SAS 3100-2019

Generating Sales Insights from Predictive Models

Krzysztof Dzieciolowski, Concordia University

ABSTRACT

In a competitive business environment, the ability to predict customers' behaviors is imperative to the success of every company. The probabilities of customers' behaviors are often estimated by predictive models. By rank-ordering the model predictions (scores), we can identify who in the customer population should be targeted to maximize the sales rate. However, the predictive models do not explain why the targeted customers are more likely to buy. As such, models may be ineffective in generating expected business outcomes and could be excluded from the deployment by decision makers.

In this paper, we propose a new inputs-ranking method to disentangle customers' motives to buy. The multi-step approach is based on a correlation analysis between customer attributes and model-ranked predictions. The highest-correlated model inputs can be then selected to reveal individual customers' reasons to buy. The resulting list of high-ranking customers can be deployed with sales together with the customers' individualized insights. Once the attributes are selected, a suitable common-language explanation can be used thus helping build a relationship with the customer. In this paper, the method of inputs-ranking is described together with an illustrative example as well as a heuristic SAS code for generating the described results.

INTRODUCTION

Predicting the likelihood of a binary outcome is imperative for business decision makers. The examples of such events include sales, customer churn, defaulting on loans, making an insurance claim and many others. The challenge of building and deploying predictive models in a business environment is associated not only with the development of appropriate machine-learning techniques, but also to gain an understanding into what "drives" customers to buy, churn or to make a claim. The models provide scores and an overall order of importance of the selected inputs, but the insight into what makes individual customers likely to purchase is missing. In this paper, we suggest an approach to identify and interpret customer-specific inputs that are correlated with customers' predicted values. These inputs can be leveraged to build customized insights and recommendations for each customer. The proposed approach is making models' deployment in sales and marketing more effective.

While there is rich literature assessing predictive models' performance and calibration, including graphical methods, as described recently by Austin (2014), there has been relatively little attention given to the evaluation and interpretation of models' inputs. The image of the predictive model as a "black box" is as prevalent as it was some 20 years ago at the start of the machine-learning revolution. The traditional dilemma of "To Explain or To Predict" as described by Breinman (2001) and Shmueli (2010) have contributed to viewing predictive model as a black-box that have only one purpose – deliver the best possible predictions. The questions whether we need to predict or explain is no longer relevant – both approaches are valid and important in a broad context of developing optimal predictions and deployment-driving explanations. Building a solid understanding of model variables and customer insights is a foundation of both a successful model development as well as its deployment. Organizational objectives of Customer Relationship Management and Sales require an in-depth understanding of customer-specific "drivers" that can be communicated with downstream channel agents where the models are implemented.

In this paper we address a question of model interpretability, not only in terms of its parameters and performance, but also through a better understanding of its inputs. The proposed inputs–ranking methodology helps select main customer attributes and generates individually-customized insights. The correlations between inputs and predictions can be used as a heuristic method of selecting a small number of attributes to "explain" the model and individual motivators to buy. The methodology of interpreting and visualizing the relationship between the dependent variables (target) and individual inputs is proposed.

We introduce inputs-ranking methodology to select model drivers for their interpretation. The visualization of the relationship between these drivers and predicted values is subsequently discussed along with the illustrative business example. Then conclusions and recommendations for further research follow.

RANKING MODEL INPUTS

Inputs selection is one of the major steps in constructing predictive models. Common methods of variable selection include regression iterative methods like stepwise, forward or backwards variables selection, decision trees, PCA and others. These methods are used to construct the model based on the principle of parsimony in order to identify the smallest possible subset of inputs with satisfactory performance. Our interest here is not in selecting variables for the model but, once they are obtained, in interpreting them in the context of their importance for a model's predictions.

The objective of the analysis is to identify inputs that are highly correlated with predicted values. However, these correlations are subject to high variance when calculated in the population. The novelty of our approach is that we suggest to obtain these correlations not in the whole population but in the semideciles. The ranking of the inputs has a smoothing effect on both the inputs and predicted values, thus allowing for the extraction of a meaningful relationship between them. More work is needed to test the statistical significance of the difference between of inputs' variance in the population and semideciles. The illustration of this difference for the Home Equity data is provided in section 2.

Here is an outline of the steps needed to define inputs-ranking methods that identify independent variables with the highest correlation with the model's predicted values.

1. Construct a predictive model for the target response Y and p-inputs $X=\{X_1,X_2,\ldots,X_p\}$.
    a. We assume a binary event being predicted but the same approach can be applied for multinomial and continuous targets.
    b. Identify a subset W of X consisting of inputs which are generally selected for the model.
    c. Rank model's predictions into fixed number of groups, e.g. 20 semi-deciles.
2. Calculate means (or medians) of model's inputs and predicted values in each rank. Then, obtain Pearson correlations r_i between the means of inputs from W and predicted values for all inputs i=1,…,p.
    a. Note that the number of observations for the Pearson correlation coefficients is 20 if we defined ranks as semideciles. If the variable is categorical, then the correlation of the proportion of each category in ranks with the predicted values can be obtained. Let's denote $\bar{Y}$ to be a mean of predicted values in the ranks.

$$r_{X,Y} = Corr(\bar{X}, \bar{Y})$$

where $\bar{X}$ is a mean of X in the ranks.

3. Sort correlations r in the descending order and plot the inputs against the ranks of the predicted values.
    a. Identify a cut-off point to select, say, a small number of insight-generating variables. A subject-domain knowledge would be very helpful in this task, for which collaboration of data scientists with marketing and sales is encouraged.
    b. Select the inputs which satisfy a criterion of high (absolute) value of the correlation coefficient, e.g. $r_{X,Y} > 0.80$, and plot their means against the ranks.

As a result of the input-ranking procedure, you obtain a list of inputs that are highly correlated with the predicted values in ranks.

VISUALIZING THE RELATIONSHIP BETWEEN INPUTS AND PREDICTED VALUES

The visualization of the relationship between input X and predicted values is accomplished through plotting its mean $\bar{X}$ versus the ranks of $\hat{Y}$. We illustrate the process of input-ranking from the previous section (in steps 1, 2 and 3) with the example of Home Equity data (DMAHMEQ) located in the library SAMPSIO of SAS® Enterprise Miner.

STEP 1. CONSTRUCT A PREDICTIVE MODEL FOR HOME EQUITY BAD LOAN

The purpose of the model is to predict a bad Home Equity loan represented by a binary target variable "bad" in the Table 1. There are 5,960 observations and 12 inputs (independent variables).

Variable	Model Role	Measurement	Description
bad	target	binary	default or seriously delinquent
clage	input	interval	age of oldest trade line in months
clno	input	interval	number of trade (credit) lines
debtinc	input	interval	debt to income ratio
delinq	input	interval	number of delinquent trade lines
derog	input	interval	number of major derogatory reports
job	input	nominal	job category
loan	input	interval	amount of current loan request
mortdue	input	interval	amount due on existing mortgage
ninq	input	interval	number of recent credit inquiries
reason	input	binary	home improvement or debt consolidation
value	input	interval	value of current property
yoj	input	interval	years on current job

Table 1. Home Equity loan variables

An exploration of the input variables reveals data issues such as missing values and skewed distributions as shown in Display 1. These issues are dealt with the standard impute and transformations which improve some of the examined models.

Display 1. Home Equity Variables

In a subsequent analysis, we create a number of models from which the most robust seems to be the Gradient Boosting model as shown in Figure 1.

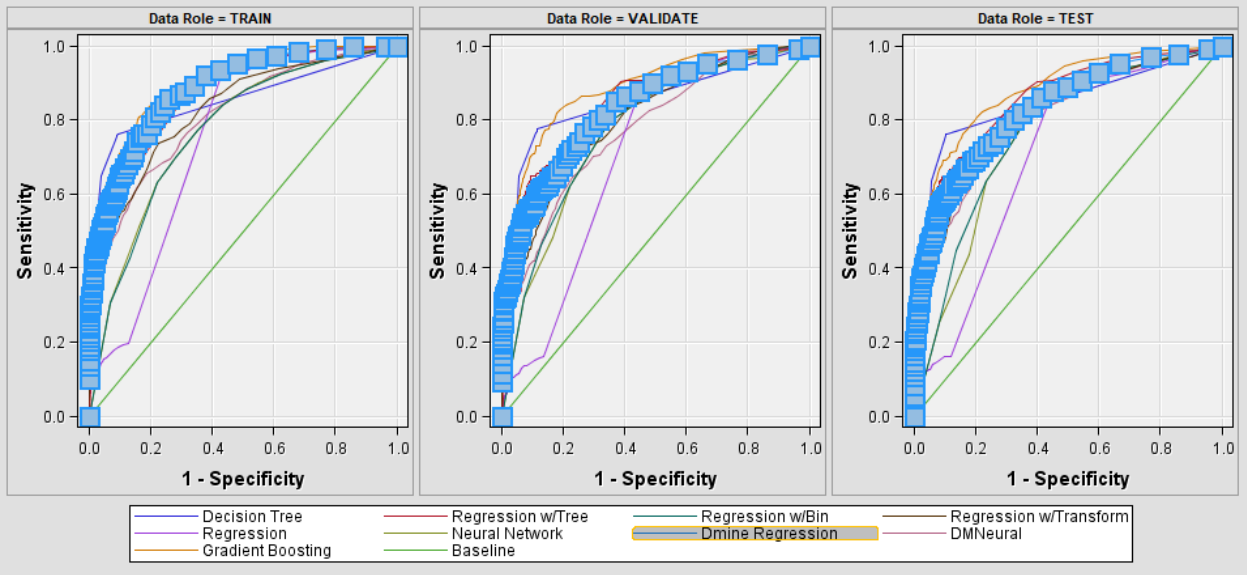

Figure 1. Gradient Boosting model for a Home Equity loan

Model's lift is satisfactory and suggest no overfitting as shown in Figure 2.

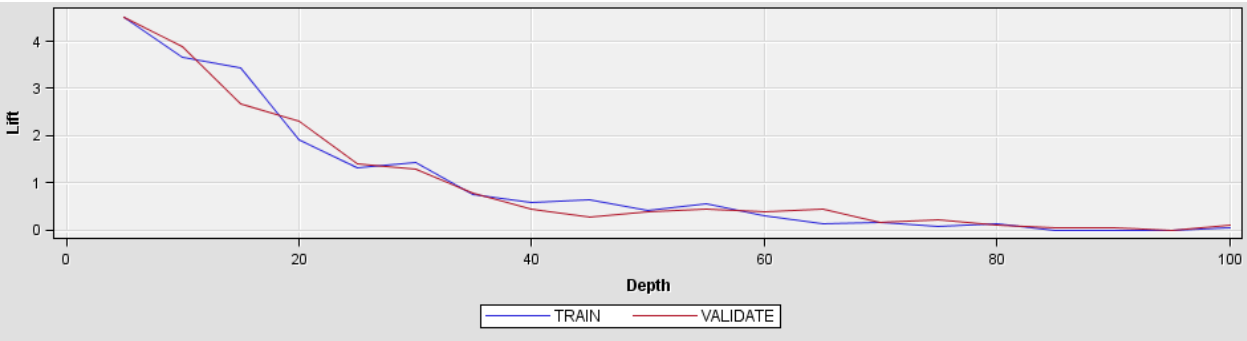

Figure 2. Lift for Gradient Boosting model

The evident differences in the distributions of the Gradient Boosting predicted values for bad and good loans is demonstrated by the corresponding box plots in the Figure 3. The distribution for bad values (1) is shifted towards larger values relative to those of good loans (0).

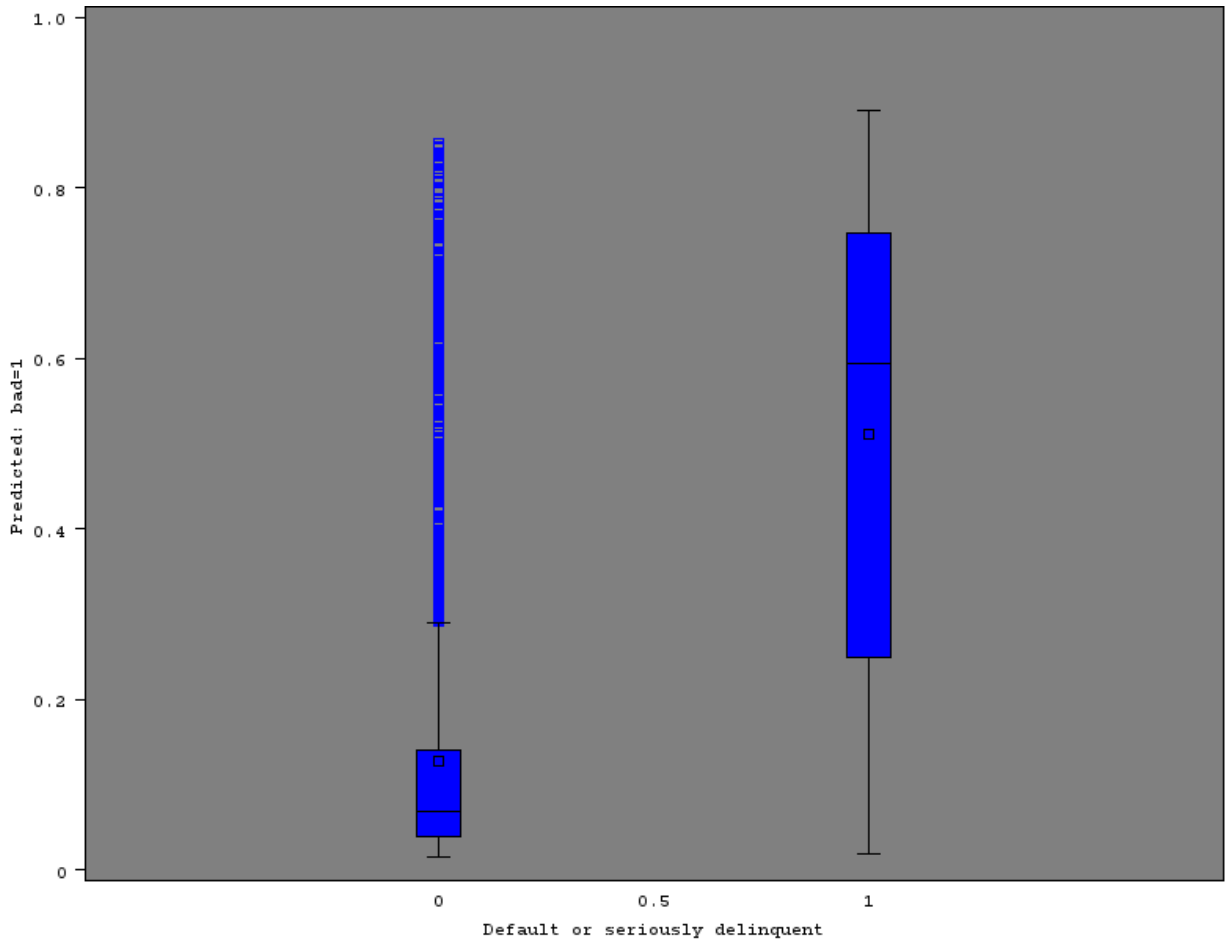

Figure 3. Box plots of the Gradient Boosting predicted values for bad and good loans of Home Equity data.

The variables' importance is shown in the Table 2.

STEP 2. CALCULATING CORRELATIONS BETWEEN INPUTS AND PREDICTED VALUES IN RANKS

First, let's note that the importance order for model inputs is markedly different than the order based on correlations in the ranks. This suggests that the relationships based on correlations cannot be deduced from the model alone. Secondly, we observe that the highest correlation (1.0) is, as expected, between the target variable and its predicted values as seen in Table 1.

As for the inputs, the highest correlation (0.96) is "Debt to income ratio". It is followed by the correlations of "Number of major derogatory reports", "Number of recent credit inquiries" and "Number of delinquent trade lines". These three all exceed the threshold of 0.80 and can be used for interpreting bad loans predictions. In addition, we observe that correlations of the variables in the ranks are substantially higher than in the population. For example, this difference for "Debt to income ratio" is 0.72, indicating a strong smoothing effect (variance reduction) of using ranks' means.

Model Importance	Correlations order in Ranks	INPUT	Rank correlations	Population correlations	Difference: Rank Corr-Population Corr
Target		Default or seriously delinquent	1.00	0.59	0.41
1	1	Debt to income ratio	0.96	0.25	0.72
10	2	Number of major derogatory reports	0.88	0.44	0.44
5	3	Number of recent credit inquiries	0.82	0.31	0.51
2	4	Number of delinquent trade lines	0.81	0.60	0.36
8	5	Amount of current loan request	0.73	0.09	0.64
4	6	Value of current property	0.58	0.05	0.53
3	7	Age of oldest trade line in months	0.56	0.23	0.33
7	8	Amount due on existing mortgage	0.47	0.08	0.38
6	9	Years on current job	0.29	0.08	0.21
9	10	Number of trade (credit) lines	0.05	0.01	0.04

Table 1. Correlations of the target and inputs with the predictive values in the Ranks and Population (excluding categorical inputs).

STEP 3. PREDICTIVE PROFILES OF INPUTS

We illustrate the relationship between inputs and predictions in the ranks for the top ranked variable, such as "Debt to income ratio" and "Number of recent credit inquiries", respectively, in Figures 4 and 5. High predictions of bad loans suggest high values of these inputs, allowing for insights into the nature of loan risk. Such novel insights help create recommendations for sales and marketing when customized to individual customer profiles. The vertical lines specify the high-risk flags defined as 0-5 semideciles. Higher than 40 debt-to-income ratios could be considered a highly correlated with the high-risk of default on Home Equity loan. Similarly, for any inquiry about credit in Figure 5 would be flagged as a being related to a high risk for bad loan. It is up to a user to determine own meaningful thresholds to formulate insights nd recommendations to CRM and real-time marketing solutions.

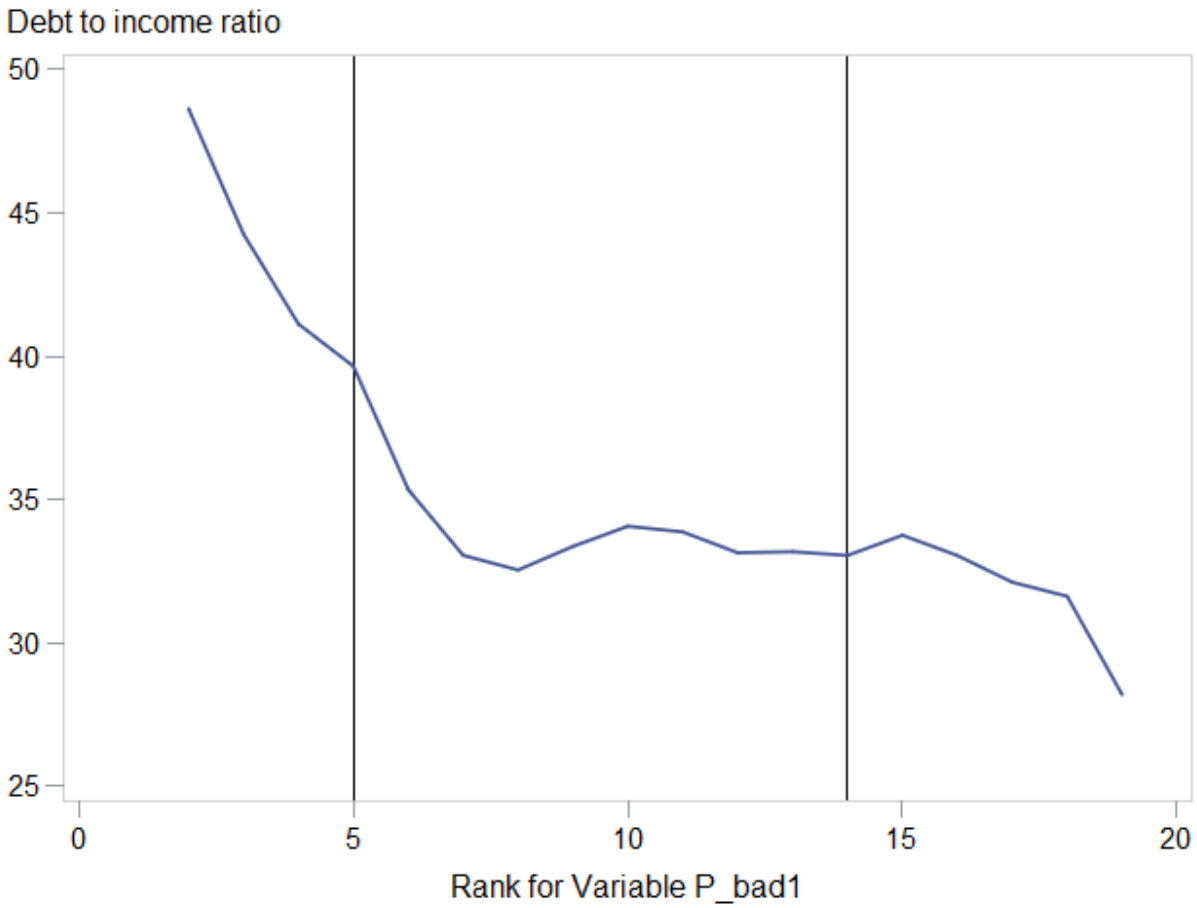

Figure 4. Predictive profile plot for "Debt to income ratio".

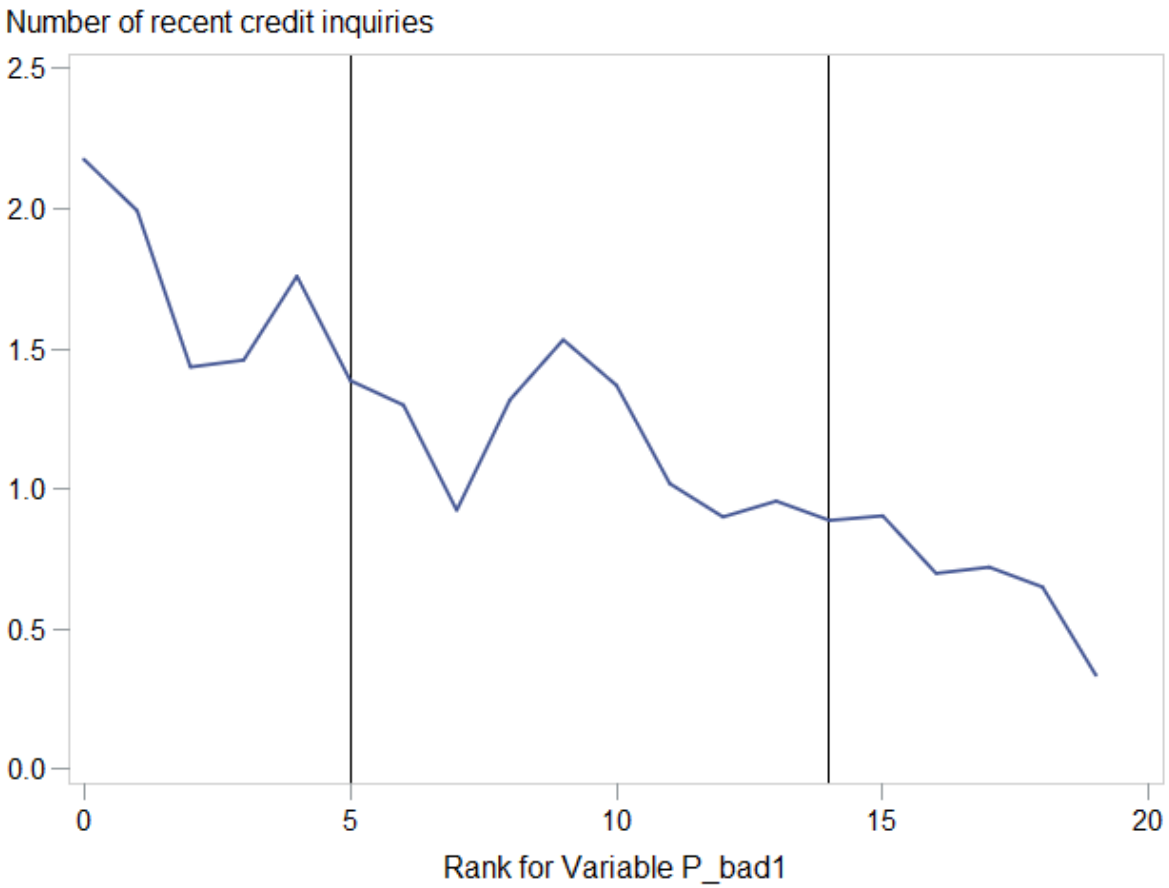

Figure 5. Predictive profile plot for "Number of recent credit inquiries",

CONCLUSION

Optimizing predictive models' performance has been a traditional focus for data scientists. This has led to boosting models' accuracy and robustness. On the other hand, the interpretation of model inputs has not attracted much attention despite the fact that often these inputs pave the way for a business to gain insights into the nature of model predictions. In this paper, we attempt to provide a methodology for bridging this gap. The suggested input-ranking and predicted profile plots allow for selecting important variables and visualizing their relationship with models' predictions. This give an insight into the "drivers" of model predictions. Visualizations of these drivers is an added benefit to help understand the relationship with model's predicted values. Our approach may also help interpret the other variables that were note selected for the model.

More work is needed to understand the statistical aspect of the proposed method of correlating inputs with the predicted values. There is an analogy between our predictive plots and the multiple regression of plots for graphing predicted values (or residuals) against the independent variables. It would also be fruitful to better understand the effect of averaging predictions and input in the ranks on reducing their variance which, in turn, allows for more stable extraction of the significant relationships in the data. Quantifying such effects will help create more effective interpretations and successful deployments of models in business practice.

REFERENCES

Austin P. et al (2014), Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers, *Stat Med.* 2014 Feb 10; 33(3): 517–535.

Breiman, L. 2001 Statistical modeling: The two cultures. *Statistical Science*, Vol. 16, 199–215.

Shmueli G., 2010, To Explain or To Predict? *Statistical Science,* Vol. 25, No. 3, 2010, 289–310