Learning to Predict 7a Charge Offs

This example will illustrate how the recipe in the KMDS repository can be applied to develop a machine-learning solution to predict 7a loan charge-offs based on 2023 data. This could be useful to screen loans this year (2024). This will be a direct application of the recipe template in the KMDS repository.

Caveat Emptor

The goal here is to illustrate how KMDS can be used with a reasonable size machine learning task. The focus here is not the modeling per se, it is how the modeling observations are captured. If you are a loan modeling expert, please bear this in mind when you review. With that said, the approach is not an unreasonable one.

Exploratory Data Analysis

1. In this exercise, there will be limited model selection, so the choice of the workflow is the *KnowledgeExtractionExperimentationWorkflow*
2. The details of the subset of attributes from the raw data file that are candidates for modeling are identified and logged.
3. Exploratory analysis reveals missing values for some attributes, the attributes and the strategy to address missing values are logged.
4. The majority of attributes are categorical. One hot encoding is a common strategy to encode categorical values in modeling. However, an evaluation of the cardinality of each of the categorical values reveals that the resulting one-hot encoding vector will be too large for most libraries. Therefore, modeling should use a data representation that will account for this fact. This finding is logged.
5. Chargeoffs are rare. Less than 5 percent of the loans guaranteed by the SBA were charged off in 2023. The fact that there is an *imbalance* in the target attribute we want to predict is logged. The modeling approach we choose to apply to the task must account for this property.
6. Some data is held out for model evaluation, and the rest is used for model development. The proportion held for model evaluation is logged. The data files are saved as *parquet* for efficiency.

Data Representation

1. The central question for data representation is the cardinality of the feature space.
2. *Feature Hashing* is a common and effective way to deal with this. See this lecture for an explanation and this talk for another perspective.
3. The parameter of relevance here is the dimension we want to use for the feature space. This is a hyper-parameter. I have used 1024 here.
4. Feature hashing is applied to both hold-out (test) and model development datasets. Post featurization, the data is saved in *csr* format. This is very

efficient and helps keep the size of the data files small.

Modeling

1. Ensemble approaches are commonly used in applications with class imbalance. Please see (Le Borgne et al. 2022). This will be the approach we use for model selection. We will evaluate both bagging and boosting approaches to ensembling. This is the model selection experiment.
2. The bagging approach is based on random forests. A nice description of the approach is available in this technical report
3. The boosting approach is based on this paper.
4. Both methods are available in the *imbalanced-learn*(Lemaître, Nogueira, and Aridas 2017)
5. (Weiss 2013) is a good reference for learning from imbalanced data. Accuracy is a misleading measure since we can get good accuracy by simply predicting the majority class in all instances. Balanced accuracy is one measure of performance for imbalanced learning. This is used here. A more interesting performance measure is the *sensitivity*, which measures the performance of the learner to pick the minority class.
6. In this limited evaluation, though bagging produces better-balanced accuracy, the boosting approach produces better sensitivity. So we might get false positives, but we do better catching the chargeoffs. This is a surface-level treatment. Cost-sensitive evaluation is a very practical and evolving area of research.
7. Costs and probabilities can be used to compute the utility of each possible prediction. We can then pick the decision with the highest utility. For example, the utility of approving the loan would be the product of the guarantee fee the SBA receives and the probability of the loan being paid in full. The utility of rejecting the loan would be the probability of the loan being a charge-off and the amount the SBA provides a guarantee for on the loan. See (Elkan 2001) and (Sheng and Ling 2006) for a discussion of how classifiers can be made cost-sensitive. This requires the accurate calibration of probabilities from the classifier to get good results. There exist methods such as Platt scaling to recalibrate a classifier to be more accurate. So a more rigorous approach to developing a model would look at these issues. The decision methodology that is used depends on the risk tolerance of the users of this application. Therefore, a solution can be developed only with collaboration and input from the end user.

Report Generation

Report generation is similar to the analytics exercise and simply involves loading the knowledge base and accessing the observation lists for each type of observation.

References

Elkan, Charles. 2001. "The Foundations of Cost-Sensitive Learning." In *International Joint Conference on Artificial Intelligence*, 17:973–78. 1. Lawrence Erlbaum Associates Ltd.

Le Borgne, Yann-Aël, Wissam Siblini, Bertrand Lebichot, and Gianluca Bontempi. 2022. *Reproducible Machine Learning for Credit Card Fraud Detection - Practical Handbook*. Université Libre de Bruxelles. https://github.com/Fraud-Detection-Handbook/fraud-detection-handbook.

Lemaître, Guillaume, Fernando Nogueira, and Christos K. Aridas. 2017. "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research* 18 (17): 1–5. http://jmlr.org/papers/v18/16-365.html.

Seiffert, Chris, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2009. "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40 (1): 185–97.

Sheng, Victor S, and Charles X Ling. 2006. "Thresholding for Making Classifiers Cost-Sensitive." In *Aaai*, 6:476–81.

Weiss, Gary M. 2013. "Foundations of Imbalanced Learning." *Imbalanced Learning: Foundations, Algorithms, and Applications*, 13–41. https://storm.cis.fordham.edu/gweiss/papers/foundations-imbalanced-13.pdf.