Advanced Analytics in Business / Big Data Platforms & Technologies: Lab Report Group 1

Jeroen Frans, Jari Peeperkorn, Steven Van Goidsenhoven, Wouter Dossche, Hans Weytjens

April 28, 2020

1 Assignment 1: paper review

2 Assignment 2: predictive modeling

2.1 Exploratory Analysis of the Data

- small dataset, many variables

2.2 Preprocessing

- reduce number of categorical levels
- eliminate features (correlation matrix not all that helpful)
- log
- 1. data selection: decided against using external data
- 2. exploration: used correlation plots, histograms (for targets)
- 3. cleaning: eg. capital to small..., filling missing values (making sure it works on unseen data "other")
- 4. transformations: logs, categoricals one-hot, skewness correction
- 5. feature engineering: string splits of cat vars, ...
- 6. feature selection: correlation matrix (and later in process feature importance analysis in models)

2.3 Models

- it's a regression problem
- goal: achieve performance difference on different parts of domain, so that ensemble benefits
- used Huber (=automatic MAE), support vector machine, random forest regression (quintessential bagging technique), gradient boosting regression, k-nearest neighbors, extreme gradient boosting regressor
- tested and rejected MLP, RANSAC, SGD regressor
- adding more models did not help ensemble. Probably to similar. A really different one could have helped
- regularization: used in all models, important parameter / also pruning

2.4 Ensemble

- linear combination, parameters computed with Huber regression: didn't matter all that much
- brought 10 points (compared to best model)

2.5 Evaluation

- selection of metric, score measure: MAE (for models) which is what Seppe error (for ensemble) is. But some models (random forest, gradient boosting) only optimize for accuracy!
- out-of-sample
- models and ensemble (?): 10-fold cross-validation (not a test set, because of small dataset size), NO optimization on 50% test set

2.6 Approach

- min_price, max_price: separately. But ended up using same models for them, but different parameters in the ensemble
- iterative
- feature importance analysis: somehow helpful, known caveats (p. 6.36)
- models optimzed separately, stacked approach inefficient (at least when including hyperparameter tuning for the constituting models)
- prediction average of 10 to reduce variance

2.7 Results and Post-Mortem Analysis

- why we fell back KISS (Seppe). We don't have KISS, but hoped ensemble would robust against overfitting and noise whilst increasing performance
- winning group: more feature design. Models don't make so much difference
- what we did not do: outlier analysis (anomaly detection), introduce domain knowledge, binning (categorical vars or outliers), revisit choice of min_price, max_price: separately, regression performance (p 4.56: check residuals, variables with extreme coefficients, sign of coefficients), opening the black box (other than feature importance)

2.8 Tools

We used Python in Jupyter Notebooks, sharing our work via GitHub. We relied heavily on scikit-learn (sklearn) for our models and pipelines. We found the sklearn models libraries very well stocked with performant yet flexible (many useful hyperparameters) models. The sklearns pipelines ar an excellent tool to streamline the XXX process. Implementing custom-made models is still relatively easy. Unfortunately, information cannot be passed between the steps in a pipeline, and the number of columns cannot be reduced. Therefore, some wrapping was required to
Problem: shared_reformat: preprocessing done on training and evaluation set (in cross validation terms) together ????

2.9 Other

- explain why so many attempts on leaderboard
- do some of the 'what we did not do' now?

3 Assignment 3: Spark Streaming with text

4 Assignment 4: ??