indicates that authors whose reviews are deemed helpful also tend to travel more
- For ”Over_100_helpful_votes” the most important predictors were also: ”author_num_cities”, ”author_num_reviews”
  - ”author_num_cities” suggests that the authors who travel more are more likely to provide reviews that are considered helpful
  - ”author_num_reviews” indicates that the number of reviews author has written is related to the amount they are found helpful

Model Selection Criteria

- Ultimately, the dataset was **unbalanced** so **SMOTE** was used to balance the dataset

- **Decision tree** and **Logistic Regression** was used
  - **Logistic Regression** is espesially well suited for nominal attributes

- Ensemble method like **Random Forest** could provide better performance compared to single decision tree by reducing overfitting

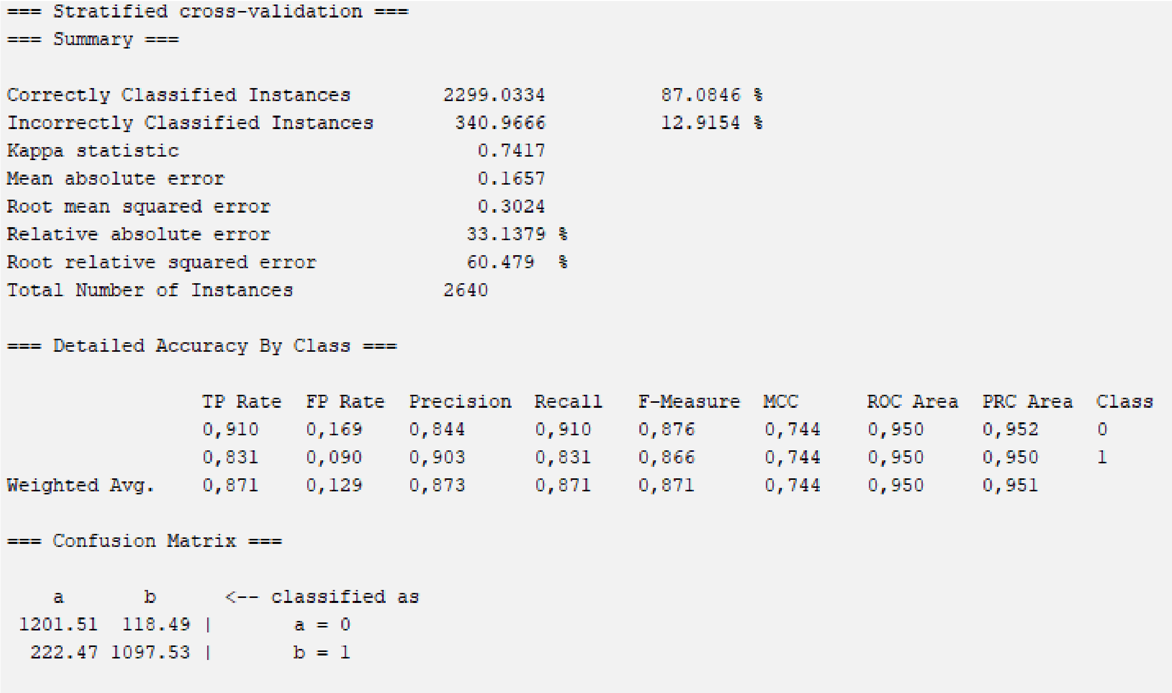
- **Cost Matrix** was utilized to gain more information about relative importance or business impact of each type of classification error

- 10-fold cross-validation was used for all of the models

Logistic Regression

- Logistic regression was used for the nominal variable ”**over_100_helpful_votes**” to predict whether a review author has received over 100 helpful votes

- Model had overall accuracy of 87.0846%
- Classes are performing pretty similarly to eachother
- Precision and Recall are the most significantly differing
- This variable was notably more unbalanced, which leads to lower accuracy

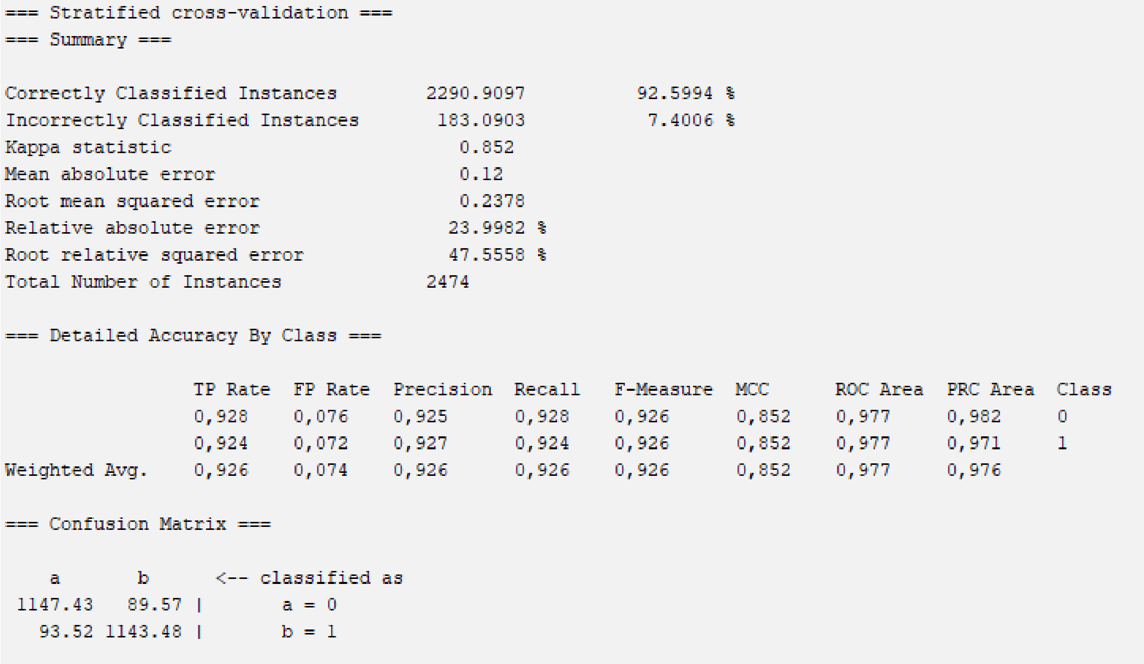


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2299.0334          87.0846 %
Incorrectly Classified Instances       340.9666          12.9154 %
Kappa statistic                          0.7417
Mean absolute error                      0.1657
Root mean squared error                  0.3024
Relative absolute error                 33.1379 %
Root relative squared error             60.479  %
Total Number of Instances             2640

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,910    0,169    0,844      0,910    0,876      0,744    0,950     0,952     0
                 0,831    0,090    0,903      0,831    0,866      0,744    0,950     0,950     1
Weighted Avg.    0,871    0,129    0,873      0,871    0,871      0,744    0,950     0,951

=== Confusion Matrix ===

    a       b      <-- classified as
 1201.51  118.49 |       a = 0
  222.47 1097.53 |       b = 1
```

Logistic Regression (cont’d)

- Logistic regression was used for the nominal variable ”**Visited_More_Than_15_cities**” to predict whether a review author has visited over 15 cities

- Model had overall accuracy of 92.5994%
- Classes are pretty well performing compared to eachother
- Compared to logistic precision of previous variable, the accuracy is notably better



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2290.9097          92.5994 %
Incorrectly Classified Instances       183.0903           7.4006 %
Kappa statistic                          0.852
Mean absolute error                      0.12
Root mean squared error                  0.2378
Relative absolute error                 23.9982 %
Root relative squared error             47.5558 %
Total Number of Instances             2474

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,928    0,076    0,925      0,928    0,926      0,852    0,977     0,982     0
                 0,924    0,072    0,927      0,924    0,926      0,852    0,977     0,971     1
Weighted Avg.    0,926    0,074    0,926      0,926    0,926      0,852    0,977     0,976

=== Confusion Matrix ===

    a       b     <-- classified as
 1147.43   89.57 |      a = 0
   93.52 1143.48 |      b = 1
```

Key Findings from the Logistic Regression Models

- ”Visited_More_Than_15_cities” :The logistic regression model shows a decent capability to classify instances based on whether an author has visited more than 15 cities, but it is more reliable in predicting those who have not visited more than 15 cities than those who have. The relatively high false positive rate for predicting visits to more than 15 cities indicates a need for model refinement or consideration of additional features that could improve the prediction accuracy for this class.

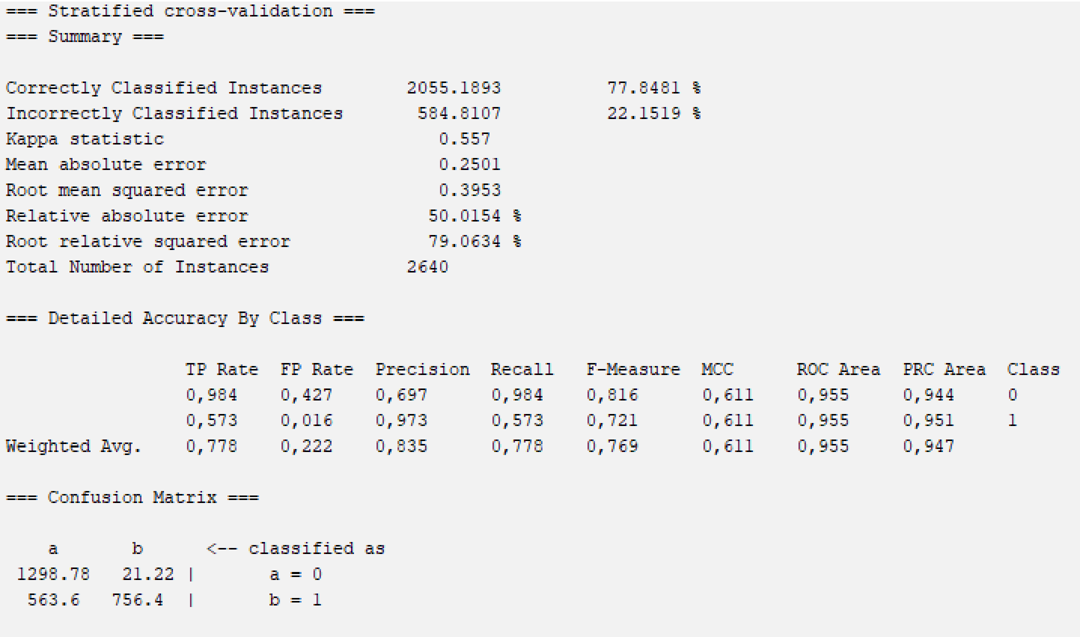
- ”Over_100_Helpful_Votes” : Although logistic regression in general is highly accurate, it can be problematic in capturing the positives for the less frequent class (authors with over 100 helpful votes). This would mean that the model is generally robust but can further improve with efforts of feature engineering, rebalance techniques, or even alternative model exploration to raise its predictive performance towards identification of the influential review author by the number of his helpful votes.

Random Forest

Aalto-yliopisto

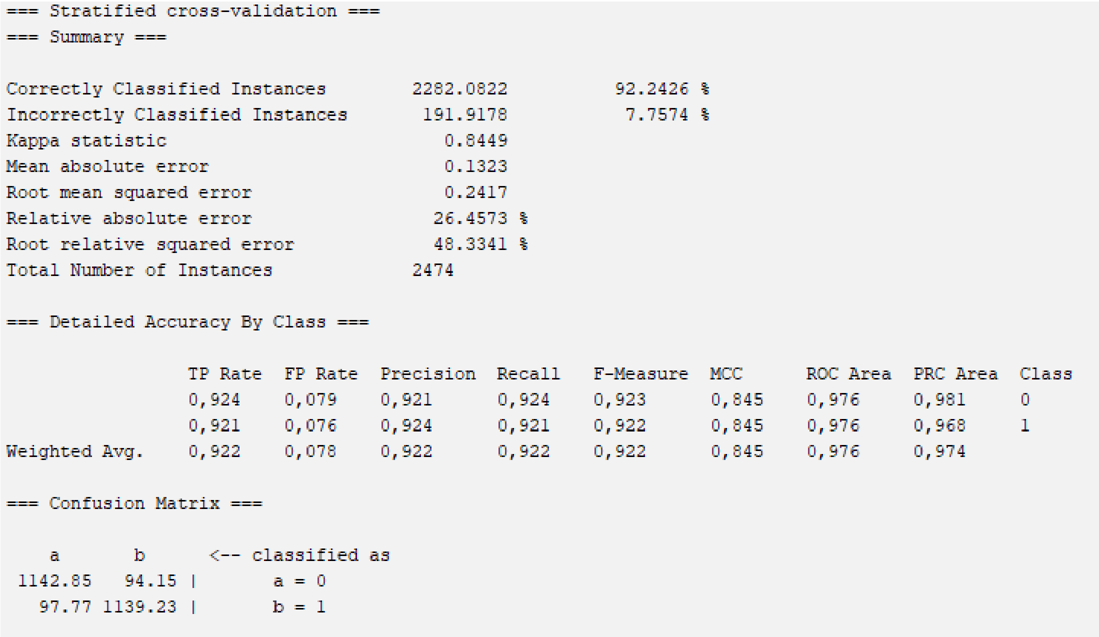
- Random Forest model was worse for "**over_100_helpful_votes**" since this variable was more unbalanced
- This model was performing relatively well for "**visited_more_than_15_cities**"

over_100_helpful_votes



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2055.1893          77.8481 %
Incorrectly Classified Instances       584.8107          22.1519 %
Kappa statistic                          0.557
Mean absolute error                      0.2501
Root mean squared error                  0.3953
Relative absolute error                 50.0154 %
Root relative squared error             79.0634 %
Total Number of Instances             2640

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,984    0,427    0,697      0,984    0,816      0,611    0,955     0,944     0
                 0,573    0,016    0,973      0,573    0,721      0,611    0,955     0,951     1
Weighted Avg.    0,778    0,222    0,835      0,778    0,769      0,611    0,955     0,947

=== Confusion Matrix ===

    a       b     <-- classified as
 1298.78   21.22 |       a = 0
  563.6   756.4  |       b = 1
```

visited_more_than_15_cities



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2282.0822          92.2426 %
Incorrectly Classified Instances       191.9178           7.7574 %
Kappa statistic                          0.8449
Mean absolute error                      0.1323
Root mean squared error                  0.2417
Relative absolute error                 26.4573 %
Root relative squared error             48.3341 %
Total Number of Instances             2474

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,924    0,079    0,921      0,924    0,923      0,845    0,976     0,981     0
                 0,921    0,076    0,924      0,921    0,922      0,845    0,976     0,968     1
Weighted Avg.    0,922    0,078    0,922      0,922    0,922      0,845    0,976     0,974

=== Confusion Matrix ===

    a       b     <-- classified as
 1142.85   94.15 |       a = 0
   97.77 1139.23 |       b = 1
```

Cost Matrix

- Cost Matrix was implemented to achieve more information about business impact of misclassification

- This was the example scenario for Cost Matrix :
    - Sending marketing materials to a non-influential reviewer costs the company €10 (FP) but missing an influential reviewer might mean a lost opportunity cost of €100 (FN) because they could have influenced more bookings or improved brand visibility.
    - Cost matrix was structured as follows:
        - Cost of false positive: 1 (because it costs €10, this is the baseline)
        - Cost of false negative: 10 (reflecting that it's ten times worse to miss an influential reviewer)

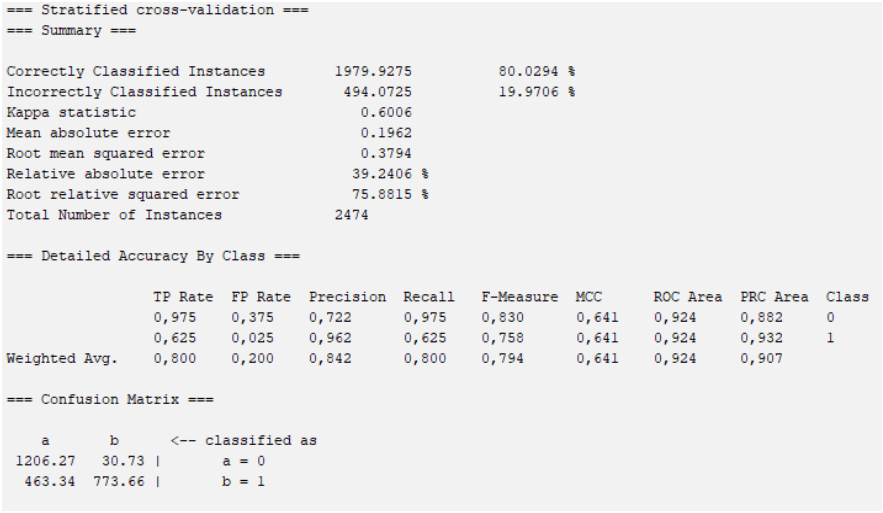
- Decision Tree was used as the Model
    - The Model was used for both new binary variables ”over_100_helpful_votes” and ”Visited_More_Than_15_cities” to gain maximum amount of information

Cost Matrix (cont’d)

- The first model **(visited_more_than_15_cities)** has overall the worst accuracy but balanced both sensitivities in classifying both classes. The statistics for Kappa for this model show moderate agreement of predicted and actual classes, showing steady power of prediction. In addition, the moderate balance of F-Measure details how the model effectively juggles precision and recall, which is critical from the viewpoint that in such a real-life setting, misclassification might be costly.

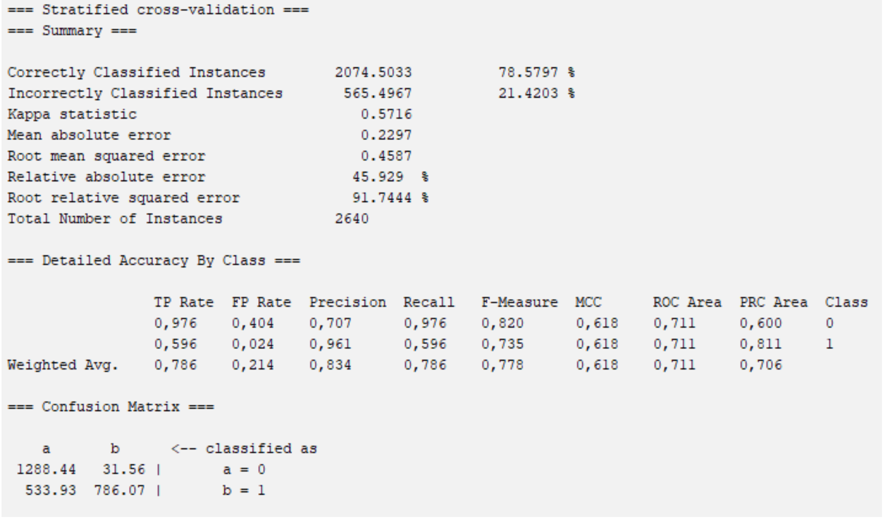
- The second model **(over_100_helpful_votes)** shows a worse accuracy. The lower F-Measure for this class, with the other average model performances, indicates that the model is generally reliable but may be improved for better recall, probably by employing some advanced technique like Cost Sensitive Learning or Class-Weight Adjustment. Such improvements would make the model even more sensitive to the minority class and thereby bring down the risk even further of missing influential reviewers.

visited_more_than_15_cities



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       1979.9275         80.0294 %
Incorrectly Classified Instances      494.0725         19.9706 %
Kappa statistic                          0.6006
Mean absolute error                      0.1962
Root mean squared error                  0.3794
Relative absolute error                 39.2406 %
Root relative squared error             75.8815 %
Total Number of Instances             2474

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,975    0,375    0,722      0,975    0,830      0,641    0,924     0,882     0
                 0,625    0,025    0,962      0,625    0,758      0,641    0,924     0,932     1
Weighted Avg.    0,800    0,200    0,842      0,800    0,794      0,641    0,924     0,907

=== Confusion Matrix ===

    a       b     <-- classified as
 1206.27   30.73 |      a = 0
  463.34  773.66 |      b = 1
```

over_100_helpful_votes



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       2074.5033         78.5797 %
Incorrectly Classified Instances      565.4967         21.4203 %
Kappa statistic                          0.5716
Mean absolute error                      0.2297
Root mean squared error                  0.4587
Relative absolute error                 45.929  %
Root relative squared error             91.7444 %
Total Number of Instances             2640

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,976    0,404    0,707      0,976    0,820      0,618    0,711     0,600     0
                 0,596    0,024    0,961      0,596    0,735      0,618    0,711     0,811     1
Weighted Avg.    0,786    0,214    0,834      0,786    0,778      0,618    0,711     0,706

=== Confusion Matrix ===

    a       b     <-- classified as
 1288.44   31.56 |      a = 0
  533.93  786.07 |      b = 1
```

Key Insights

- The key feature predictors for influential reviews and customer satisfaction were identified to be **"Author_num_reviews," "Author_num_helpful_votes,"** and **"Author_num_cities.**

- Best model for both new variables were **Logistic Regression**
- Worst model for both new variables were **Cost Matrix with Decision Tree**

- The authors with the highest number of reviews and helpful votes influence the potential customer a lot, and their satisfaction is reflected greatly in business results.

- One of the aspects on which this reviewer influence sign was noted was exhibited by the number of cities visited and its strong statistical relationship, indicative of the expectation of separate expectation of experienced travelers

Business Recommendations

- **Enhanced Reviewer Engagement**: Focus on engaging with reviewers who have visited more than 15 cities and those whose reviews receive significant helpful votes. These reviewers are likely to be key influencers within the community.
- **Targeted Response Strategies**: Prepare focused response strategies to the feedback of the most influential reviewers. Addressing their concerns and feedback on time would, to some extent, open an opportunity that may help in increasing customer satisfaction and influencing prospective customers.
- **Monitoring and Analysis**: Continue to monitor and analyze the reviews with an eye on key themes related to good scores and intentions of revisits. Use this feedback for operational excellence, e.g., develop service quality and room cleanliness.
- **Continuous Improvement**: Leverage insights from experienced travelers to refine service offerings.