DIPLOMA THESIS 2019-2020

Development of Model-based Recommender Systems using Classification Algorithms

Zoi Chalil, 246672

Supervisor: Nikos Karacapilidis, Professor

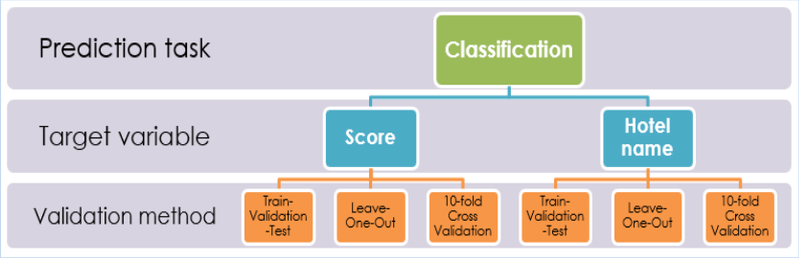
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

DEPARTMENT OF MECHANICAL ENGINEERING AND AERONAUTICS
Laboratory of Industrial Management and Information Systems

ABSTRACT

Recommender Systems were developed to address the information overload problem resulted from the rapidly increasing use of the Internet, in order to meet the individual needs of each user. Although having a wide scope of application, they are mainly used in the field of e-commerce to provide the user with the opportunity to easily find the items that interest him the most, which benefits both the companies and the customers. The main objective of this thesis is to develop models, based on machine learning algorithms, that generate recommendations. At first, the theoretical background of Recommender Systems is described and the most well-known approaches for generating personalized recommendations are presented. The basic data mining procedures used in this context are analyzed and then applied to our own dataset, which consists of user ratings for various hotel units. The experimental analysis is divided into two cases of recommendations. In the first one, the goal is to predict the rating a user would give to a hotel, and in accordance we decide whether a recommendation of the given hotel will be made or not. In the second one, we try to directly predict the hotel that a user will visit. A detailed approach on how to address the recommendation problem in either case, along with the arising challenges are presented. For the development of these models, supervised learning techniques are proposed and more specifically classification algorithms are engaged. Finally, the predictions of each classification algorithm, for the two cases, are evaluated and the final conclusions are drawn, followed by suggestions for possible modifications in future work.

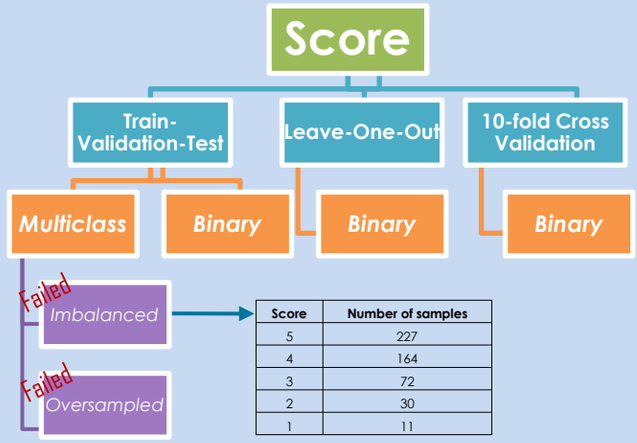
IMPLEMENTATION OF CLASSIFICATION ALGORITHMS FOR HOTEL RECOMMENDATIONS



Design of the experimental analysis

- 504 users, 21 hotels, 5-point rating scale
- Pre-processing: Feature selection, Encoding categorical variables, Feature transformation
- Machine learning algorithms: Decision tree, Random forest, Logistic regression, Naïve Bayes, SVM, KNN
- Implementation: Python language, Jupyter notebook, Scikit-learn, Pandas, Numpy, Matplotlib and Seaborn libraries
- Parameter tunning for both cases ➔ using the validation set in Train-Validation-Test split

TARGET VARIABLE: SCORE



Score	Number of samples
5	227
4	164
3	72
2	30
1	11

Experiments conducted for target variable Score

Multiclass classification yielded low accuracy training scores ➔ transformation to binary classification: 5&4="Like", 1,2&3="Dislike" ➔ Better accuracy scores

Mean Training Accuracy						k-fold (binary)
DT	RF	LR	NB	SVM	KNN	
0.776 (0.001)	0.813 (0.007)	0.811 (0.007)	0.791 (0.009)	0.977 (0.003)	1.000 (0.000)	

Evaluation on train data using k-fold cv in a binary classification setting ("Score")

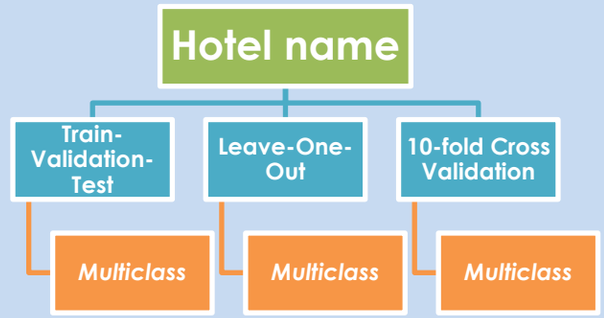
Mean Testing Accuracy						k-fold (binary)
DT	RF	LR	NB	SVM	KNN	
0.772 (0.011)	**0.774** (0.022)	0.751 (0.039)	0.767 (0.037)	0.764 (0.032)	0.755 (0.038)	

Evaluation on test data using k-fold cv in a binary classification setting ("Score")

Random Forest and Decision Tree models obtain the highest accuracy scores (77%).

This is still a classification problem with imbalanced classes ➔ in real-life cases, additional evaluation metrics should be considered in order to better capture the classifiers' performance. Two additional metrics proposed for such cases, are Confusion Matrix and F1 score

TARGET VARIABLE: HOTEL NAME



Experiments conducted for target variable Hotel Name

Mean Training Accuracy						k-fold
DT	RF	LR	NB	SVM	KNN	
0.705 (0.017)	0.824 (0.012)	0.778 (0.007)	0.739 (0.008)	0.868 (0.006)	0.995 (0.001)	

Evaluation on train data using k-fold cv in a binary classification setting ("Hotel name")

Mean Testing Accuracy						k-fold
DT	RF	LR	NB	SVM	KNN	
0.604 (0.031)	0.630 (0.039)	**0.636** (0.037)	0.598 (0.047)	0.611 (0.032)	0.460 (0.052)	

Evaluation on test data using k-fold cv in a binary classification setting ("Hotel name")

In this case the accuracy of the classifiers on unseen data is considered moderate to low. A reason may be the large number of classes (21 hotels), almost the same as the number of features, along with the fact of not enough observations (24 user ratings per class), for each class to be considered well-represented.

CONCLUSIONS

Challenges faced in the prediction task = overfitting & imbalanced classes

Classification model proposed to be used for hotel recommendations = a Random Forest classifier, with target variable for prediction the user rating ("Score"). When this system is provided with a hotel and its features/amenities, it combines them with the target user's profile, and predicts if he would like it or not, i.e. if the prediction of a hotel is "YES", then that hotel will be recommended.

Future work suggestions: consider feature importance in feature selection, better parameter setting, implementation of Ensemble models