OpinRank

Introduction

This document describes the application of the model developed in the MovieLens project to part of the information contained in the OpinRank dataset from the UCI Machine Learning Repository. More specifically, slightly modified versions of the files containing the hotel review summaries are used here.

The main goal is to reliably predict the variable `overall_rating` (decimal). Out of all the available predictors, only the following decimal variables are being considered: `cleanliness`, `room`, `service`, `location`, `value` and `comfort`. The remaining predictors are not expected to contain relevant information, but coming up with a more comprehensive approach to fully maximise the aforementioned variables is a matter of future work.

The analysis described in this paper goes through the following steps:

- Calculating the ideal format for each predictor (i.e., number of significant digits).
- Determining the most relevant predictor for each city.
- Using the information above to predict the target variable for each city.

The accuracy of the predictions is assessed using the root mean square error (RMSE). The obtained values (0.3123577 and 0.2776807) indicate an excellent performance.

Analysis

The final model from the MovieLens project was precisely meant to deal with datasets containing personal opinions together with other variables somehow related to them. In any case, it is still quite preliminary and should not be expected to be able to deal with notably different conditions right away.

The main changes introduced with respect to the MovieLens project are the following:

- The train/test partition ratio is now 0.2 in order to better account for the appreciably smaller amount of data.
- The `lambda` value is always 0. Regularisation only makes sense with bigger and more variable datasets.
- Only one type of bias is considered, but its calculations have to be dynamically adapted to different variables.
- The input information (e.g., train/test partitioning) has to be redefined at various points. The test set is only used for the final predictions, divided into two groups of cities. All the intermediate calculations are exclusively relying on the information contained in the training dataset.

The version of the model used in this paper can be defined with the following equation:

$$rating = aver_r + bias + \epsilon$$

All the calculations are performed at the city level. The exact predictor used to calculate the corresponding `bias` is also being defined for each city.

The number of decimal digits is quite important and variable for all the predictors. This reality can have a relevant impact on the reliability of the calculations and, consequently, needs to be addressed.

The model is tested against a reasonable range of significant decimal digits (i.e., 0 to 3) for all the possible scenarios (i.e., all the cities for each variable). The number of significant digits whose predictions get the

lowest average RMSE value across all the cities is assumed to be the best one for the corresponding variable. The number of significant decimal digits used for all the predictors are listed in the following table.

Table 1: Significant digits.

Variable	Digits
cleanliness	0
room	0
service	0
location	0
value	0
comfort	1

The next step is to determine the predictor to use to calculate the bias for the given city. The model is once again run for all the scenarios with the lowest RMSE indicating the best variable. The tables below show the predictors chosen for all the cities and the RMSE values which were used to perform those selections.

Table 2: Best predictors.

City	Predictor
beijing	room
chicago	cleanliness
dubai	value
las-vegas	room
london	cleanliness
montreal	cleanliness
new-delhi	value
new-york-city	service
san-francisco	cleanliness
shanghai	cleanliness

Table 3: Training set RMSE values.

City	Cleanliness	Room	Service	Location	Value	Comfort
beijing	0.2307616	0.2176740	0.2539358	0.2458402	0.2232576	2.3715139
chicago	0.2610983	0.2812545	0.3358685	0.4384611	0.3368911	1.5221086
dubai	0.3562643	1.0748647	0.4009234	0.4696602	0.3417341	1.0288113
las-vegas	0.3081844	0.2557895	0.3359179	0.4081305	0.3393506	1.4119575
london	0.3018856	0.3328636	0.3405948	0.5970042	0.3687337	0.9522912
montreal	0.1913841	0.2354326	0.3075468	0.4993096	0.3807923	1.5758826
new-delhi	0.4888912	0.3010464	0.5222091	0.6414523	0.2853605	1.2270627
new-york-city	0.2480788	0.2072061	0.1972769	0.3413482	0.2865790	0.3194230
san-francisco	0.2610333	0.2616852	0.2676135	0.3890292	0.2927765	0.8159671
shanghai	0.2020740	0.2320703	0.2365814	0.2392958	0.2608831	2.2241155

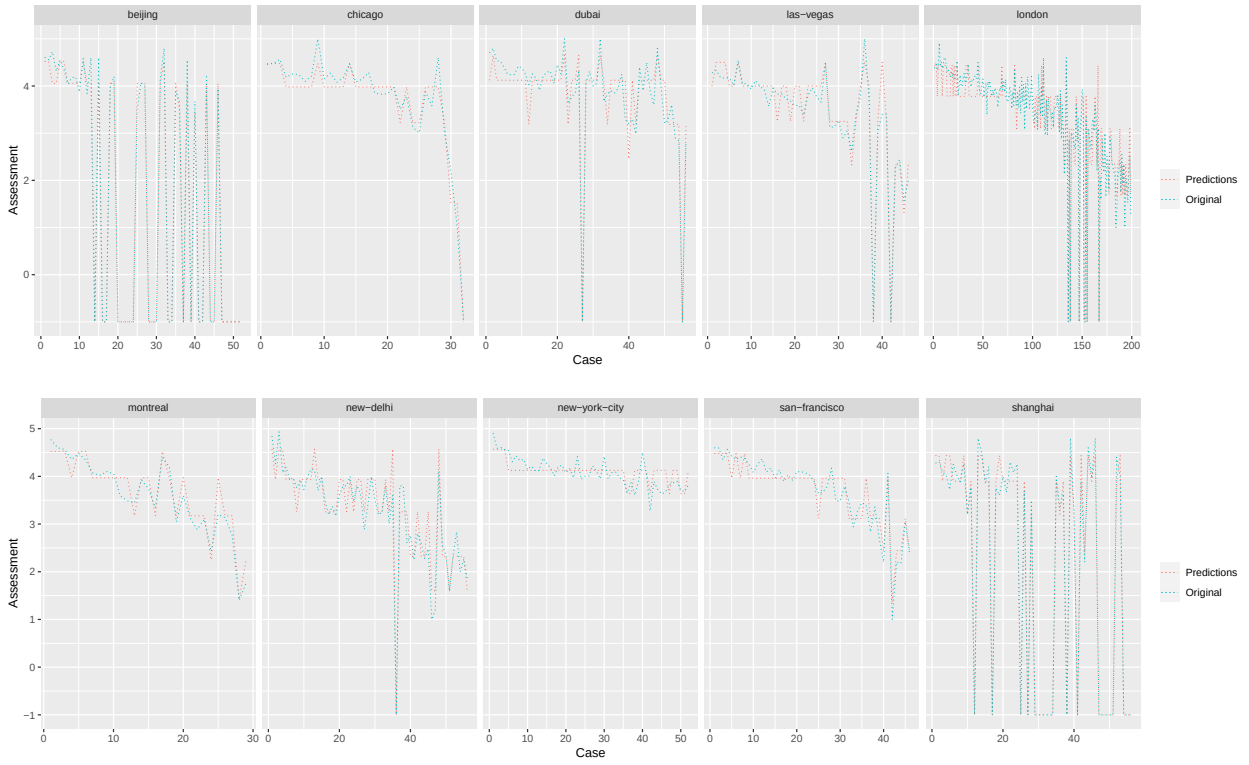
Some of the values in the datasets seem to be invalid, but there is no easy way to confirm that assumption. Additionally, the performance of the model is already very good and, consequently, the impact from removing those apparent outliers is likely to not even be noticeable. The default data cleaning actions of the original model, namely removing values outside the boundaries of the training set, have been performed.

The model can now generate the final predictions from the main test dataset by relying on the best parameters for each city.

Results

The performance of the model under the proposed conditions is certainly good. In the MovieLens project, for example, a RMSE value below 0.86490 was considered a success and the values here (0.3123577 and 0.2776807) are noticeably better than that. Note that the final results have been divided into two subgroups for eminently practical reasons.

The correlations between the predictions and the real values are also very strong (0.98 and 0.983). This aspect can even be visually confirmed by looking at the graphs below.



Conclusion

This model can be safely assumed to understand the reality described under the proposed conditions. And, with that, the main goal of this document has been fulfilled: the suitability of the MovieLens model for predicting similar situations is confirmed.

Relevant improvements are certainly possible. In fact, the contents of this document could even be considered a rather simplistic application of a very flexible approach. It would be possible, for example, to account for various predictors at the same time or to target other variables. Trying to further maximise different aspects of this approach, with this or other datasets, is a matter of future work.