MODEL EXPLAINABILITY IN PREDICTIVE ANALYTICS: A FURTHER INVESTIGATION

Preprint · April 2021

DOI: 10.13140/RG.2.2.36118.68166

1 author:

Jacob Dineen
University of Virginia
6 PUBLICATIONS **2** CITATIONS

SEE PROFILE

MODEL EXPLAINABILITY IN PREDICTIVE ANALYTICS: A FURTHER INVESTIGATION

Abstract

In our exploratory investigation into a strategy for prediction model explainability, we uncovered inconsistencies across the various techniques and prediction methods, as well as anomalies between the actual and predicted values in some forecasting scenarios. This has the undesirable effect of casting doubt upon the efficacy of existing explainability approaches. To address these problems, we undertake here detailed "drilldown" analyses to gauge more precisely the effects of sample size and "what if" perturbations on prediction results. In combination with using a second academic data-set, we obtained much more encouraging results with respect to the uniformity of explainability effects. This has motivated us in our quest to generate "reference narratives" for explaining models to the end user community of organizations and their clients.

1. Introduction

A long-standing problem in predictive analytics has been the disconnect between modelers (statisticians, mathematicians and data scientists) at the model development stage and end-users at the organizational and decision-maker levels for whom the models are ostensibly built in the first place. The latter community typically does not have the analytical background necessary for a full comprehension of the resulting model artifacts whereas the former community often lacks the requisite knowledge of the problem domain driving the requirements of decision-makers. This "cultural divide" too frequently has the undesirable consequence of diminishing the utility of modeling to its intended audience.

This communication problem has migrated into the arena of machine learning (ML) and artificial intelligence (AI) in recent times, giving rise to the need for and subsequent emergence of *explainable AI (XAI)*. XAI has arisen from growing discontent with "black box" models, often in the form of neural networks and other emergent, dynamic models (e.g., agent-based simulation, genetic algorithms) that generate outcomes lacking in transparency. Applications such as facial recognition have met with stern resistance as, too often, mistaken identifications have led to unnecessary and serious disruption in individuals' lives. As another example, pre-existing bias in historic data-sets used for building ML algorithms (MLAs) has resulted in some people being denied loans because of their race and/or gender. This is a very sensitive issue for companies for whom the public perception of policy fairness and impartiality is critical to their business and well-being.

A recent well-received book by John Kay and Mervyn King, "Radical Uncertainty: Decision-Making Beyond the Numbers" [4] highlight similar problems for economic models. The authors suggest a need for "reference narratives" which are effectively stories that can be marshaled to address the overriding objective of unraveling "what's going on here?" We adopt this perspective as our long-term strategy for model explainability.

We have addressed this problem in a preliminary fashion exploring model transparency within the circumscribed context of predictive analytics [5]. Our objective is to expand upon our initial analyses with an eye towards the eventual reconciliation of the two stakeholder communities mentioned above. Model transparency is likely to become a growing area of interest as ML/AI models continue to develop, and organizations and users will continue to demand improved accountability. Translating mathematical and data science expertise into decision-making expertise remains a significant obstacle in gaining organizational acceptance of model artifacts. We believe that advances in model explainability and interpretation are essential to bridge this gap.

2. Previous Work and Methodology

Our objective here is to expand upon our preliminary investigation of explainability techniques, detailed in [5]. In that work, we ran four different prediction models and then compared the following explainability techniques for each model: Local Interpretable Model Explanation (LIME [9]), SHAP (SHapley Additive exPlanations [6,7]), GAM (General Attribute Model [3]), and an SKLearn neural net [8], all applied to a static Lending Club loan applications training data-set, consisting of 80,000 observations (Table 1). The intent of our paper was to examine the consistency of the feature sets across the four techniques and to assess the comparability of these approaches across all four of the different models, both in static and dynamic (i.e., predictive) modes.

Within the context of the loan application data-set, we discovered that the SHAP technique to be the most robust technique for explainability. SHAP and GAM were relatively consistent with respect to identifying the operative *feature sets* in the training data-set but the neural networks did not track as well. We further

discovered that in the dynamic case involving actual prediction, the forecasts did not always align with the static results as well as desired. This has led us to drill down and sharpen our analysis as we describe below.

Our preliminary investigation showed inconsistencies in feature importance across the different explainability measures, and equally disturbing, when comparing actual predictions with the static expectations. In this work, we examine the dynamic, predictive case by performing perturbation analyses on the features which appear to be most influential in model explainability. In an ideal world, these perturbation-based predictive cases should track the corresponding static cases.

Because of the computation costs of Shapley estimators, most Shapley examples have relatively small sample sizes. In that context, we also want to look at the effect of increased sample size on the Shapley estimators to see how well they hold up.

Our approach then is as follows:

- We look at the sensitivity of predictions to sample size, i.e. as the sample size increases, do the prediction-to-expected differences remain, and if so, to what degree? Using SHAP estimators, we show the effect of straight sample size on the feature importance dimension of features identified as major influencers. We expect ideally that sample size would not have a large impact on feature importance measures.
- As a way to minimize sampling bias, we stratify the overall sample into randomly selected equal-sized Subsamples to test the variance of the predictions across these Subsamples. Again, we would like to see the Subsample results match up with the overall sample results.
- We examine the dynamic, predictive case by performing perturbation analyses on the features which appear to be most influential in model explainability. We perturb selected features over the range -0.5 to 0.5 in increments of 0.1 for each of the segments. Again we expect these perturbation-based predictive cases to track the corresponding static cases.
- To verify the results are not 'data-set' dependent, we then recreate these same analyses on a second, different financial data-set.

In this analysis, in addition to the changes discussed above, we have modified the previous experiment as follows:

- In our previous work, we experienced some leakage from the data pipeline which we have largely eliminated. For example, we considered features that were explicitly tied to the active status of a loan. In a production environment where a model would yield a score determining viability and efficacy of a consumer to fulfill the basic requirements of the loan, these features would not be available, which is analogous to a user cold start problem prevalent in recommender system tasks. This is task-dependent, however, as dynamic loan evaluation models could be used throughout a loan's duration to signal potential changes in behavior towards delinquent status.
- We focus solely upon the SHAP model explainability in the current analysis for the following reasons:
    a. The four techniques we examined in [5] did not reveal consistent results across the board. We have thus decided to focus upon only a single technique, namely Shapley Explanations, which [7] have shown to be a more widely used and general explainability strategy.
    b. SHAP provides convenient graphical displays for feature importance which are readily understandable to end-users as well as data scientists as we show in our analyses.
    c. A disadvantage of SHAP is the costly computational time as the sample size increases.
- We consider only 3 prediction techniques (GBC, RF, Logit), discarding the neural net (NN) methodology Deep learning has experienced tremendous industrial adoption this decade, but often require large amounts of data to approximate a functional mapping from input to output space. The largest sample we consider here is 80k, and we were unable to achieve performance on evaluation metrics comparable to generalized linear models or tree-based algorithms. With a moderately extensive grid search of both hyperparameters and neural network architecture, the 'best' model on any Lending Club class-balanced sample was only moderately better than a random baseline model, while other considered algorithms yielded a size-invariant holdout accuracy greater than 80%.

3. Data-Sets[1]

We analyze two data-sets:

1. Lending Club Loan Application consisting of active and past loans. Those completed loans that have been fully paid or have no existing derogatory marks are classified as 'good loans' whereas 'bad loans' are instances where an individual has either defaulted or is currently delinquent. What we want to predict is whether an individual loan is "good" (GoodLoan), i.e. it does not have associated factors such as payment defaults, late payments, high balances, etc which would constitute being a BadLoan. This data-set was used in our initial investigation and contains 80K observations, 5 continuous variables and 16 categorical variables.
2. Census Income from 1994 containing demographic variables from which we want to predict whether income for an individual exceeds $50K/year. This dataset is a popular data-set frequently used in academic papers for assessing MLP (multilayer perceptron) effectiveness. It contains ~30K observations, 6 continuous variables and 8 categorical variables.

For each data-set, we perform these different analyses:

- Feature importance; we compare feature importance from the entire sample (80K for Lending Club and 30K for Census Income) to Subsamples (20K for Lending Club and 10K for Census Income)
- Using perturbation analysis, assess the model operation in dynamic, prediction mode by perturbing the major features identified from -50% to 50% by increments of 10% to see what effects this has upon explainability. The perturbation process varies slightly depending on whether the feature is continuous or categorical. As with Feature Importance, we consider the impact of sampling on the perturbation process.
- Sample Size Analysis. We considered the impact of sample size, but report these only for Lending Club. (The results are similar to the Subsample analysis.)

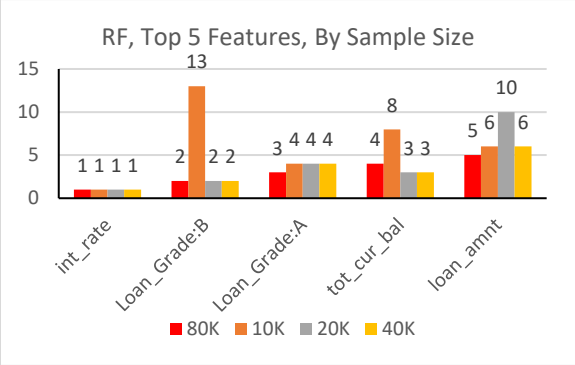
4. Lending Club Loan Application

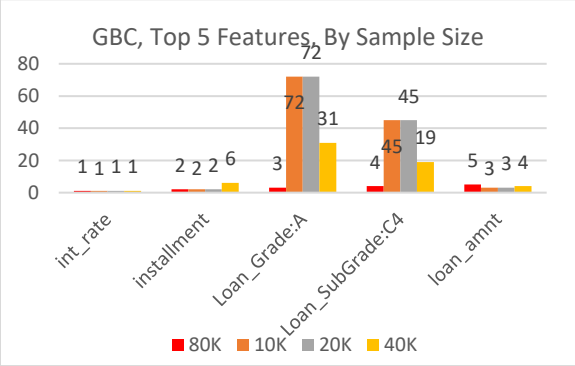
4.1 Explainability Comparisons

In our previous paper, we demonstrated that the Shapley estimates of feature importance were relatively consistent (across estimators, but that the sensitivity in predictions (via perturbation analysis) varies quite widely from expectations based on the predicted feature importance. In particular, we observed insensitive response functions for all the estimators (except logit). To examine this discrepancy further, we perform the sample size and perturbation analyses as described above. We construct for each feature a cardinal metric, namely the Shapley Value relevant to the scenario being analyzed. This enables us to represent feature sets for a scenario in a single graph for easier comparison as we show below.

4.2 Sample Size Analysis

Figure 1 compares the feature importance rank from the SHAP scores of the five most important features using samples ranging from 10K to 80K.





[1]Reference links to data-sets:

- Lending Club Loan Applications: https://www.kaggle.com/wendykan/lending-club-loan-data#LCDataDictionary.xlsx;
- Census Income: http://archive.ics.uci.edu/ml/data-sets/Census+Income

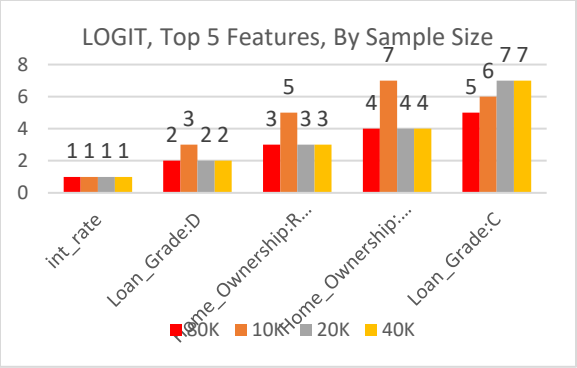
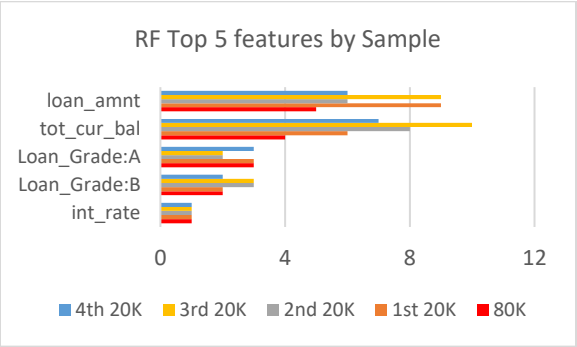


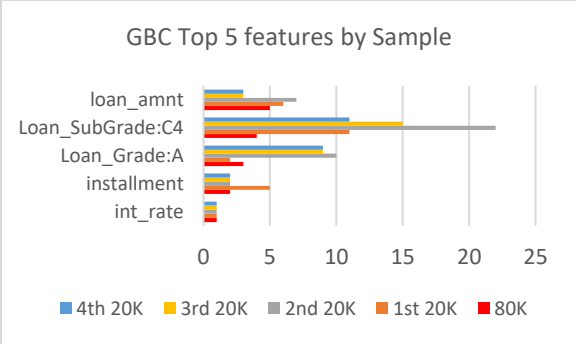
Figure 1. Top5 Feature Importance by Sample Size by Prediction Technique

With the exception of the smallest sample (10K) and the Loan Grade binary features in GBC, we see a high-degree of similarity in the feature importance. Given the similarities between the sample size analysis and the Subsample analysis, we will end the discussion of sample size analysis here to save space.

4.3 Subsample Analysis

Figure 2 compares the feature importance rank from the SHAP scores of the five most important features (using the entire 80K sample) to the feature ranks of each of these features for the four 20K samples (the sample size we used in the previous paper).





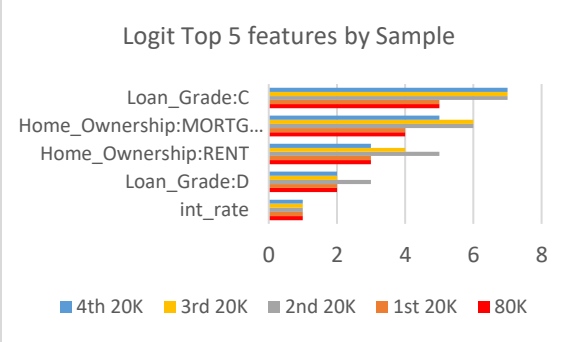


Figure 2. Top5 Feature Importance by Subsample by Prediction Technique

The five most important features from the overall sample is utilized. Looking at the RF (random forest) results as an example, we see that *int_rate* is the most important feature for each of the Subsamples. Loan Grade B is the second most important feature in two of the Subsamples, and is the third most-important feature in the other two Subsamples.

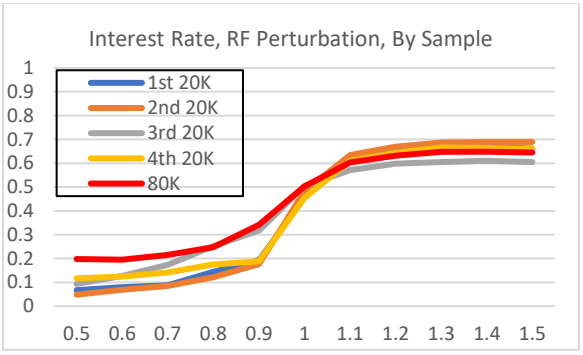
On the plus side, all estimation techniques and samples agree on *int_rate* as the most important feature. There is less uniformity as feature rank falls, particularly if we consider the Top10 (not shown here) rather than Top5 feature sets. For example, for the random forests (RF) technique, *tot_cur_bal* is the 4th most important feature for the entire 80K training set, but only 6th - 10th for the four 20K samples. Overall, there is less disparity between the models than in our previous analysis. We suspect that this is related, at least in part, to the improved data pipeline.

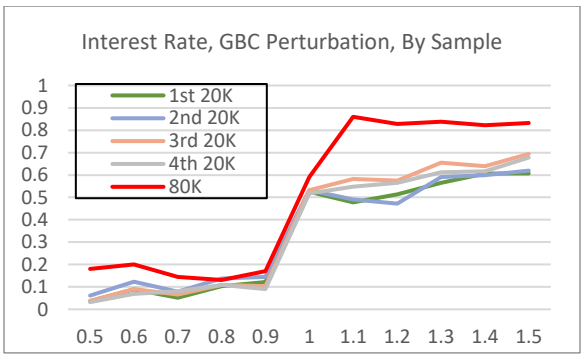
4.4 Perturbation Analysis

There are two cases to consider in the perturbation analysis. For continuous features, we simply increase (or decrease) the feature value by a fixed percentage, e.g., by (say) 50%. In the categorical case, we "turn on" features that were previously "off". This re-coding of binary features also requires that other binary features are changed at the same time.

In the continuous case, we run the model for 10 different scenarios for each of the three different prediction techniques, perturbing the variable in question in increments of .1 ranging from -.5 to .5 to see what, if any, effects, this has upon feature importance (Figure 2). Although we have run the perturbations for the Top10 features in terms of importance, we only show the two examples below for ease and compactness of representation. Note that the y-axes in all cases below represent the prediction probability of a 'good loan'. For a perturbation value of 1, the prediction is the sample enumeration estimate of the predicted probability. For a perturbation value of 1.5, the prediction is for the case where the feature of interest, *Interest Rate* in the case of Figure 3A, is increased by 50%.

Figure 3 compares predictions for two Features across perturbation values for the 3 estimation techniques and the overall 80K sample to the four 20K Subsamples.





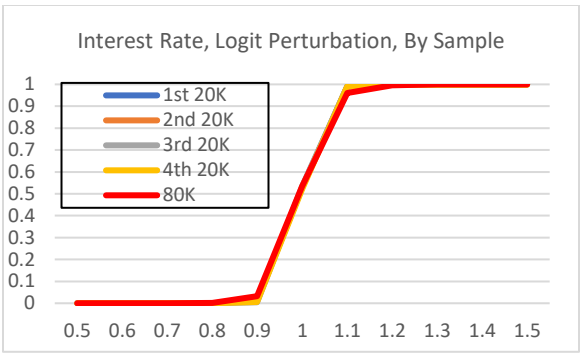
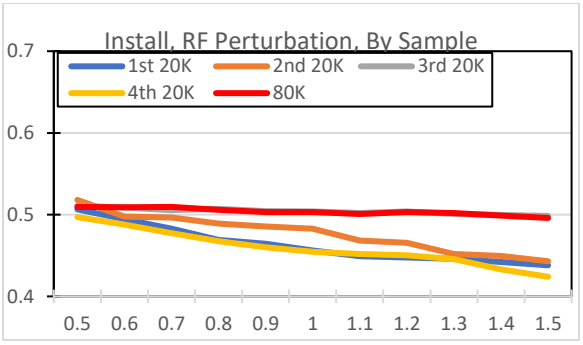
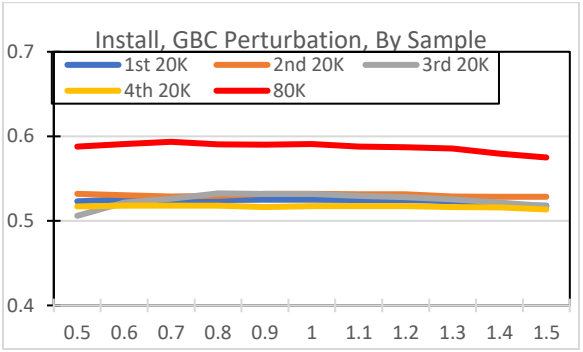


Figure 3A. Perturbations for *Interest Rate* by Prediction Technique by Subsample

In Figure 3A, we see 'universal agreement' in Logit: the 5 prediction response functions are essentially identical. For Random Forest (RF), we likewise see similar results (more similarity for increases in the *int_rate* than for decreases). GBC shows the largest differences—especially on the 'increase' in the interest rate. Note also that the GBC response functions generally show slight sign reversals for slight increases from the base (up to a 20% increase for some samples).





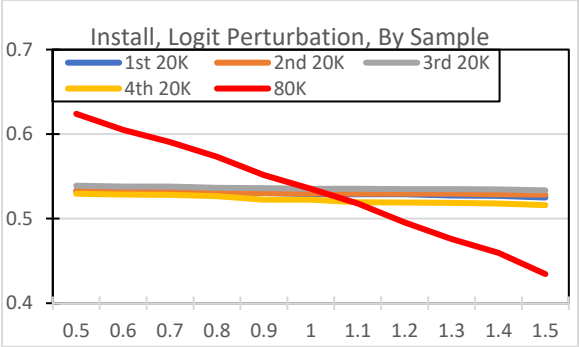
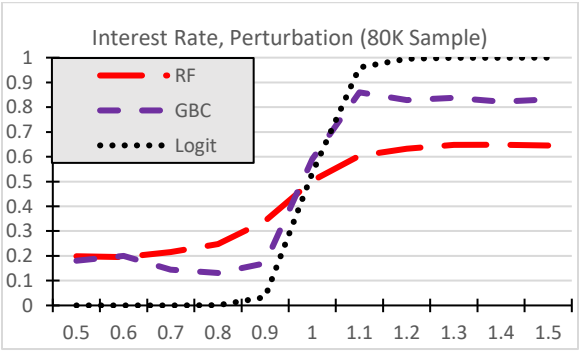


Figure 3B. Perturbations for *Install* by Prediction Technique by Subsample

In Figure 3B for *install*, there is much less agreement---both between the samples and between the estimation techniques. For Random Forest (RF), the Subsamples are responsive than is the overall sample. For GBC, the response functions are all 'flat', but the overall sample is separated from the Subsamples. For logit, we the see the overall sample is fairly sensitive to changes in install, but the Subsamples are insensitive (or flat).

Figure 3A is more like we might have have hoped in the sense that we similar response functions for all estimators), whereas Figure 3B is more similar to our previous analysis that showed large discrepancies between predicted and actual feature importance.

Figure 3C summarizes the response function (for the entire 80K sample), across the 3 estimation techniques for features *Interest Rate* and *Install*. The most important feature, *Interest Rate*, has similar response across the three techniques---though there are slight 'sign reversals' in GBC 'outside' the (0.9, 1.1) interval.



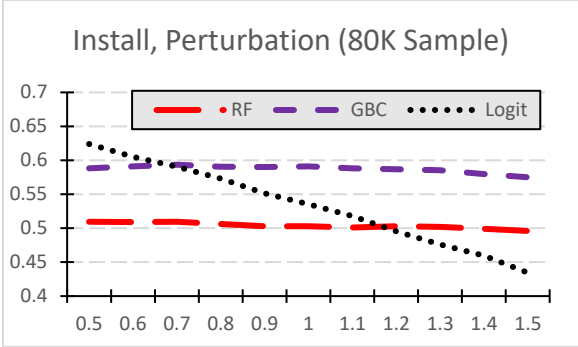
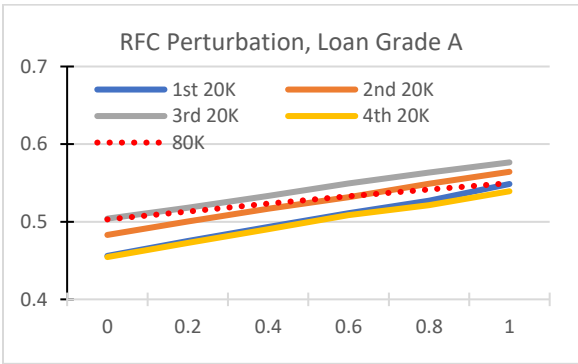
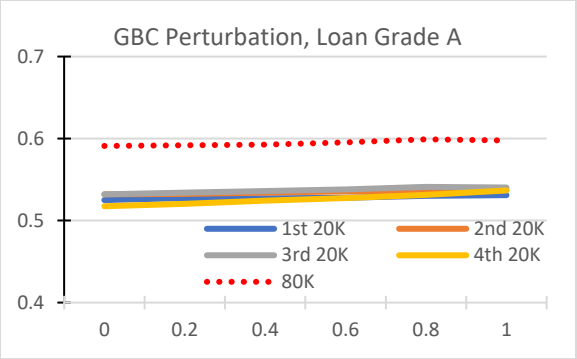


Figure 3C. Perturbation of *Interest Rate* and *Install* Features across Prediction Techniques for 80K Sample

Figure 3C highlights the broad similarities for the response functions between estimation techniques for the most important feature (*int_rate*). There is basically no response for GBC and RF, while logit shows some response to *installation* (8th most important feature for RF, 2nd most important feature for GBC and 12th most important feature for logit).

For the categorical perturbation, we change the number of observations "turned-on" by steps of .2 (from 0 to 1). That is, a step of .2 turns on 20% more observations (than are on in the data-set); 1 then means that we have doubled the proportion of "on's" compared to the actual values.





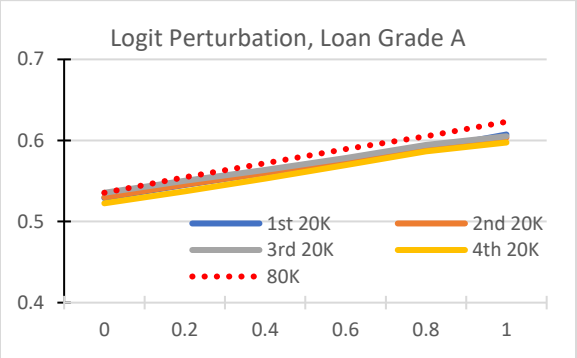
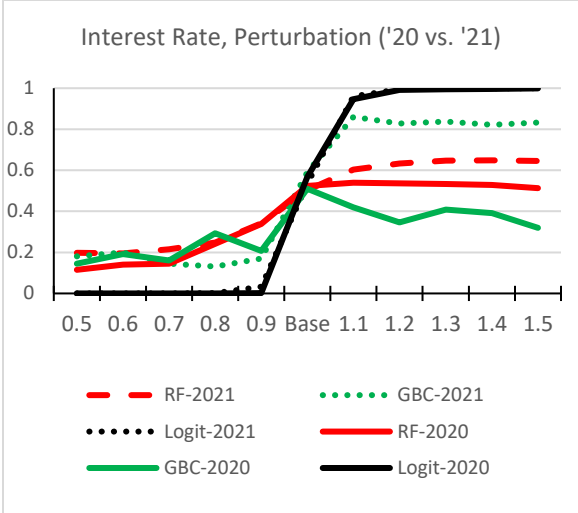


Figure 4. Perturbation of *Loan Grade A* by *Subsample.*

Figure 4 details the perturbation for the categorical feature Loan Grade A (3rd most important feature for RF and GBC and the 9th most important feature for logit). We see that the response functions for RF and logit behave similarly. The surprise here is GBC—both the flatness (given its feature importance) and the fact that the entire sample is so 'different' from the Subsamples (in terms of level).



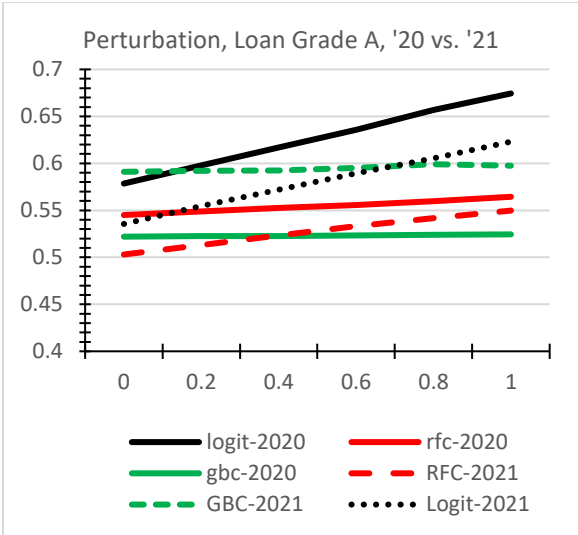


Figure 5. Perturbation Results (2020 vs 2021).

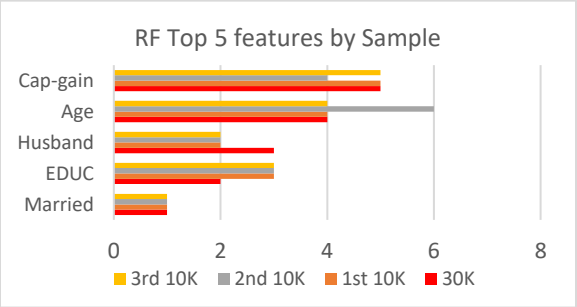
Figure 5 compares the current results to our results from the previous paper for *int_rate* and *Loan Grade A*. We see that the newer models generally behave closer to expectations in the sense that feature importance better correlates with the response functions and there are fewer anomalies in the form of sign-reversals.

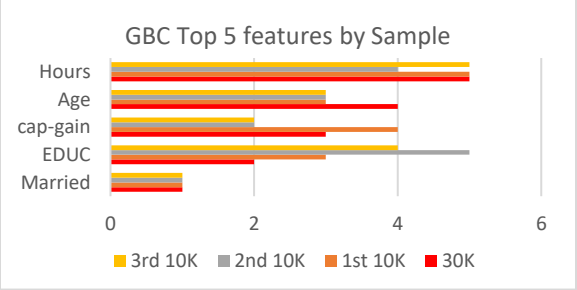
5. Census Income

5.1 Subsample Analysis

Figure 6 below compares the feature importance rank from the SHAP scores of the five most important features (using the entire 30K sample) to the feature ranks of each of these features for the three 10K samples. We see that all estimation techniques find *Married* to be the most important feature.

Interestingly, logit shows the most disparity between the entire sample and the Subsamples; though the results are more uniform in this case.





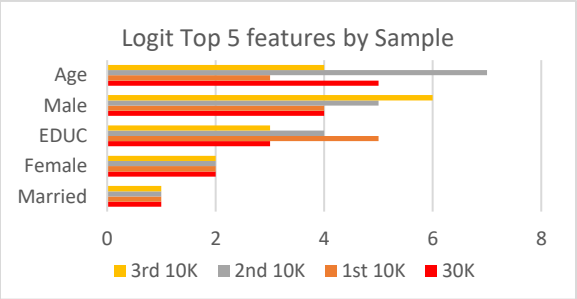
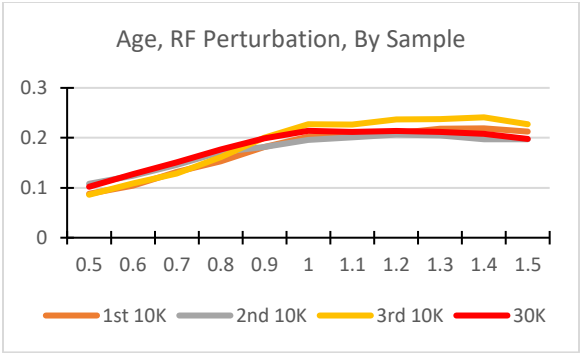
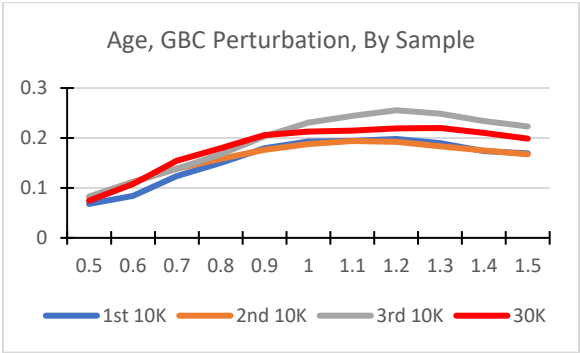


Figure 6. Feature Importance by Subsample Size by Prediction Technique for Census Income Data-Set

5.2 Perturbation Analysis: Continuous Features

As before in the continuous feature case, we exercise the model for 10 different scenarios for each of the three different prediction techniques, perturbing the variable in question in increments of .1 ranging from -.5 to .5 to see what, if any, effects, this has upon feature importance. Figures 7 and 8 below compare predictions for two selected features across perturbation values for the 3 estimation techniques and the overall 30K sample as compared to the three 10K Subsamples.





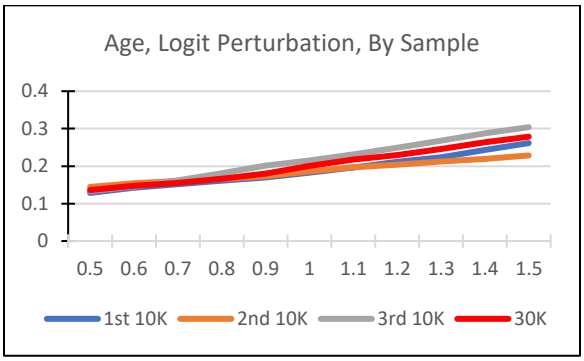


Figure 7. *Age* Feature (continuous) by Prediction Technique by Subsample for Credit Income Data-Set

In Figure 7, we see 'nearly-uniform agreement in the response function, both across estimation techniques and samples. For RF, response function is definitely flatter for increases than decreases. For GBC, the response function once again demonstrates modest sign-reversals (for some increases). Unlike Figures 2A and 2B, the response function for the entire samples falls within the Subsamples. This may be partially explained by the differences in the data-sets; demographics and income are less heterogenous across households than are good loans and financial features.

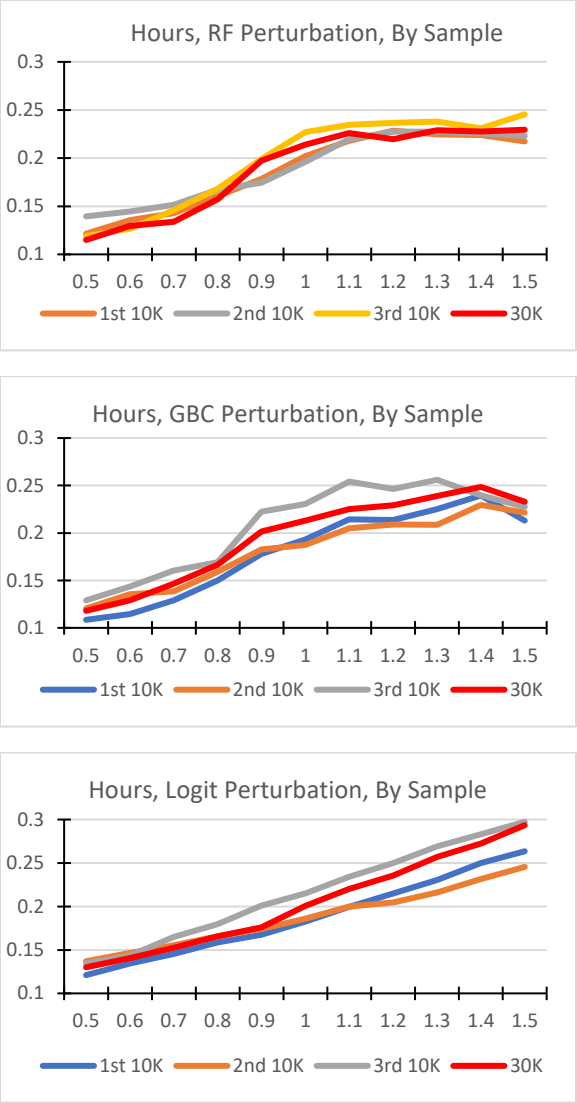


Figure 8. *Hours-per-week* Continuous Feature by Prediction Technique by Subsample

Figure 8 like Figure 7 shows fewer differences in response across the Subsamples and estimation techniques, a further encouraging sign.

5.3 Perturbation Analysis: Categorical Features

Figure 9 below details the perturbation for the categorical feature *Never Married* (9th most important feature for RF, 15th for GBC and the 6th most important feature for logit). We see 'nearly-uniform' agreement in the response functions for all estimation techniques. Once again, the Census dataset appears to be better behaved; perhaps it is more accurate to say that the response function (in the perturbation analysis) conform better to expectations (based on the SHAP feature importance rankings/scores).

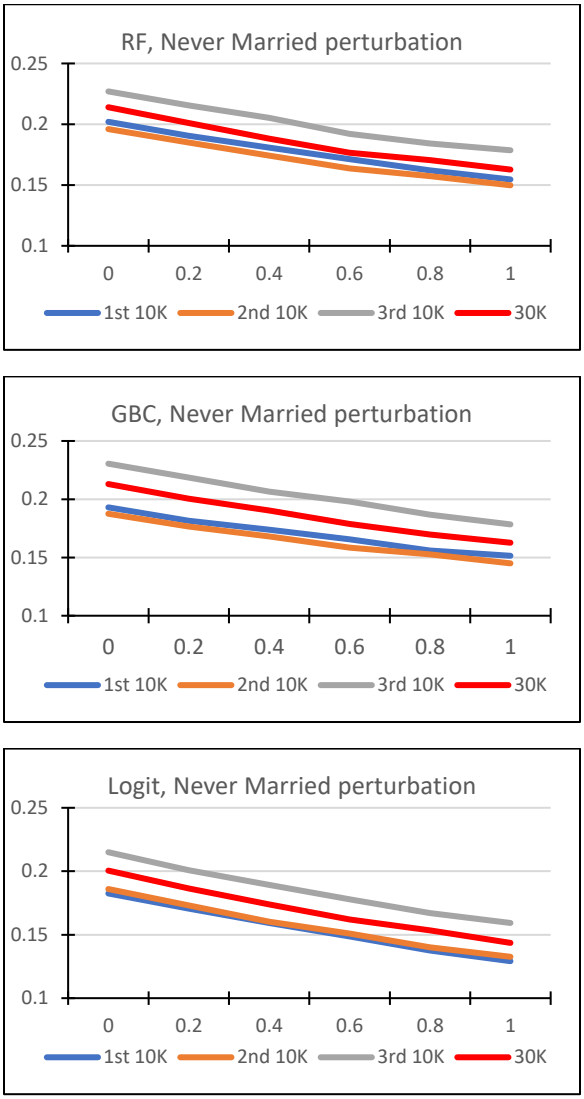


Figure 9. *Marital Status* Categorical Feature by Prediction Technique by Subsample

6. Summary and Conclusions

Our original intent was to utilize an actual firm-specific finance data-set where we could look carefully at the intersection of feature-importance and regulatory constraints. Unfortunately, for circumstances beyond our control, mainly the COVID virus, we were unable to get access to a dataset in a timely manner. As a result, we utilized another public dataset for purposes of comparison.

We generally found similar results across data-sets, although the Census data-set did more closely align

with expectations. Our Subsample analyses seemed to dispel concerns about sampling bias, and the disparity between static and dynamic forecasting scenarios was much less pronounced than our initial analyses which is an encouraging development. Thus, we believe we are on the right track to making model explainability a realistic and valuable tool in the overall decision-making landscape.

Our eventual goal is to develop a methodology which facilitates the generation of "reference narratives" as a way to answer the question "what's going on here?" in any particular model and decision-making setting. "Reference narratives" are essentially stories we can tell to end users to help explain the data and models that utilize the data [4]. In short, we are looking for ways to bridge the gap between the mathematician/ data scientist/statistician and organizational model users.

Our next step in this quest is to apply what we have learned from our analyses here to utilize an actual finance data-set that is the basis of a firm's decisions in a regulatory environment. This will allow us to look more closely into the impacts of regulation on the use of feature importance. We also want to look into potential metrics for the 'differences' in feature importance between estimation techniques, e.g., pairwise distance or similarity.

7. References

[1] Friedman, J.H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 2001, Vol. 29, No. 5, 1189–1232.

[2] Ho, T.K. Random decision forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[3] Ibrahim, M., Louie, M., Modarres, C., Paisley, J. Global explanations of neural networks: Mapping the landscape of predictions. *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*, Honolulu, HI, Jan 27-28, 2019. arXiv:1902.02384v1

[4] Kay, J. and King, M. *Radical Uncertainty: Decision-Making Beyond the Numbers.* W.W. Norton and Company, 2020.

[5] Kridel, D., Dineen, J., Castillo, D., Dolk, D. Model interpretation and explainability: Towards creating transparency in prediction models. *Proceedings of HICSS-53*, 2019.

[6] Lipovetsky, S. and Conklin, M. Analysis of regression in game theory approach. In: *Applied Stochastic Models in Business and Industry* 17.4 (2001), pp. 319–330.

[7] Lundberg and Lee. A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.

[8] Reagen, B., Whatmough, P., Adolf, R., Rama, S., Lee, H., Lee, Hernández-Lobato, H., Wei, G-Y., Brooks, D. Minerva: Enabling low-power, highly-accurate deep neural network accelerators. *ISCA 2016.*

[9] Ribeiro, M., Singh, S. and Guestrin, C. Model-agnostic interpretability of machine learning. In *Human Interpretability in Machine Learning workshop*, ICML '16, 2016.

[10] Train, K. *Discrete Choice Methods with Simulation*. Cambridge University Press 1st ed., 2003 2nd edition, 2009.