Unfold Income Myth: Revolution in Income Models with Advanced Machine Learning Techniques for Better Accuracy

ABSTRACT

Consumer IncomeView is the Equifax next-gen income estimation model that estimates a consumer's individual annual wage income. This model provides an optimal income solution to our clients by incorporating premier multi-source data assets and advanced machine learning modeling techniques. As a result, the new model significantly improves the scorable population rate compared with the older models and significantly enhances the prediction accuracy. The results of Consumer IncomeView have been successfully validated by an array of new proprietary accuracy metrics for model performance measurement, both on internal out-of-time and clients' data. This paper is to describe the design, development and main results of the model.

INTRODUCTION

While huge strides have been made in the management of delinquency and bankruptcy risk, a significant gap remains in the marketing and risk management toolkit. Understanding a consumer's income level strengthens customer relationships across the entire account lifecycle. Many existing products use unverified income data that has been self-reported through consumer surveys or in government censuses, which generally did not yield a satisfactory result. A reasonably accurate modeled income estimation is urgently demanded to meet market needs. Such product will mainly be used by financial industry (FI) and non-FI sectors for various risk-based and account management programs such as prescreen criteria, credit-line assignments,

cross-sell/up-sell of products, improved segmentation and targeting, ability to pay and debt management. However, Equifax Consumer IncomeView cannot be used for adverse action or risk-based pricing decisions. Ten years ago, Equifax launched the first generation of the Personal Income Model (PIM) to the U.S. market, which was a huge success and provided in-depth income insights to identify the best places to deploy key resources. However, the model performance deteriorates over time and a new model that incorporates novel data assets and state-of-the-art machine learning techniques was required. Recently, Equifax developed and implemented next-gen income model for U.S. market, Equifax Consumer IncomeView which was designed to estimate a consumer's annual wage income at an individual level (income value of 20-300 corresponding to an estimated annual income in thousands). It enables customers to better target their product offerings to consumers who more closely match credit requirements of the offer. This paper presents the design, methodology and results obtained from the development of the Equifax Consumer IncomeView model.

METHODOLOGY

1. Modeling Data: Sources and Validation

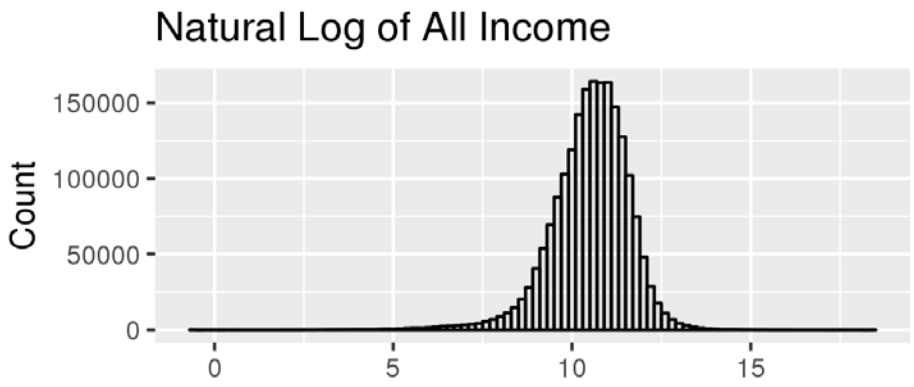
Equifax Workforce Solutions (WFS) is a subsidiary of Equifax, Inc. that provides employment and income verifications for over 4,600 U.S. employers, including over 75% of the Fortune 500 and Fortune 1000 companies maintained in The Work Number® (TWN) database. Employment and income information is provided directly by participating employers and is updated every time their employees are paid. The individual annualized salary/wage incomes were derived directly from this database, which was used as the dependent variable. Necessary exclusion steps were performed to remove the inactive records from subsequent analysis. The remaining

active employment data was further merged with Equifax proprietary consumer credit database, Equifax Automated Credit Reporting Online (ACRO), including the Equifax Advanced Decisioning Attributes consumer credit file attributes (ADA), Equifax Regulation Z ability to pay attributes (RegZ), Equifax Mortgage Consumer Modeling Attributes (MCMA), and Equifax Dimensions trended credit attributes. The overlapped population between The TWN and ACRO were used as the model population. Standard exclusion criteria were then applied to this data to further remove the invalid records, such as duplicates, deceased, fraud etc. Approximately 20MM observation points were used for the model development, and out-of-time samples of similar size were used for model evaluation. Model data was further randomly split into development and in-time validation.

2. Dependent Variable

TWN verified annualized income in dollar amount, i.e. annualized individual salary/wage before tax, was used as the dependent variable. Notable data exclusion filters were used: individuals with outlier incomes that could not be validated were excluded, as well as consumers that were retired, restricted, deceased, surviving spouse and records that had recently been added to the TWN database without payroll history. The consumers with no credit activity within 24 months, or outdated, or who were identified as deceased were excluded. Inquiry-only records were also excluded from the sample. Multiple transformation schemes, such as box-cox power transformation, have been explored for the dependent variable and an internal research shows that log transformation provided the best in-sample fit. Due to the nature of income distribution, the assumption that earnings are log-normally distributed is widely accepted.

Figure 1: Log-Transformation of the Target Variable for the 1% ACRO Sample Population



3. Independent Variables

Model attributes were sourced from various proprietary Equifax consumer credit data attributes assets, including: ADA, RegZ, MCMA and Dimensions.

Prior to model parametrization, the following variable treatment and selection steps were observed:

- Apply standard data cleansing procedures to the sourced data;
- Missing value imputation and capping/flooring; and
- Perform exploratory data analysis to understand the stability and predictive power of each attributes.

In addition, care was taken to avoid high correlation among the independent variables, which can affect model stability. Measures such as coefficient of correlation and variance inflation factor

were used to assess the relationship between the independent variables. In some cases a variable could have a strong association with the dependent variable, but the relationship may be non-linear. To address this situation, analytical team(s) used standard proprietary methodologies to accomplish the following:

- Determine the necessity of transformation for each of the independent variables;
- Determine the optimal method for the variable transformation; and
- Enable transformed variables to be included as independent variables in model development.

Other proprietary variable treatment included additional variable creation and the use of interaction terms.

As mentioned, all independent variables and their missing indicators were initially considered as independent variables. Together, these variables were run through computer-aided variable selection or reduction procedures in order to narrow down the candidate set of variables into a smaller, more manageable list. The variable list was then further refined through several more iterations to ensure that the model worked from both a statistical and business standpoint. In addition, from a statistical point of view, highly correlated variables were eliminated from subsequent regressions. Finally, variables were tested one at a time to determine the best possible combination of predictive variables.

4. Segmentation

The purpose of segmentation analysis is to determine the possibility, as well as the necessity, of defining homogeneous segments or subgroups in the population that require separate models. If

such groupings can be identified, it may be deemed necessary to build separate models for these groups to enhance the overall performance.

Decision Tree was used to select the optimal segmentation scheme and splits (Table 1). The size, significance, complexity and interpretation of various segmentation themes were evaluated to finalize the final segmentation. More than ten different scenarios of segmentation were studied, the goal was to separate the most accurate to least accurate group to provide different confidence level for income estimation on each segment. The best scheme is to use two layers of decision trees, by using different target variables. Table 1 summarized the final four segments generated from this scheme.

Table 1 summarizes the 4 segments generated from the scheme.

Segment Number	Description	Equifax ACRO Attributes
1	Low Income	• Age of Trade • Consumer Credit Capacity • Available Credits on Revolving Accounts
2	Medium - Low Income	
3	Medium –High Income	
4	High Income	

5. Modeling Methods

Over 120 different machine learning modeling techniques were explored to find the best modeling approaches, with two goals in mind: optimizing both the model prediction accuracy and interpretability. The final product is a combination of three methods; different method is used for different segment: linear regression (Ordinary Least Square baseline model), Multivariate Adaptive Regression Splines introduced by Friedman (1) and deep learning –

multiple layers Neural Network (2). The performance of each model was evaluated and compared by using various proprietary accuracy metrics innovated internally.

Multiple Linear Regression

Multiple linear regression is a proven successful modeling technique designed to model the relationship between a continuous dependent variable y and one or more explanatory variables (or independent variables) denoted X, i.e.

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots n$$

where y_i is the Log-transformed annualized individual income in dollar amount, and $x_{i1}, x_{i2} \ldots x_{ip}$ are the Equifax proprietary consumer credit attributes as independent variables. Statistical estimation and inference in linear regression focuses on the coefficients βs. The elements of this parameter vector are interpreted as the partial derivatives of the dependent variable with respect to the various independent variables. The linear regression model uses the explanatory variables to estimate the transformed dependent variable takes on a given value, and then transform it back to dollar value. Model is evaluated on statistically significance on R-square, P-value, etc., and multicollinearity issues among the independent variables were also properly handled.

To select the significant independent variables and prevent model overfitting, LASSO (least absolute shrinkage and selection operator) technique (3) is used, which arises from a constrained form of ordinary least squares regression where the sum of the absolute values of the regression coefficients is constrained to be smaller than a specified parameter. More precisely for a given parametert, the LASSO regression coefficients $\beta = (\beta_1, \beta_2 \ldots \beta_m)$ are the solution to the constrained optimization problem

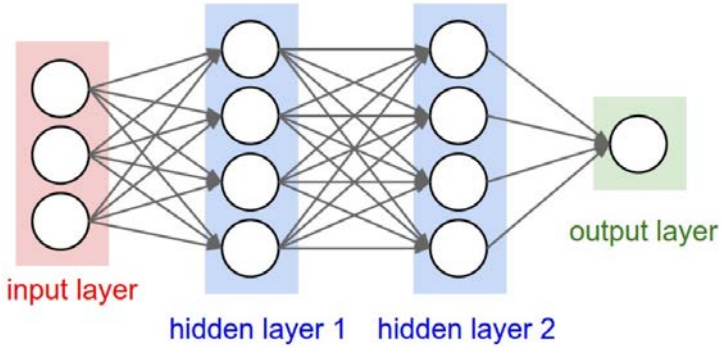
$$minimize||y - X\beta||^2 \quad subject\ to \sum_{j=1}^{m} |\beta_j| \leq t$$

Provided that the LASSO parameter t is small enough, some of the regression coefficients will be exactly zero. Hence, the LASSO can be used as selecting a subset of the regression coefficients for each LASSO parameter. By increasing the LASSO parameter in discrete steps, a sequence of regression coefficients are obtained, where the nonzero coefficients at each step correspond to selected parameters. The algorithm that core modeling team implemented in SAS is a stepwise like procedure with a single addition to, or deletion, from the set of nonzero regression coefficients at any step.

Neural network

A neural network is a series of algorithms, which assemble many "neurons," and output the prediction neuron, as shown in figure 4 below. The leftmost layer of the network is called the input layer, and the rightmost layer is called the output layer, the middle layers of nodes are the hidden layers. In this model, we chose to use two hidden layers to optimize the model prediction accuracy preventing the overfitting issue.

Figure 4: Neural Network Model Configuration



The input neurons take the input x_1, x_2, x_3, x_4, and an intercept "(+1)" term or bias units, then output

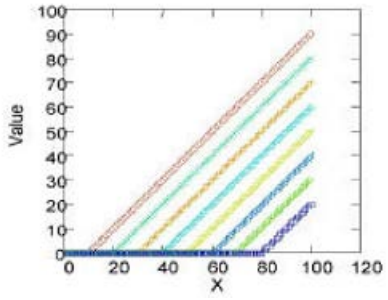
$$h_{W,b}(x) = f(W^T x) = f(\sum_{i=1}^{3} W_i x_i + b),$$

Where f (.) is the activation function, parameters W is weight function. In the model, the hidden layers used hyperbolic tangent activation function. Model is evaluated on statistically significance on errors, such as % average training error, % average error, and % maximum error etc. To prevent the stability of the optimization process, which means, prevent being trapped at a local max, the model was fit by using 5 random starting points (assigning weight) with maximum iteration = 1000. Research have been conducted to increase the number of the starting points, however model accuracy is not improved, and model is not better, while the computation time is significantly prolonged.

Multiple Adaptive Regression

Multiple adaptive regressions constructs nested "hockey-stick" spline basis functions in an adaptive way by automatically selecting appropriate knot values for different variables, and it obtains reduced models by applying model selection techniques. The method does not assume parametric model forms and does not require specification of knot values. The bases are constructed by using truncated power functions (hockey stick function) as follow:

$$(x - s)_+ = \max(x - s, 0)\,, \qquad s\ is\ one\ of\ the\ knots\ for\ x$$



The final income prediction $\hat{y}$ is a piecewise linear combination all bases:

$$\hat{y} = \sum_{i=1}^{k} \beta_i Basis_i$$

Similar to the forward selection in linear regression model, pairs of corresponding basis functions were selected and added to the model. The pair that resulted in the largest reduction in the residual sum of square was added. The next phase was backward elimination of a single basis function whose elimination minimizes the generalized cross validation criterion (GCV), a function of the residual sum of squares. Backward elimination iterates until all terms except the intercept are eliminated and then the model with the minimum GCV was chosen. ADAPTIVEREG procedure is used to fit the final model. Like other nonparametric non-linear regression procedures, the ADAPTIVEREG algorithm can yield complicated models that involve high-order interactions in which many knot values or subsets are considered. Besides the basis functions, both the forward selection and backward selection processes are also highly nonlinear. Because of the trade-off between bias and variance, the complicated models that contain many parameters tend to have low bias but high variance. To select models that achieve good prediction performance, GCV was used:

$$\mathrm{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}_i}{1 - \mathrm{trace}(\mathbf{S})/n}\right)^2 = \frac{\mathrm{RSS}}{n(1 - \mathrm{trace}(\mathbf{S})/n)^2}$$

where y is the response, $\hat{f}$ is an estimate of the underlying smooth function, and S is the smoothing matrix. The effective degrees of freedom for the smoothing spline can be defined as the trace of S. Friedman uses a similar quantity as the lack-of-fit criterion,

$$\mathrm{LOF} = \frac{\mathrm{RSS}}{n(1 - (M + d(M-1)/2)/n)^2}$$

where d is the degrees-of-freedom cost for each nonlinear basis function and M is total number of linearly independent bases in the model.

Based on the accuracy performance on each segment as well as the whole model, we finally came up with the following model strategy for Consumer IncomeView, which listed in the Table 2.

Table 2: Modeling approaches for each segment.

Segment	%Population	Equifax ACRO Attributes
Seg1: Low Income	10%	Linear Regression with LASSO
Seg2: Low-Medium Income	64%	Neural Network: 2-hidden layer 10 nodes & 5 nodes
Seg3: Medium-High Income	21%	Neural Network: 2-hidden layer 10 nodes & 5 nodes
Seg4: High Income	6%	Multiple Adaptive Regression

RESULTS

Performance Metrics

To assess the performance of Consumer IncomeView^{TM}, Equifax examined the accuracy of the predicted income by the following innovative accuracy metrics: Windowed Percent Error (WPE), One-tail Accuracy, Capture Rate and Classification metric. These metrics were designed and implemented primarily for business applications.

1. **WPE:** measures percentage of predicted income falls into ±x% of actual income values. This accuracy is pinpoint accuracy measurement. WPE-20 (i.e. ±20% of actual income) is commonly used in the industry.
2. **One-tail accuracy**: One tail (cut-off threshold measure) accuracy is innovated for business application:
    - It measures how accurate the model estimates a consumer's income higher than $x.

- For example, if the model estimate a consumer's income is >$60k (say point estimate=62k), it is 79% accurate that his/her true income is also >$60k, which is the worst case scenario

3. **Capture rate**: is used to support one-tail accuracy to provide more comprehensive view in the following:
    - If consumers true income is higher than $x, what percent of the predicted income is higher than $x
    - Evaluate what percent of the true income the model can correctly capture directionally

    One-tail accuracy and capture rate are combined accuracy measurement, they should be considered together as one measurement criteria.
4. **Classification metrics**: measures the effect of correctly predicted income for both upward and downward one tail accuracy. It provides a comprehensive view of how accurate the model can predict correctly on the directional for both tails.
5. **Concordance**: Concordance is nonparametric, which measures "rank-ordering" properties of a statistical model. Concordance computation logic is below (calculation is completed for each record pair within a statistically significant sample of records):

    If income1 < income2 and predicted1 < predicted2, then concordant;

    If income1 > income2 and predicted1 > predicted2, then concordant;

    Otherwise, discordant.

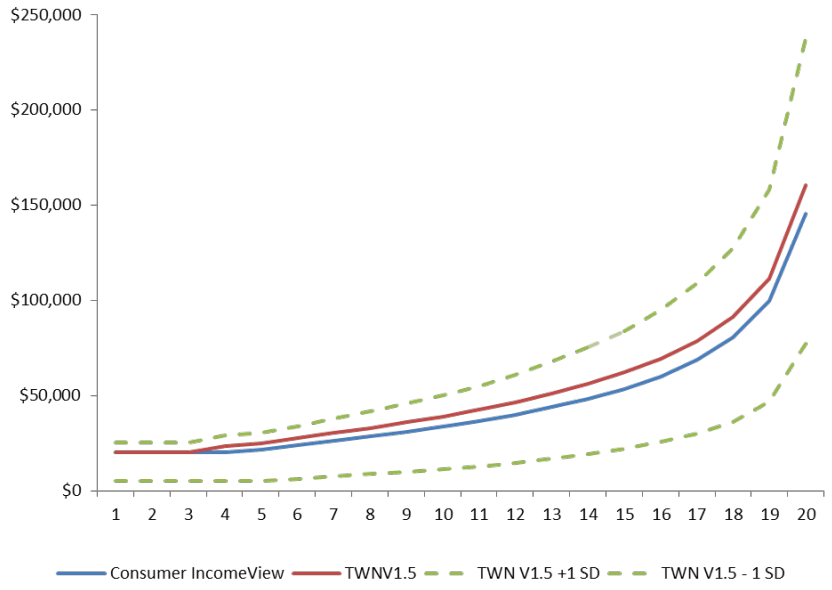
The final concordance measure is expressed as a percentage of correctly ranking pairs of income records, e.g. if the concordance statistic is 70.3%, then 70.3% of the pairs of predictions were rank-ordered correctly.

Model Results

Income Distribution

Consumer IncomeView outputs individual income scores in the range from 20-300 in the unit of one thousand dollars. Based on the out-of-time 2016-Dec validation samples, Figure 5 compares the distributions of predicted income vs. the actual income in vingtiles. The median incomes estimated by Consumer IncomeView correspond very closely with the median of actual incomes.

Figure 5 Predicted vs Actual Income distribution comparison.



Accuracy Performance

When compared with the older Equifax income model PIM3, Figure 6 shows that Consumer IncomeView provides 31% incremental WPE-20 accuracy lift (Table 2) and 10% incremental scorable rate lift (Figure 6).

Table 2 Model performance on WPE-20 by segments

Details	Low Income	Medium - Low Income	Medium - High Income	High Income
WPE20	67.0%	38.7%	29.0%	29.80%

Figure 6 Overall Model Results

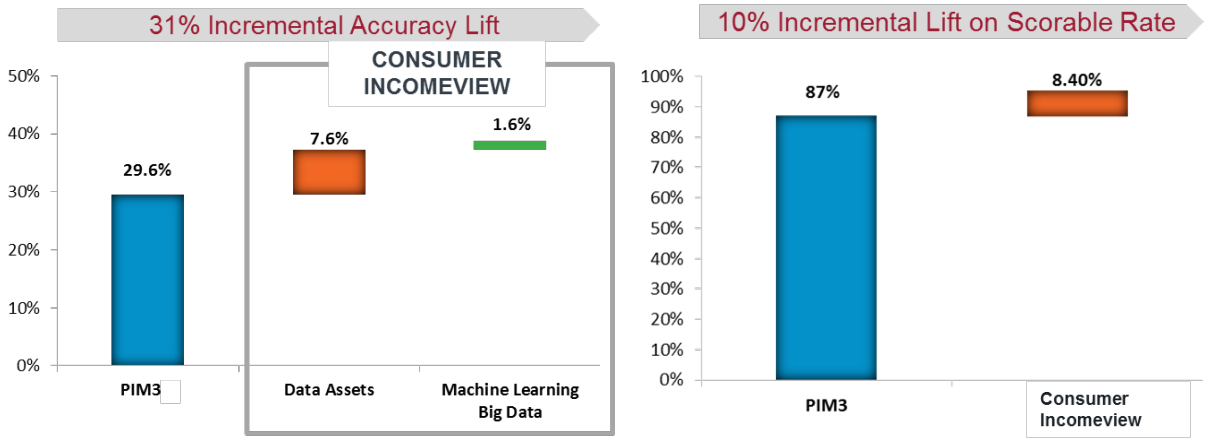
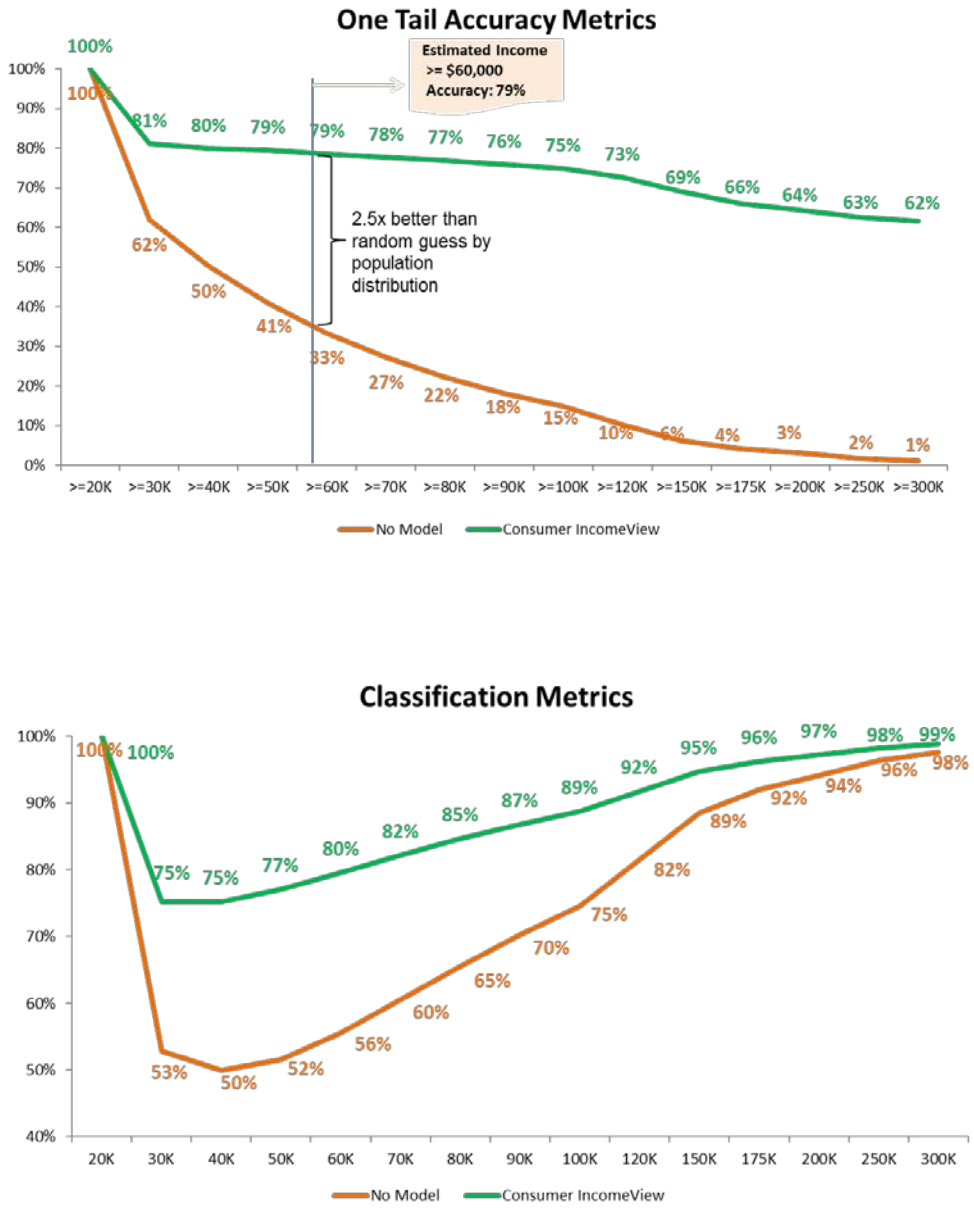
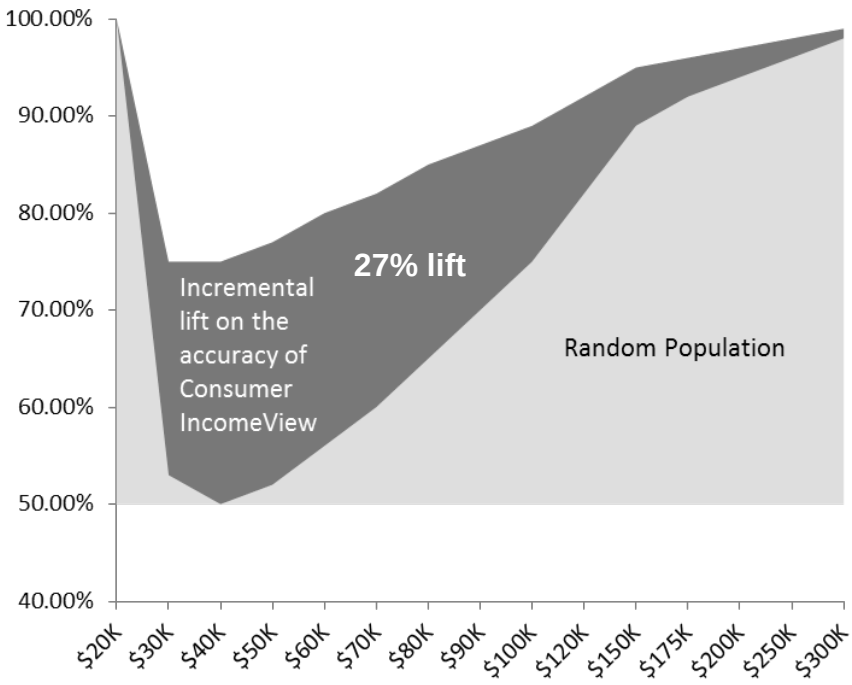


Figure 7 shows the One-tail and Classification Accuracy of the Consumer IncomeView. Compared with the general U.S. population distribution, the new solution significantly improves the One-tail (upwards) and classification accuracy. For instance, the One-tail has a ~2.5 times higher accurate rate on the threshold of $60K. It means that if a consumer has a $60k+ modeled income, there is ~79% of the chance that that person's actual income is indeed $60k+, which is much higher than the general population rate of ~33%. Consumer IncomeView also achieves a

minimum 75% classification accuracy, for example, for income=60k, the Consumer IncomeView can predict correctly 80% for consumers' income directionally, either >=60k or <60k, and significantly improved from general population without model.

Figure 7 One-tail and Classification Accuracy





Furthermore, we can calculate the area under the two classification accuracy curves (random population vs Consumer IncomeView) and obtain the overall classification accuracy over all income range from $20K -- $300K. Compared with the benchmark random U.S. population (without a predictive model), the Consumer IncomeView has ~27% incremental lift on the overall classification accuracy over all predicted income ranges.

Finally, concordance statistics of the Consumer IncomeView is shown in Table 3. When or concern focuses on the overall rank-ordering rather than individual income estimate, the nonparametric concordance metrics can assess the overall model performance. Compared with the older Equifax PIM3 model, Consumer IncomeView™ significantly improved the concordance statistics, from 67.7% to 71.2%, generating a 5.2% incremental lift.

Table 3 Concordance Statistics

Details	PIM3	Consumer IncomeView™	Lift
Concordance	67.70%	71.20%	5.2%

DISCUSSION

In this paper we describe the design and development of the Equifax Consumer IncomeView model. This enhanced solution was built on the enriched Equifax proprietary consumer credit attributes, including the powerful newly developed trended credit attributes, featuring premier monthly consumer credit data up to 24 months of extended financial account history. Compared with the older Equifax PIM3 model, the Consumer IncomeView significantly improves the overall WPE-20 and expands the scorable population. When measured by the innovative One-tail and classification metrics, this new model also outperforms the older PIM3 model by a big margin. It is quite reasonable to believe that the performance improvement is due to incorporating both highly predictive attributes and machine learning techniques.

Consumer IncomeView has also been validated by both in-time validation and out-of-time validation. Table 4 shows the out-of-time validation. Segmentation distribution is almost the same as the model development sample, i.e. 2-layers segmentation scheme validation holds well. WPE-20 has solid validation on out of time data, both segmentally and overall, and one-tail and classification accuracy on out of time validation hold very well (not shown).

REFERENCE

1. Friedman JH. Multivariate Adaptive Regression Splines. The Annuals of Statistics. 1991, 1-67.
2. Bishop CM. Neural Networks for Pattern Recognition. Oxford: Oxford University Press. 1995.
3. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B. 1996, 58:267-288.