Ain Shams University

Faculty of Computer and information science

Machine learning

Team ID: SC_7

Project name: Hotel Rating Prediction

Project Description:

Predicting the reviewer score on a hotel depending on some features.

Name	Section	ID
احمد اسامة عبدالرزاق حسن (team leader)	1	20201700020
الاء مصطفى صادق على	2	20201700137
شهد اشرف احمد عبداللاه	4	20201700407
شهد محمد كامل مصطفي	extra	20201700409
احمد اديب معضماني	1	20201701044
مصعب ثابت محمد عبد الآخر	7	20201700848

1- The Preprocessing:

We worked on every column.

- **(lat & lng -> address)**
  After trying to find a pattern in Hotel Address column we decided to take the latitude and longitude and connect the model to google maps to return an object that has the city, country to put them in columns to the next stage.

- **(tags)**
  Splitting on the “,” and dropping the extra punctuation marks then putting each of the room type, trip type, nights spent and the people in a new column to later then encode.

- **(Trip Type)**

  After splitting the "tags" column into different categories, we found that the "trip type" category had only two values: "Leisure trip" and "Business trip". Some rows were empty, so we performed logistic regression to classify the values based on the provided tags into the aforementioned two classes.

- **(Review_Date)**
  Turning it to date time then putting day, month and year then putting each into a column to encode.

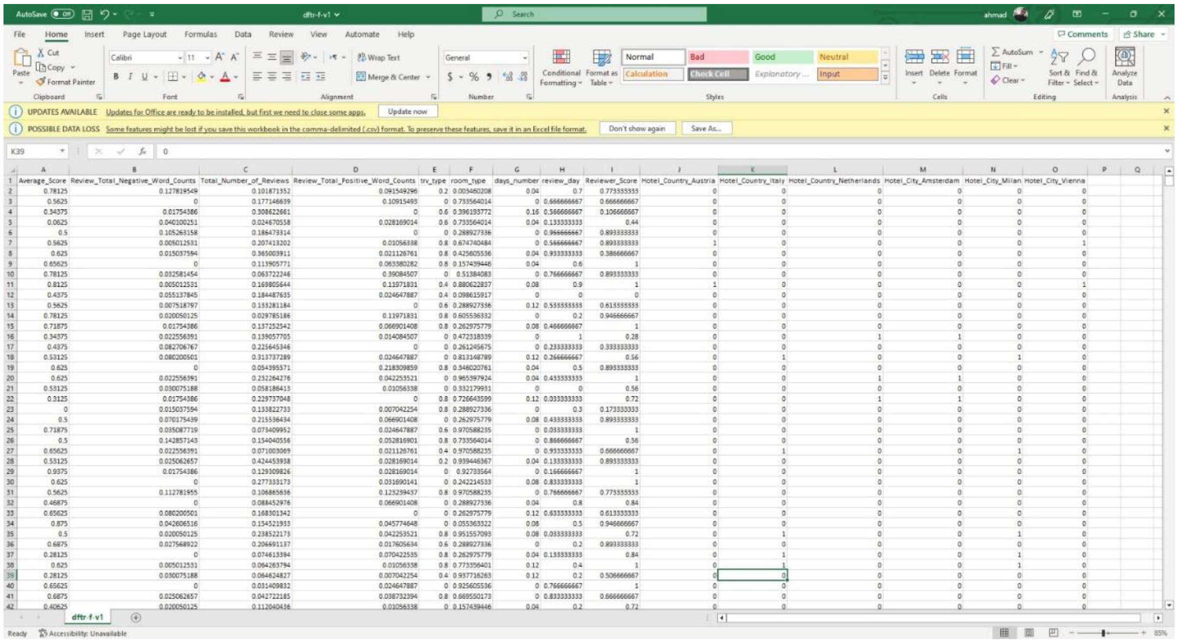
After that dropping the nulls, we moved to the next part.

Encoding:

Using Label and one hot Encoding to turn all our data to numerical to use them properly.

Scaling:

Re scaling the data from 0 to 1 so that all columns become on one scale.



UPDATES AVAILABLE Updates for Office are ready to be installed, but first we need to close some apps.

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

K39 | 0

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Average_Score	Review_Total_Negative_Word_Counts	Total_Number_of_Reviews	Review_Total_Positive_Word_Counts	trv_type	room_type	days_number	review_day	Reviewer_Score	Hotel_Country_Austria	Hotel_Country_Italy	Hotel_Country_Netherlands	Hotel_City_Amsterdam	Hotel_City_Milan	Hotel_City_Vienna
2	0.78125	0.127819549	0.101871352	0.091549296	0.2	0.003460208	0.04	0.7	0.773333333	0	0	0	0	0	0
3	0.5625	0	0.177146639	0.10915493	0	0.733564014	0	0.666666667	0.666666667	0	0	0	0	0	0
4	0.34375	0.01754386	0.308622661	0	0.6	0.396193772	0.16	0.566666667	0.106666667	0	0	0	0	0	0
5	0.0625	0.040100251	0.024670558	0.028169014	0.6	0.733564014	0.04	0.133333333	0.44	0	0	0	0	0	0
6	0.5	0.105263158	0.186473314	0	0	0.288927336	0	0.966666667	0.893333333	0	0	0	0	0	0
7	0.5625	0.005012531	0.207413202	0.01056338	0.8	0.674740484	0	0.566666667	0.893333333	1	0	0	0	0	1
8	0.625	0.015037594	0.365003911	0.021126761	0.8	0.425605536	0.04	0.933333333	0.386666667	0	0	0	0	0	0
9	0.65625	0	0.113905771	0.063380282	0.8	0.157439446	0.04	0.6	1	0	0	0	0	0	0
10	0.78125	0.032581454	0.063722246	0.39084507	0	0.51384083	0	0.766666667	0.893333333	0	0	0	0	0	0
11	0.8125	0.005012531	0.169805644	0.11971831	0.4	0.880622837	0.08	0.9	1	1	0	0	0	0	1
12	0.4375	0.055137845	0.184487635	0.024647887	0.4	0.098615917	0	0	0	0	0	0	0	0	0
13	0.5625	0.007518797	0.133281184	0	0.6	0.288927336	0.12	0.533333333	0.613333333	0	0	0	0	0	0
14	0.78125	0.020050125	0.029785186	0.11971831	0.8	0.605536332	0	0.2	0.946666667	0	0	0	0	0	0
15	0.71875	0.01754386	0.137252542	0.066901408	0.8	0.262975779	0.08	0.466666667	1	0	0	0	0	0	0
16	0.34375	0.022556391	0.139057705	0.014084507	0	0.472318339	0	1	0.28	0	0	1	1	0	0
17	0.4375	0.082706767	0.225645346	0	0	0.261245675	0	0.233333333	0.333333333	0	0	0	0	0	0
18	0.53125	0.080200501	0.313737289	0.024647887	0	0.813148789	0.12	0.266666667	0.56	0	1	0	0	1	0
19	0.625	0	0.054395571	0.218309859	0.8	0.346020761	0.04	0.5	0.893333333	0	0	0	0	0	0
20	0.625	0.022556391	0.232264276	0.042253521	0	0.965397924	0.04	0.433333333	1	0	0	1	1	0	0
21	0.53125	0.030075188	0.058186413	0.01056338	0	0.332179931	0	0	0.56	0	0	0	0	0	0
22	0.3125	0.01754386	0.229737048	0	0.8	0.726643599	0.12	0.033333333	0.72	0	0	1	1	0	0
23	0	0.015037594	0.133822733	0.007042254	0.8	0.288927336	0	0.3	0.173333333	0	0	0	0	0	0
24	0.5	0.070175439	0.215536434	0.066901408	0	0.262975779	0.08	0.433333333	0.893333333	0	0	0	0	0	0
25	0.71875	0.035087719	0.073409952	0.024647887	0.6	0.970588235	0	0.033333333	1	0	0	0	0	0	0
26	0.5	0.142857143	0.154040556	0.052816901	0.8	0.733564014	0	0.866666667	0.56	0	0	0	0	0	0
27	0.65625	0.022556391	0.071003069	0.021126761	0.4	0.970588235	0	0.933333333	0.666666667	0	1	0	0	1	0
28	0.53125	0.025062657	0.424453938	0.028169014	0.2	0.939446367	0.04	0.133333333	0.893333333	0	0	0	0	0	0
29	0.9375	0.01754386	0.129309826	0.028169014	0	0.92733564	0	0.166666667	1	0	0	0	0	0	0
30	0.625	0	0.277333173	0.031690141	0	0.242214533	0.08	0.833333333	1	0	0	0	0	0	0
31	0.5625	0.112781955	0.106865636	0.123239437	0.8	0.970588235	0	0.766666667	0.773333333	0	0	0	0	0	0
32	0.46875	0	0.088452976	0.066901408	0	0.288927336	0.04	0.8	0.84	0	0	0	0	0	0
33	0.65625	0.080200501	0.168301342	0	0	0.262975779	0.12	0.633333333	0.613333333	0	0	0	0	0	0
34	0.875	0.042606516	0.154521933	0.045774648	0	0.055363322	0.08	0.5	0.946666667	0	0	0	0	0	0
35	0.5	0.020050125	0.238522173	0.042253521	0.8	0.951557093	0.08	0.033333333	0.72	0	1	0	0	1	0
36	0.6875	0.027568922	0.206691137	0.017605634	0.6	0.288927336	0	0.2	0.893333333	0	0	0	0	0	0
37	0.28125	0	0.074613394	0.070422535	0.8	0.262975779	0.04	0.133333333	0.84	0	1	0	0	1	0
38	0.625	0.005012531	0.064263794	0.01056338	0.8	0.773356401	0.12	0.4	1	0	1	0	0	1	0
39	0.28125	0.030075188	0.064624827	0.007042254	0.4	0.937716263	0.12	0.2	0.506666667	0	0	0	0	0	0
40	0.65625	0	0.031409832	0.024647887	0	0.925605536	0	0.766666667	1	0	0	0	0	0	0
41	0.6875	0.025062657	0.042722185	0.038732394	0.8	0.669550173	0	0.833333333	0.666666667	0	0	0	0	0	0
42	0.40625	0.020050125	0.112040436	0.01056338	0	0.157439446	0.04	0.2	0.72	0	0	0	0	0	0

2- Feature Selection (Correlation):

our data is numerical and categorical so after research we found that the best methods to handle them is Pearson and ANOVA correlation.

- ANOVA Correlation was applied on room type and year of the review as they are categorical and we want the output to be numerical.
- As for Pearson we tried it on the rest of the columns as they are numerical and we want the output to be numerical.

3- Models:

We tried more than one model to choose the best of them according to accuracy.

Model name	MSE	Accuracy
Linear	0.03357117405195145	0.2826225059200652
Random Forest	0.030295117059987713	0.3526280872476978
SVR	0.034258421636806695	0.2679368130851052