Estimating Family Income from Administrative Banking Data: A Machine Learning Approach

By DIANA FARRELL, FIONA GREIG, AND ERICA DEADMAN

* JPMorgan Chase Institute, 601 Pennsylvania Ave NW, 6th FL, Washington, DC 20004 (institute@jpmchase.com).

The JPMorgan Chase Institute aims to use administrative banking data to publish insights that are representative of the US population. To do this, we require a method to reweight our samples of Chase customers to reflect key characteristics of the nation, with income foremost among them. Given that we do not have full coverage of income information across our portfolio of customers, we have developed a proof-of-concept method for estimating income.

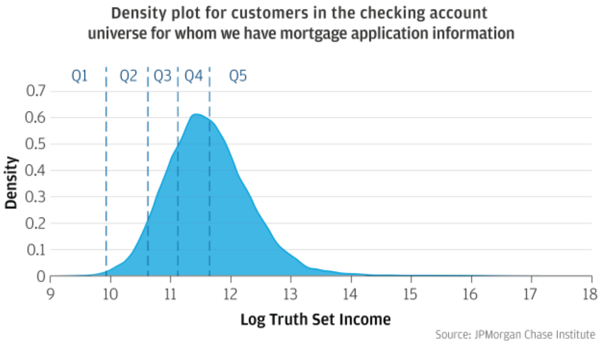
JPMC Institute Income Estimate (JPMC IIE) version 1.0 uses gradient boosting machines (GBM) to estimate gross family income based on a truth set drawn from credit card and mortgage application data. The estimation relies on administrative banking data – such as checking account inflows – in combination with ZIP code-level characteristics available through public datasets, as well as Census data at the tract level. Deposit account inflows alone are insufficient to approximate gross family income. The combination of administrative banking data with other data sources and a machine learning approach yielded a significantly more accurate prediction of income.

I. Data

The goal of JPMC IIE is to predict gross family income of Chase checking account customers each year from 2013 to 2017. We aggregate data to the primary account holder level and restrict the prediction exercise to customer-months that have sufficient checking account activity to establish a relationship with the bank. In addition to administrative data from families' banking activities, we also leverage ZIP code level attributes from publicly available data.

A. Dependent Variable

We use two sources of ground truth income for modeling: income obtained from mortgage and credit card applications. Each source has its strengths and weaknesses. Income obtained for mortgage applications undergoes a verification process and is therefore an accurate representation of a family's gross income. However, the set of families applying for mortgages tend to be more affluent than those that do not. As shown in Figure 1, our mortgage income data skews toward high-income families, with close to 50 percent of the sample falling into the top income quintile as defined by American Community Survey (ACS) income data, and only 6 percent in the first two quintiles.

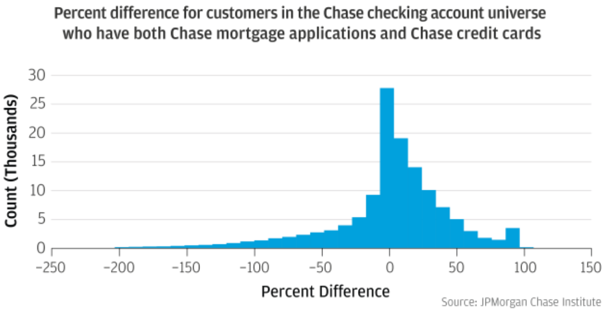

Density plot for customers in the checking account universe for whom we have mortgage application information

Source: JPMorgan Chase Institute

FIGURE 1. DENSITY PLOT OF MORTGAGE VERIFIED INCOME

On the other hand, while credit card applicants cover a broader range of incomes, data obtained through that application process are self-reported and may be less accurate than mortgage-verified incomes. Comparing the two income values for customers for whom we had both verified income from a mortgage application and stated income from a credit card application revealed that customers tended to state more income on their credit card applications than was verified on their mortgage applications. The median percentage difference of credit card stated income minus mortgage verified income was positive among customers who applied for both products within the same year (Figure 2).

Differences may represent income from unverifiable sources, such as cash, or real income changes during the year. To avoid losing this information, we average the two sources of income when both are present. Finally, we create our modeling truth set by log-transforming the dependent variable to address positive skew in income distribution.


Percent difference for customers in the Chase checking account universe who have both Chase mortgage applications and Chase credit cards

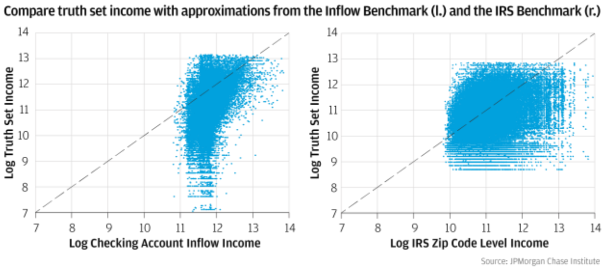
Source: JPMorgan Chase Institute

FIGURE 2. DISTRIBUTION OF STATED INCOME MINUS MORTGAGE VERIFIED INCOME

B. Benchmark Income Measures

To better understand the incremental value of a machine learning approach, we constructed two naïve approximations of income: (1) the Inflow Benchmark sums checking account inflows that we categorize as income, adjusted to approximate pre-tax income; (2) the IRS Benchmark uses ZIP code level average IRS-reported income to proxy income for each individual based on their reported ZIP code.

The relationship between the benchmark measures and our truth set is presented in Figure 3. Both benchmarks yield high mean absolute error (MAE) values: 162 percent for the Inflow Benchmark and 103 percent for the IRS Benchmark. These high error rates confirm the need for a more comprehensive approach to income estimation and will be used for comparison as we develop our machine learning approach.


Compare truth set income with approximations from the Inflow Benchmark (l.) and the IRS Benchmark (r.)

Source: JPMorgan Chase Institute

FIGURE 3. PERFORMANCE OF BENCHMARK INCOME MEASURES

B. Independent Variables

The features used to predict income originate from sources both internal and external to the bank. Internally, we include four main groups of features: (1) Customer information, such as age and location; (2) Checking account attributes, including inflow categorizations and account balances; (3) Credit card attributes, such as credit limit and number of cards; (4) Attributes of other accounts, including loan information and total liquid assets across deposit accounts.

We also use features from ZIP code level characteristics available through public datasets, such as the Internal Revenue Service (IRS) Statistics of Income (SOI) dataset, and Zillow rental information, as well as Census data at the tract level.

We aggregate account features at the annual level, capturing the maximum, minimum, average, range, and total of each feature within the calendar year. In total this yields 400 raw candidate features for model training per year, which we then treat to remove missing values, handle skewed distributions, and standardize ranges.

C. Sample Construction

Our modeling sample includes customers for whom we have information on either mortgage income or credit card income at some point during 2013 through 2017.

We perform three steps to improve upon the accuracy and representativeness of our final income truth set. Because checking account inflows represent take-home income after taxes and other deductions and may not represent the customer’s overall income, we expect true income to always be greater than income inflows. To address accuracy, we remove from our sample customers whose truth set income is less than income inflows into their checking account. In our remaining sample, we remove customers with income in the top or bottom percentile of the truth set in order to train the model without undue influence of extreme observations. Finally, to address sample representativeness, we stratify the sample by ACS quintiles, selecting 50,000 customers from each. This yields a final modeling sample of 250,000 customers each year.

II. Modeling Approach

We embarked on this proof-of-concept effort to assess the feasibility and use of a gross income estimate derived via machine learning. We initially considered several modeling techniques in order to identify which is best aligned with our goals. Most weight was given to finding a method that is performant in quick iterations as we build toward a minimum viable product (MVP) estimate.

Our set of candidate techniques included gradient boosting machines, random forests, elastic net linear regression, and support vector regressors. After initial runs, we selected gradient boosting machines (GBM) as the approach best suited to our MVP project goals.

A. Gradient Boosting Machines (GBM)

GBM is a machine learning algorithm based on ensembles of decision trees. The GBM algorithm generates a sequential series of weak learners (shallow trees), iteratively improving the estimate with each new tree.

GBM, like other tree-based methods, is capable of fitting relationships with no requirements around the underlying functional forms. This frees the modeler from detecting and specifying nonlinearities in variable relationships, as the algorithm seamlessly captures complex curvatures and interactions.

However, the flexibility that makes GBM so appealing also makes it prone to over-specification, wherein the model is fit so well to the particular idiosyncrasies of the training data that it fails to generalize out of sample. In other words, GBM may generate a model that yields impressive performance metrics during development, but very poor metrics when applying the model to new data – the very goal of developing an estimate.

We prevent over-fitting through hyper-parameter tuning, selecting the preferred level of model complexity to ensure model generalizability. Our GBM tuning focused on four hyperparameters: number of estimators (trees) generated; maximum depth of each tree; maximum features to consider for each split within a tree; and minimum sample size needed to allow a further split.

B. GBM vs. Regression-Based Methods

There are two primary critiques of GBM relative to traditional parametric techniques, such as linear regression. The first is that tree-based methods come with a high risk of over-specification. While true, that risk can be reliably managed with standard approaches, including hyperparameter tuning and deliberate separation of development data. We describe our management of these processes in sections IIA and IIIA, respectively.

The second common critique is that modeled relationships are less interpretable with GBM than regression-based approaches. Without an easily readable scoring equation, it is difficult to understand the contribution of individual inputs to the final estimate. However, there are other techniques available for understanding the relationships captured by the model. For example, partial dependence plots (PDPs) can be used to graphically represent the marginal effect of each input on the predicted outcome.

III. Performance Assessments

The purpose of JPMC IIE is to provide an estimate of gross family income that we can use to segment and reweight populations by income quintile. Thus, in optimizing the performance of JPMC IIE, we aim to minimize MAE of the point estimates and also to predict the correct income quintile accurately. Here we present results for version 1.0 of JPMC IIE.

A. Data Approach

In order to train a model that is generalizable to different populations of our research universe, we separated our modeling data into three groups:

1. Training Set (60 percent of sample): Used to fit the models in order to determine the form of the relationship between income and the feature set.

2. Validation Set (20 percent of sample): Used in parallel with the training set, to tune hyperparameters and guard against overfit.

3. Testing Set (20 percent of sample): Used to assess the predictive power of the final model, on observations not used for training or hyperparameter tuning.

B. Results

We focus our attention on MAE and quintile prediction accuracy: the proportion of each predicted income quintile classified correctly (e.g., belonging to the same truth set income quintile), based on ACS quintile boundaries. By all metrics, results are fairly consistent across years in our testing set, yielding an average MAE of 41 percent and an accurate quintile prediction 55 percent of the time. We also observe small differences between the MAE on the training and testing sets (38 percent and 41 percent, respectively) indicating that the estimates are not overfitting to the training sample.

As most of JPMCI's research uses income on the ACS quintile basis, we prioritized consistent accuracy across those quintiles. Figure 4 shows classification across truth set income quintiles, within each predicted income quintile. We observe that the accuracy rate is consistent across predicted quintiles, and misclassifications tend to be concentrated in the adjacent quintiles: 90 percent or more of the observations were classified in the correct or an adjacent income quintile.

	Q1 TRUE	Q2 TRUE	Q3 TRUE	Q4 TRUE	Q5 TRUE
Q1 Predicted	56.8%	32.1%	7.8%	2.5%	0.8%
Q2 Predicted	9.3%	55.3%	29.3%	5.6%	0.5%
Q3 Predicted	1.1%	17.2%	53.3%	26.0%	2.4%
Q4 Predicted	0.3%	4.0%	24.4%	56.4%	14.9%
Q5 Predicted	0.1%	1.0%	7.8%	35.4%	55.8%

Source: JPMorgan Chase Institute

FIGURE 4. ACS QUINTILE ACCURACY BY PREDICTED QUINTILES

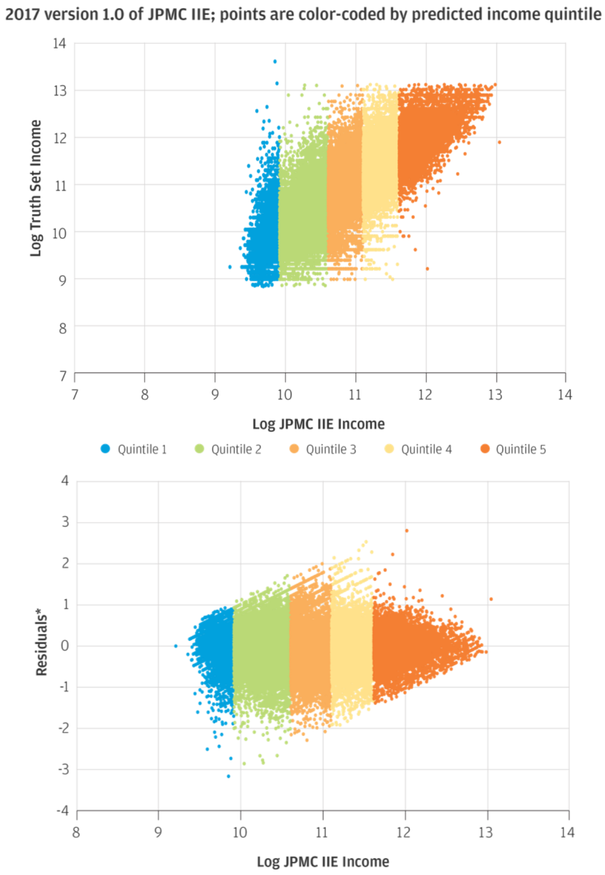
The consistent accuracy across quintiles is the result of stratifying our modeling data by ACS income quintile (described in section IC). In contrast, a model trained on a random, un-stratified sample underperforms for lower income quintiles while performing better for middle and high income quintiles (Figure 5). This is because our un-stratified training set naturally over-represents families in the third and fourth income quintiles.

Although accuracy rates within a predicted quintile are consistent across quintiles, the model exhibits asymmetric errors when assessed at a more granular level. Figure 6 shows that low predicted income values are skewed slightly toward underpredicting their corresponding truth set income.

QUINTILE ACCURACY	FINAL MODEL	UN-STRAT. MODEL	DIFFERENCE
Q1 Predict	56.8%	28.5%	-28.3%
Q2 Predict	55.3%	49.0%	-6.3%
Q3 Predict	53.3%	73.2%	19.9%
Q4 Predict	56.4%	67.5%	11.1%
Q5 Predict	55.8%	59.8%	4.0%

Source: JPMorgan Chase Institute

FIGURE 5. ACS QUINTILE ACCURACY BY STRATIFICATION OPTIONS



*Residuals represent truth set minus estimate

Source: JPMorgan Chase Institute

FIGURE 6. ESTIMATED INCOME VS. TRUTH SET INCOME, AND CORRESPONDING RESIDUALS

C. Case Study: IIE version 1.0 Research Use

We test the estimate's ability to reweight the sample for the JPMCI Healthcare Out-of-pocket Spending Panel (HOSP). Figure 7 compares out-of-pocket health spending levels, reweighting our HOSP sample to match each state's joint age and income distribution using age and truth set income (orange line). Applying JPMC IIE and age (black line) yields estimated spend levels much closer to this true weighting than either an unweighted sample (blue line) or a sample weighted exclusively by age (green line). We conclude that the use of JPMC IIE is valuable in weighting the HOSP sample to more closely represent each state population by age and income.

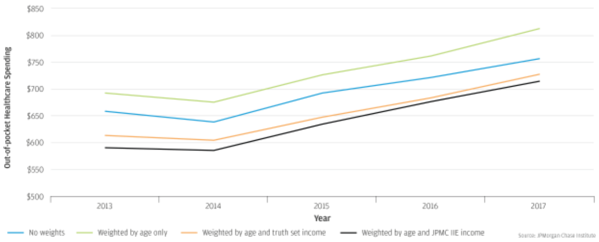


FIGURE 7. OUT-OF-POCKET HEALTH SPENDING ACROSS YEARS, BY DIFFERENT WEIGHTING SCHEMES

IV. Directions for Future Improvement

With a validated version 1.0 of JPMC IIE in place, there is room for enhancement and expansion of scope. Due to the proof-of-concept nature of version 1.0, certain avenues of exploration were put on hold for future iterations, including: data expansion, feature refinement, and insight exploration.

Data Expansion: Version 1.0 of JPMC IIE relies on the de-identified Chase checking account universe as its base population. It cannot be applied to credit-only customers who do not have a Chase checking account, limiting the research projects where it can be of use. JPMC IIE could be expanded to enable estimation of credit-only customers.

Feature Refinement: We spent minimal time on feature engineering for version 1.0, but thorough data exploration and treatment is critical for establishing a well-performing model. Thoughtful assessment of feature aggregation may yield powerful predictors that the gradient boosting algorithm cannot easily approximate, such as ratios, trends over time, or other functions of multiple features.

Insight Exploration: Future assessments could yield a deeper understanding of the relationships captured by the model, and how individual features impact income predictions. Beyond understanding feature relationships, the model could be explored holistically to gain perspective on areas of caution through two approaches: demographic monitoring to assess whether modeled income exacerbates demographic biases; and residual monitoring, to assess the model for systematic weaknesses.

We look forward to continued exploration in this space, iterating toward additional insights of value to the broader academic, policy, and data science communities.

REFERENCES

Farrell, Diana and Fiona Greig. "On the Rise: Out-of-Pocket Healthcare Spending in 2017." JPMorgan Chase Institute, 2018.