Millstone report for machine learning part

Predict product backorders

By: Hediyeh Ledbetter

The training data set the amount of backoredered products are less than 1% of the whole products, so the data is very imbalanced. Therefore I add up the data that "went on backorder" to this sample. I used supervised learning to predict “went on backorder” product according to what they have reordered. The result shows the accuracy of 99%.

Data trained in two supervised models logistic regression, and random forest (Bagging-based

Ensemble).Comparing these two models in logistic regression model first because data was ImBalance it showed high accuracy. That reason is, Logistic regression produces an estimated probability that a particular instance is from the positive class. It caused the classifier to over-predict positive instances. For some classifiers, it is not a significant problem, but I expect that logistic regression might be more sensitive. After balancing the data set, I used regularization with my logistic regression model and used cross-validation to select the regularization hyper-parameter to find a suitable threshold that maximizes the F1 score (or some other metric).

A logistic regression model is searching for a single linear decision boundary in the feature

space, whereas a decision tree is essentially partitioning the feature space into half-spaces using axis-aligned linear decision boundaries. The net effect is that it is a non-linear decision boundary, possibly more than one.It is nice when a single hyperplane does not readily separate the data points, but on the other hand, decisions trees are so flexible that they can be prone to overfitting.

To combat this, I used the random forest. Logistic regression tends to be less susceptible (but not immune!) to overfitting.I used ROC AUC score since it gives the probability of an estimator ranking a positive example higher than a negative example. This way it can evaluate the models before selecting a threshold for the decision function. I looked at ‘Precision’ as validation criteria because it is crucial that as many of the records predicted are correct as possible, so it was not wasted working on false positives. RandomForestClassifier had good Precession score compare to all other classifiers.Model Validation: Model trained on 80% of the data, and 32% is the test data below is the precision score of 10 folds.The features importance of random forest model is sales\_3\_month and lead\_time.

Random Forest precision and recall score Precision-recall curves show how Precision and Recall metrics compete depending on the threshold defined for the decision function of the model. Following is the ROC curve for the case in hand.

After calculating AUC and precision-recall scores, RandomForestClassifier has better AUC

and precision value. I decided to tune and validate these models.For tuning the parameters of the model, I used a mix of cross-validation and randomized search.

Based on the performances of the predictive models, I found tuned RandomForestClassifier as

the most suits predictive model to choose in this project. I recommend this model to the client

as it has Auc 99% and a precision score of 1.