2/08/2022

Danielle sherman

CTO, Blackwell Electronics

Overview:

As per your request, I have compiled a “lessons learned” report detailing the technical take-aways dealing with R. To be more specific, utilizing R to answers a series of analytical and predictive tasks in classification, regression, and market basket analysis.

**Classification (Brand prediction)**:

This task consisted of running and optimizing different decision tree classification methods. Specifically, I ran and optimized models using the following methods: C5.0, RandomForest, Stochastic Gradient Boosting, and Partial Least Squares. The Stochastic Gradient Boosting models yielded the best results.

The task was very helpful and informative in terms of working with problems that require regression modeling. I will note that working with corrupt data is not ideal, but that’s how data is, not always uniform.

**Multiple Regression in R**:

This task entailed the analysis of provided historical sales reports and the creation of predictive regression models to forecast the sales volume for select product types. I utilized the following methods: Random Forest Regression, which Quantile Random Forest performing the best, Support Vector Machine Regression, with no reliable models, and Gradient Boosting (Regression), with no reliable models.

The specific issues with SVM and GBM models was documented in an early reply, but the biggest issue was the negative values within the models, which is an indicator of a poor prediction.

The task provided a chance to practice regression modeling with caret in R. I enjoyed the task and think it is a good approach for future projects. However, the dataset provided was quite small and caused a high occurrence of overfittings. Therefore, I would say the predictions are not the strongest.

**Market Basket Analysis**:

This tasks required a market basket analysis on Electrondex’s sales history. The analysis was created to help determine the profitability of acquiring Electondex. I needed to clean the dataset and remove redundant rules in order to provide a more accurate analysis.

I utilized the Apriori algorithm within the arules package to asses product association rules. I build a great deal of models in order to find the strongest rules. Out of thousands of transactions, the strongest rules occurred no more than 20 times. In the future, I would be much more confident in recommending or not recommending whether to aquire a company if the sales history was at last 6 months worth. That being said, we are able to glean some actionable data, albeit not the most reliable or confident.

**Conclusion:**

Blackwell should follow the same path in terms of how data analytic projections are concerned. The only caveat is the lack of data. I also realized that it is not always possible to provide larger datasets and/or data sets with no corrupt data. Fortunately, models can be tune and adjusted, i.e. 2 crossfolds instead of 10, to account for smaller datasets. That being said, I look forward to assisting you in future data analytic projects.

GRANT JOHNSON