Predicting the Accuracy of Regression Models in the Retail Industry[1]

Fábio Pinto[2] and **Carlos Soares**[3]

Abstract. Companies are moving from developing a single model for a problem (e.g., a regression model to predict general sales) to developing several models for sub-problems of the original problem (e.g., regression models to predict sales of each of its product categories). Given the similarity between the sub-problems, the process of model development should not be independent. Information should be shared between processes. Different approaches can be used for that purpose, including metalearning (MtL) and transfer learning. In this work, we use MtL to predict the performance of a model based on the performance of models that were previously developed. Given that the sub-problems are related (e.g., the schemas of the data are the same), domain knowledge is used to develop the metafeatures that characterize them. The approach is applied to the development of models to predict sales of different product categories in a retail company from Portugal.

1 Introduction

The retail industry is a world of extreme competitiveness. Companies struggle on a daily basis for the loyalty of their clients through diverse marketing actions, while providing better products, better prices and better services. The growing need for analytic tools that enhance retailers performance is unquestionable, and Data Mining (DM) is central in this trend [2].

Sales prediction is one of the main tasks in retail. The ability to assess the impact that a sudden change in a particular factor will have on the sales of one or more products is a major tool for retailers. DM is one of the approaches for this task.

In early approaches to predict sales, a single model could be used for a whole business. As more detailed data becomes available, retail companies are dividing the problem into several sub-problems (predict the sales of each of its stores or product categories). The same trend can be observed in several industries [6].

In this approach, there are obviously many similarities between the sub-problems. Not only is the structure of the data typically the same across all sub-problems (e.g., the variables are the same and their domains are similar) but also the patterns in that data may have similarities (e.g., the most important variables across different problems are also similar). Therefore, the process of model building should not be independent. The knowledge obtained from generating a model for one sub-problem can and should be applied to the process of developing the model for the other sub-problems. Different approaches can be used for that purpose, two of them being MtL and transfer learning [1].

Our goals with this work is to use metalearning (MtL) to predict the performance of one model based on the performance of models that were previously developed to predict sales of product categories in a retail company in Portugal, and unveil the attributes that are more important for that prediction. The paper is organised as follows. In Section 2, we briefly survey the concept of MtL and the importance of metafeatures. Our case study is presented in Section 3. Finally, in Section 4 we expose some conclusions.

2 Metafeatures for Metalearning

MtL can be defined as the use of data about the performance of machine learning algorithms on previous problems to predict their performance on future ones [1]. For more information on MtL, we refer the reader to [5, 1].

One of the essential issues about MtL are the metafeatures that characterize the problem. Which metafeatures contain useful information to predict the performance of an algorithm on a given problem? Much work has been done on this topic (e.g., [3]). Typically the work on MtL includes problems from different domains, so the metafeatures need to be very generic (e.g., number of attributes and mutual information between symbolic attributes and the target). However, in more specific settings, metafeatures should encode more particular information about the data, which probably contain useful information about the performance of the algorithms.

3 Case Study

The base-level data used in this study was collected to model monthly sales by product category in a Portuguese retail company. We also gather 9 variables that describe store layout, store profile, client profile and seasonality. Six regression methods from R packages were tested: Cubist, NN, SVM, Generalized Boosted Regression, MARS and Random Forests (RF). The DM algorithm with the most robust performance was RF.[4] The models for the 89 categories were evaluated using the mean percentage error (Eq. 1) where f_i is the predicted value and a_i the real value.

[1] This work is partially funded by the ERDF – European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project "FCOMP - 01-0124-FEDER-022701"

[2] Faculdade de Economia, Universidade do Porto, Portugal, email: fabiohscpinto@gmail.com

[3] INESC TEC/Faculdade de Economia, Universidade do Porto, Portugal, email: csoares@fep.up.pt

[4] The *randomForest* package was used to fit RF models.

$$MPE = \frac{\sum_{i=1}^{n} |\frac{f_i - a_i}{a_i}|}{n} \quad (1)$$

The estimates were obtained using a sliding window approach where the base-level data spans two years. For the majority of the models, one and a half year (approximately 75% of the data) was used as training set and the remaining half year was used as test set. For categories containing just one year of data, the first 9 months were used as training set and the remaining instances as test set. The results are summarized in Table 1.

Table 1. Summary of base-level results (MPE)

Min.	1st Q.	Median	Mean	3rd Q.	Max.
0.072	0.091	0.116	0.212	0.211	1.209

Modelling the variance in results across different product categories is important not only to predict the performance of models but also to understand it. A better understanding of the factors affecting the performance of the algorithm may lead us to better results. For that purpose, we used a MtL approach with the following problem-specific metafeatures:

- Number of instances
- Type of sliding window
- Variables that capture the amount of variation in store layout
- Variables that capture the diversity of store profile in the data
- Variables that capture the amplitude[5] of sales in the test set
- Variables that capture the amplitude of store layout attributes in the training and test sets

The metadata contained 89 examples (corresponding to the 89 categories) described by 14 predictors. The meta-level error of RF for regression was estimated using 10-fold cross-validation.[6] The number of trees was set to 500, as improvement in performance was not found for larger values; and 3 values where tested for the m_{try} parameter, which controls the number of variables randomly sampled as candidates at each split [4]. The results are summarized in Table 2.

Table 2. Meta-level results obtained with all metafeatures in terms of Relative Mean Squared Error and the R^2 and their standard-deviations (SD).

m_{try}	RMSE	R^2	RMSE SD	R^2 SD
2	0.148	0.66	0.0847	0.215
8	0.146	0.663	0.0877	0.234
14	0.15	0.65	0.0871	0.224

The best results were obtained with the m_{try} parameter set to 8. Even with a standard deviation of 0.23, a value of 0.66 for R^2 gives us confidence about the capacity of the regression metamodel in predicting the performance of future RF models.

The next step is to identify the metafeatures that contributed the most to this result. The RF implementation that we used includes a function to measure the importance of predictors in the classification/regression model. We applied the algorithm on all 89 instances.

[5] We calculate "amplitude" of a variable by dividing the largest value by the smallest.

[6] The package *caret* for R was used for 10-fold cross validation estimation. Size of each sample: 81, 80, 80, 81, 80, 79, 80, 80, 79 and 81.

Table 3. Importance of Variables.

Rank	Variable
1st	Max(sales)/Min(sales) in training set
2nd	Mean number of changes in shelf size of category
3rd	Number of instances

The R^2 obtained was 0.93. The importance of the variables for this model is summarized in Table 3. These results show evidence that the metafeatures that represent the amplitude of the dependent variable in the training set and the amount of variance in data are very informative for predicting the accuracy of regression models.

Finally, we used these results for meta-level variable selection. We re-executed the meta-level experiments, again estimating the performance of the RF for regression with 10-fold cross-validation, with only the three most informative metafeatures. The results are summarized in Table 4.

Table 4. Meta-level results obtained with three selected metafeatures in terms of Relative Mean Squared Error and the R^2 and their standard-deviations (SD).

m_{try}	RMSE	R^2	RMSE SD	R^2 SD
2	0.138	0.678	0.0552	0.234
3	0.14	0.687	0.0571	0.23

The results shown in Table 4 are even better than those obtained previously. However, they must be interpreted carefully, as the same dataset was used to do metafeature selection and test its effectiveness, thus increasing the potential for overfitting. Nevertheless, this evidence let us believe that some of the metafeatures used before were carrying noise and that results can improve by metafeature selection.

4 Conclusions

We successfully used MtL with domain-specific metafeatures to predict the accuracy of regression models in predicting sales by product category in the retail industry. Our work shows evidence that, when possible, using domain knowledge to design metafeatures is advantageous.

We plan to extend this approach to predict the performance of multiple algorithms. Additionally, we will compare these results with the results of MtL using traditional metafeatures.

REFERENCES

[1] Pavel Brazdil, Christophe Giraud-Carrier, Carlos Soares, and Ricardo Vilalta, *Metalearning: Applications to Data Mining*, Cognitive Technologies, Springer, Berlin, Heidelberg, 2009.

[2] Thomas H Davenport, 'Realizing the Potential of Retail Analytics', *Working Knowledge Research Report*, (2009).

[3] Alexandros Kalousis, João Gama, and Melanie Hilario, 'On Data and Algorithms: Understanding Inductive Performance', *Machine Learning*, **54**(3), 275–312, (2004).

[4] A Liaw and M Wiener, 'Classification and Regression by randomForest', *R News*, **II/III**, 18–22, (2002).

[5] F. Serban, J. Vanschoren, J.U. Kietz, and A. Bernstein, 'A survey of intelligent assistants for data analysis', *ACM Computing Surveys*, (2012). in press.

[6] Françoise Soulié-Fogelman, 'Data Mining in the real world: What do we need and what do we have?', in *Proceedings of the KDD Workshop on Data Mining for Business Applications*, eds., R Ghani and C Soares, pp. 44–48, (2006).