DEC 2018

Estimating Family Income from Administrative Banking Data

A Machine Learning Approach

JPMorgan Chase & Co.
INSTITUTE

Abstract

At the JPMorgan Chase Institute, we aim to publish generalizable insights that are representative of the overall US population. To do this, we require a method to reweight research based on key characteristics, with income foremost among them. Given that we do not have full coverage of income information across our portfolio of customers, we set out to develop a reliable method for estimating income.

Using machine learning techniques, we trained an estimate of gross family income based on a truth set drawn from credit card and mortgage application data. JPMC Institute Income Estimate (JPMC IIE) version 1.0 uses gradient boosting machines (GBM) and relies heavily on administrative banking data such as checking account inflows. It predicts income with a mean absolute error (MAE) of 41 percent, outperforming comparative benchmarks, and demonstrates consistent accuracy across predicted income pentiles (average 55 percent). JPMC IIE version 1.0 is currently in use for research purposes, with results similar to truth set income when used for reweighting purposes. Future versions will seek to improve predictive power and expand the use of the estimate.

Acknowledgements

We thank Carlos Grandet and Chuin Siang Bu for their excellent project and technical leadership for this work. We thank Erica Deadman, Amar Hamoudi, and Marvin Ward for their guidance and advice. In addition, we thank Diana Farrell and Fiona Greig for their leadership throughout.

We are also grateful for the contributions of the JPMorgan Chase Institute members who reviewed and helped refine this paper and analysis: Chris Wheat, Chen Zhao, Melissa O'Brien, Chex Yu, James Duguid, Bryan Kim, Beatriz Rache, Max Liebeskind, and Tanya Sonthalia.

We would also like to acknowledge the invaluable comments and advice of research experts including Bob Hunt from the Federal Reserve Bank of Philadelphia's Consumer Finance Institute, as well as Robert Hughes, Sam Leitner, David Low, Eva Nagypal, Akaki Skhirtladze, and Ron Borzekowski at the Consumer Financial Protection Bureau.

This effort would not have been possible without the critical support of our partners from the JPMorgan Chase Consumer & Community Bank and Corporate Technology teams of data experts, including Samuel Assefa, Connie Chen, Anoop Deshpande, Senthilkumar Gurusamy, Ram Mohanraj, Karen Narang, Stella Ng, Rob Rappa, Ashwin Sangtani, Anmol Karnad, and JPMorgan Chase Institute team members including Elizabeth Ellis, Alyssa Flaschner, Courtney Hacker, Sarah Kuehl, Caitlin Legacki, Sruthi Rao, Carla Ricks, Jolie Spiegelman, Gena Stern, and Maggie Tarasovitch.

Finally, we would like to acknowledge Jamie Dimon, CEO of JPMorgan Chase & Co., for his vision and leadership in establishing the Institute and enabling the ongoing research agenda. Along with support from across the Firm—notably from Peter Scher, Max Neukirchen, Joyce Chang, Patrik Ringstroem, Lori Beer, and Judy Miller—the Institute has had the resources and support to pioneer a new approach to contribute to global economic analysis and insight.

About the Institute

The global economy has never been more complex, more interconnected, or faster moving. Yet economists, businesses, nonprofit leaders, and policymakers have lacked access to real-time data and the analytic tools to provide a comprehensive perspective. The results—made painfully clear by the Global Financial Crisis and its aftermath—have been unrealized potential, inequitable growth, and preventable market failures.

The JPMorgan Chase Institute is harnessing the scale and scope of one of the world's leading firms to explain the global economy as it truly exists. Its mission is to help decision-makers—policymakers, businesses, and nonprofit leaders—appreciate the scale, granularity, diversity, and interconnectedness of the global economic system and use better facts, timely data, and thoughtful analysis to make smarter decisions to advance global prosperity. Drawing on JPMorgan Chase's unique proprietary data, expertise, and market access, the Institute develops analyses and insights on the inner workings of the global economy, frames critical problems, and convenes stakeholders and leading thinkers.

The JPMorgan Chase Institute is a global think tank dedicated to delivering data-rich analyses and expert insights for the public good.

Contact

For more information about the JPMorgan Chase Institute or this report, please see our website www.jpmorganchaseinstitute.com or e-mail institute@jpmchase.com.

Introduction

The JPMorgan Chase Institute was established to leverage the power of administrative banking data to deepen our understanding of critical economic issues and provide timely insights to decision makers. Our data are drawn from millions of de-identified JPMorgan Chase accounts, enabling us to shed light on how consumers interact with their personal finances and their broader economic environments, without compromising our customers' privacy (see Box 1).

As in all research, our ability to extend insights gained from the Chase portfolio to the US population relies on having or approximating a sample that is representative of the broader population and being able to describe how the results differ by age, income, geography, and other key attributes. For example, if we want to measure growth in consumer spending in Houston, as we do with our Local Consumer Commerce Index, we want to make sure that the customers we observe in Houston are truly representative of that city, and we might also want to know who within Houston is contributing most of the growth.

Financial institutions have some but not all of the tools at their fingertips to segment and assess representativeness of customers. As a result of Know Your Customer Requirements stipulated by the US Patriot Act of 2001, financial institutions are required to know a person's age and residence before they can open a bank account for her. These are key attributes that could be used to reweight the population to match the nation along those dimensions.

When it comes to income, the task is much more difficult. Financial institutions are not required to know a customer's income in order to open a checking account for her. Once that account is opened, the financial institution might infer income levels from observing the inflows into that account, but even still, those flows would represent take-home income and might be incomplete or difficult to interpret as true income. Moreover, those checking account inflows would be difficult to compare to measures of pre-tax, gross family income provided in public data sets.

Meanwhile, federal credit underwriting criteria (e.g., Regulation Z §1026.51 Ability to Pay) require financial institutions to collect—and, in some cases, verify—information on income for the purposes of extending loans in order to ensure that customers have the ability to pay back the loan, estimates which might be more comparable to the public measures. In this brief we explore methodologically whether those sources of income information used for underwriting, stated by the customer and verified by the financial institution, can be used as the "ground truth" alongside vast amounts of administrative and public data to estimate gross family income for customers for whom such income information does not exist. Ultimately, the use case for this income estimate is to segment or reweight sample populations exclusively for analytical and research purposes and not for business decisions.

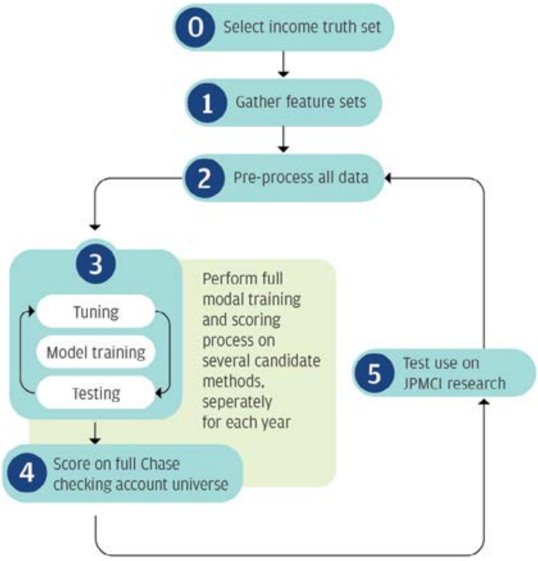
This technical note describes the methodology behind version 1.0 of the JPMorgan Chase Institute Income Estimate (JPMC IIE). Version 1.0 of the JPMC IIE is able to predict gross family income with a mean absolute error of 41 percent and assign families to the correct income quintile 55 percent of the time. Across all income quintiles, roughly 90 percent or more of the observations were classified in the correct or an adjacent income quintile. We describe efforts to overcome sparse coverage, uncertain accuracy, and systematic biases in the samples for whom such "ground truth" exists. We broadly outline the data sources used to generate 400 raw features and the steps taken to pre-process those into almost 800 candidate features used to predict income. We describe efforts to tune the model to achieve the best performance. And we report diagnostic tests used to assess the validity of version 1.0 of JPMC IIE to reweight our sample in our first use case, the JPMorgan Chase Institute Healthcare Out-of-Pocket Spending Panel. We conclude with insights from developing version 1.0 of JPMC IIE and ideas for further improvement. As with version 1.0, the goal of subsequent versions of JPMC IIE will remain unchanged: for JPMorgan Chase Institute's use in segmenting, reweighting, and analyzing populations to deepen understanding of our research insights.

Methods

Figure 1 illustrates process flow for developing version 1.0 of JPMC IIE. A brief summary follows, and the remainder of this paper discusses details and results of each step.

At a high level, we select an income truth set by assembling income data from mortgage applications and credit card information from a base population of Chase checking account customers. We then gather the feature set from sources internal and external to the bank for use as model inputs. Data pre-processing occurs prior to model training to ensure that all features are appropriately structured prior to use in the model. The model training itself progresses in an iterative loop, cycling through hyperparameter tuning, model training, and testing. After scoring the models on the full checking account universe, performance on JPMCI research use case determines whether additional rounds of processing and training are needed, and the cycle begins anew.

Figure 1 - JPMC IIE development process flow


0 Select income truth set
1 Gather feature sets
2 Pre-process all data
3
Tuning
Model training
Testing
Perform full modal training and scoring process on several candidate methods, seperately for each year
4 Score on full Chase checking account universe
5 Test use on JPMCI research

Source: JPMorgan Chase Institute

Box 1: JPMC Institute – Public Data Privacy Notice

The JPMorgan Chase Institute utilizes rigorous security protocols to ensure all customer information is kept confidential and secure. Our strict protocols and standards are based on those employed by government agencies and we work with technology, data privacy and security experts to maintain industry leading standards.

There are several key steps the Institute takes to ensure customer data are safe, secure, and anonymous, including:

- Removing all unique identifiable information—including names, account numbers, addresses, dates of birth, and Social Security Numbers—before the Institute receives the data.
- Putting in place privacy protocols for researchers, including rigorous background checks and strict confidentiality agreements. Researchers are contractually obligated to use the data solely for approved research and may not re-identify any individual represented in the data.
- Disallowing the publication of any information about an individual, consumer, or business. Any data point included in any publication based on the Institute's data may only reflect aggregate information.
- Storing data on secure servers and under strict security procedures such that data cannot be exported outside of JPMorgan Chase's systems. The data are stored on systems that prevent them from being exported to other drivers or sent to outside email addresses. These systems comply with all JPMorgan Chase Information Technology Risk Management requirements for data monitoring and security.

The Institute prides itself on providing valuable insights to policymakers, businesses, and nonprofit leaders. But these insights do not come at the expense of JPMorgan Chase customer privacy or security.

Data Sources

The goal of JPMC IIE is to predict gross family income of Chase checking account customers each year from 2013 to 2017. We restrict the prediction exercise to customer-months that have sufficient checking account activity to establish a relationship with the bank. We aggregate all of the data in the checking account universe to the primary account holder level. Broadly, the estimate predicts "ground truth" income combined from mortgage and credit card applications. Absent self-reporting errors, both sources of income reflect customers' gross family-level income. For customers present in the checking account universe, we obtained income information from mortgage applications and credit card data processed from 2013 to 2017, which we divided into separate truth sets by year. We gathered feature sets from sources both internal and external to the bank and aggregated the feature sets at the annual level for each calendar year.[1]

Truth Set and Additional Filtering

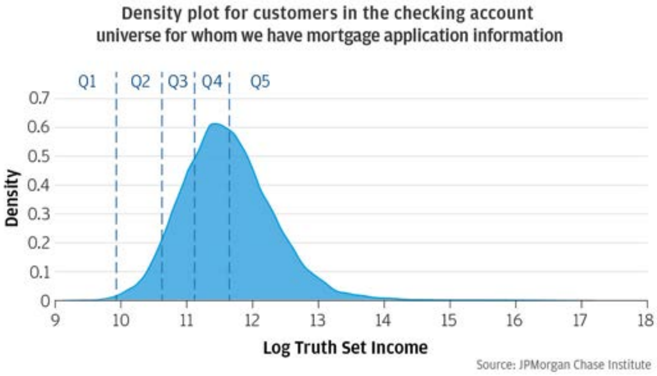
Our initial choice for income truth set was the mortgage sourced income. As this income was verified during the mortgage application process, it should be highly accurate. It became apparent, however, that this truth set was highly unbalanced and biased towards high income groups. Specifically, when we bucketed the mortgage verified income into quintiles defined by the American Community Survey(ACS), close to 50 percent of the sample was in quintile 5, compared to just 6 percent in quintiles 1 and 2 (Figure 2).[2] This poses a challenge for generalizability and our ability to estimate income for and reweight checking account customers across the entire income spectrum in the US.

Unlike the mortgage sample, the checking account universe includes customers who have no credit products. It therefore has larger representation in the lower ACS income quintiles than the mortgage truth set. In addition, mortgage customers have different underlying attributes compared to checking account customers who do not have a mortgage, which could further exacerbate the problem. For example, mortgage customers tend to be older and have higher levels of overall indebtedness than customers who do not qualify for a mortgage.[3] Given these concerns, we sought ways to increase the representativeness of our sample.

Including credit card income in the truth set was a possible solution. While the credit card income provided greater representation of the lower income groups, this source of income was customer-reported and not verified. Figure 3 shows that the median percentage difference of credit card stated income minus mortgage verified income was positive among customers who applied for the two products within the same year. In other words, customers tended to state more income on their credit card applications than was verified on their mortgage applications.

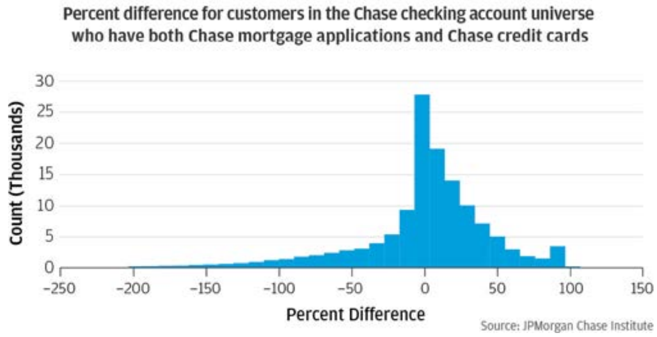
Our final truth set includes customers for whom we have information on either mortgage income or credit card income. Although customers tend to state more income on credit card applications than is verified on their mortgage applications, this may represent income from unverifiable sources, such as cash. To avoid losing this information, we averaged the two sources of income when both were present. Finally, we log-transformed income for model training and assessments to address the positive skew in the distribution of truth set income.

Figure 2 - Mortgage verified income



Source: JPMorgan Chase Institute

Figure 3 - Stated income minus mortgage verified income



Source: JPMorgan Chase Institute

Difference (as % of stated income)	
1st Quartile	-10%
Median	5%
3rd Quartile	25%
Correlation	0.44

Source: JPMorgan Chase Institute

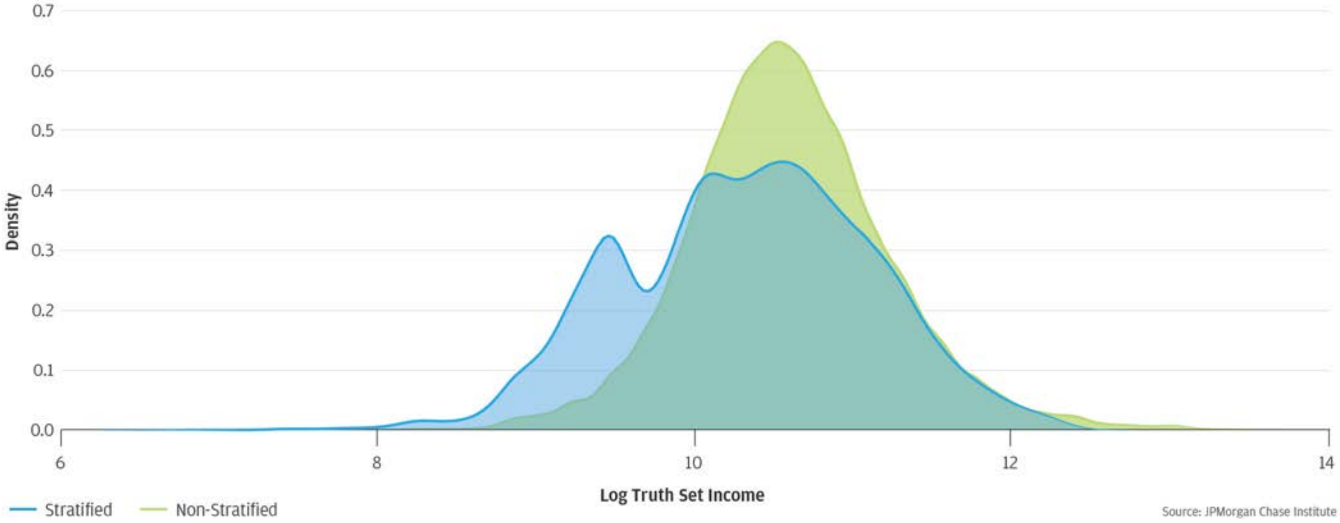
After combining the two forms of income data, we performed three steps to improve upon the accuracy and representativeness of our final income truth set. To address accuracy, we first removed customers whose truth set income was less than income inflows into their checking account.[4] We constructed a conservative view of income based on inflows by summing only the inflow transactions that we categorized as income. 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. We therefore removed customers from the sample whose truth set income was below their checking account inflows. Overall, approximately 9 percent of customers met this condition for sample removal.

Second, in our remaining sample, we also removed customers whose truth set income was in the top or bottom percentile of the truth set in order to train the model without undue influence of extreme observations.[5]

Finally, we sought to address sample representativeness, especially in the lower income quintiles. We bucketed the truth set incomes by ACS quintiles and created a stratified sample by randomly selecting 50,000 customers from each quintile to form our final truth set. This yielded a truth set of 250,000 customers each year. As shown in Figure 4, the stratified sample has better coverage across the income spectrum and improves representation among low-income families, relative to the un-stratified sample.

Figure 4 – Final income truth set used for model training

Density plot comparing the truth set before and after stratifying based on ACS income quintiles; we used this stratified sample for model training. Both densities represent the data after removing customers with (1) truth set income less than income inflows, and (2) truth set income in the 1st or 99th percentile.



Source: JPMorgan Chase Institute

Feature Set and Data Treatment

The feature set used to predict income originated from sources both internal and external to the bank. Internally, we include four main groups of features, appended to the file of the checking account primary account holder. We aggregated account features at the annual level, capturing the maximum, minimum, average, range, and total of each feature within the calendar year:[2]

- **Customer information:** Age, inferred gender, ZIP code;
- **Checking account attributes:** We categorize checking account inflows into several income categories, such as labor income. We also include features that characterize the inflow channel, such as cash deposit. We include additional checking account attributes, such as end-of-month checking account balance;
- **Credit card attributes:** Number of Chase credit cards and card attributes, such as the credit limit; and
- **Attributes of other accounts:** Number of loans and loan terms; total liquid assets across all deposit accounts.

Although checking account outflows and spending categories are also likely to be predictive of income, we exclude those from the feature set for two reasons. First, in administrative banking data, income and spending are mechanically linked, insofar as an account holder can generally only spend that which she has deposited into the account (without incurring overdraft fees). Second, we intend to use JPMC IIE to explain spending behavior in future research and cannot do so if that spending behavior is an input into JPMC IIE, as the relationship would become circular.

In addition to administrative banking data features, we also gathered 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.

In total this yielded 400 raw candidate features for model training per year. We appended these 400 features to the income truth set contemporaneously, meaning that we assembled a feature set for each year and associated it with an income truth set of the same year.

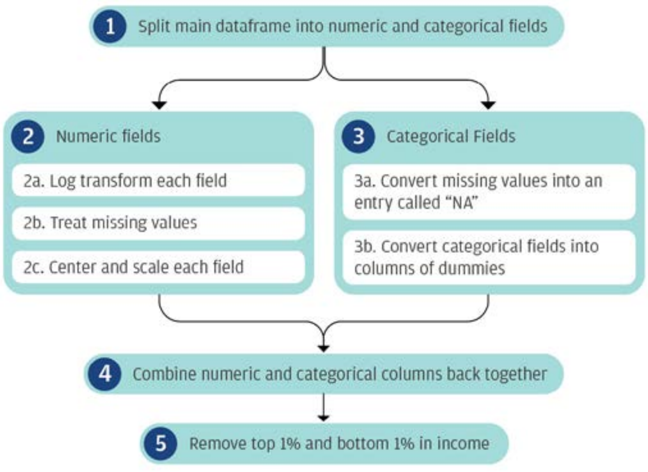
Pre-processing numeric features: Figure 5 illustrates the steps taken to pre-process our input data. Broadly, the steps to process numeric features include:

- **Applying log-transformation:** We use a log function to transform our features so that they approximate a normal distribution. Two reasons guided this decision. First, many features exhibited log-normal distributions with long right tails due to outliers. We transformed our features in order to reduce the influence of outliers in model training. Secondly, since linear regression is among the candidates in our model selection, having normal distribution in features also allows for the relationship between feature set variables and truth set income to be approximately linear. To account for zero and negative values in the feature columns, we use the function log(X + 1 – min(X)) to transform any given feature X.
- **Treating missing values:** Rather than assume that data were missing at random, we allowed for the possibility that patterns in the incidence of missing observations might contain useful predictive information. We created a missing-indicator binary variable for each feature, which takes the value of 1 for customers who did not have recorded information for that feature. These variables retain a record of missing information for use as separate candidate features. We then imputed each feature's missing entries with the average of its non-missing values.
- **Standardizing features:** In our final step, we standardized each numeric feature by subtracting its mean and dividing by its standard deviation. This is to ensure consistent units across features, which allows models that use gradient descent algorithm (e.g., linear regression with regularization, Support Vector Machine) to converge faster.[6]
- **Interaction terms:** We created interaction terms between age and the following numeric features: credit limit, checking account inflows, and revolving credit card balances. The interactions proved to be highly predictive of our ground truth income.

Pre-processing categorical features: For each categorical feature, we converted missing values into a distinct category. We then use one-hot encoding to convert each categorical feature into columns of binary variables. While this works for most features, we encountered a challenge with the treatment of ZIP code. Due to the number of distinct ZIP codes in our data set, converting each distinct value into a separate binary variable would expand our total feature count by upwards of 30,000 columns. To circumvent this, we converted each ZIP code into two numeric fields: the longitude and latitude of its centroid.

The processing described above expanded the total candidate feature set to 780 features due to the addition of missing value flags and interaction terms. The number of features raises concerns of multicollinearity, which we explored by removing features that were highly correlated with others from the feature set. As this step did not impact model performance as measured by MAE, we did not include it in our final pre-training data processing.[7]

Figure 5 – Procedures for processing the feature set prior to training the estimate


1 Split main dataframe into numeric and categorical fields
2 Numeric fields
2a. Log transform each field
2b. Treat missing values
2c. Center and scale each field
3 Categorical Fields
3a. Convert missing values into an entry called "NA"
3b. Convert categorical fields into columns of dummies
4 Combine numeric and categorical columns back together
5 Remove top 1% and bottom 1% in income

* Repeated for each year from 2013 - 2017.

Source: JPMorgan Chase Institute

Income Estimation Benchmarks

To better understand the incremental value of a machine learning approach, we constructed two naïve approximations of income for benchmarking purposes: the Inflow Benchmark and the IRS Benchmark, described below. These benchmarks grounded our ability to estimate gross family income without the use of predictive modeling. They also helped ascertain the value and predictive power of administrative banking data in estimating gross family income.

The Inflow Benchmark used inflows into customers' checking accounts that we categorized as income to proxy take-home income. We then adjusted this inflow-based measure of take-home income to approximate post-tax income by dividing by one minus the federal tax rate of each income bracket to approximate gross family income.[8]

The IRS Benchmark used ZIP code level average IRS-reported income to proxy income for each individual based on their reported ZIP code. This benchmark approximates ability to predict income using only publicly available data.

The relationship between the benchmark measures and our truth set is presented in Figure 6. Unsurprisingly, both benchmarks yielded high mean absolute errors. MAE values were 162 percent for the Inflow Benchmark and 103 percent for the IRS Benchmark.[9] Put differently, if we only used checking account inflows or publically available incomes at the ZIP code level, we would misestimate gross family income on average by over 100 percent. In coming sections, we will see that information from high-frequency deposit accounts can be powerful for predicting income in conjunction with other features in a machine learning approach, yielding a significantly lower MAE than either benchmark. Additionally, the Inflow Benchmark demonstrates that checking account inflows alone are not sufficient for predicting gross family income.

We used these MAE values for comparison purposes as we developed our machine learning approach to income estimation. Any machine-learning-based income estimates should have considerably lower MAE than these benchmarks to justify the added complexity of the approach. The remainder of this paper discusses our approach and results in detail.

Box 2: Overview of the modeling algorithms considered for training JPMC IIE version 1.0

Linear regression is a parametric statistical modeling technique with a long history of successful use in analyzing financial behaviors. Regression algorithms produce interpretable modeled relationships, while suffering from rigid assumptions around functional forms of those relationships, as well as risk of overfit. The former requires substantial data pre-processing to address; we mitigate the latter via an elastic net algorithm to penalize overfit and remove unnecessary features.

Gradient boosting machines (GBM) and random forests are machine learning algorithms 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. The random forest algorithm trains many deep decision trees, averaging the individual predictions for the final estimate. These nonparametric methods are capable of fitting relationships with no requirements around the underlying functional forms. However, the final modeled relationships are less interpretable than with regression, and can also be prone to overfit.

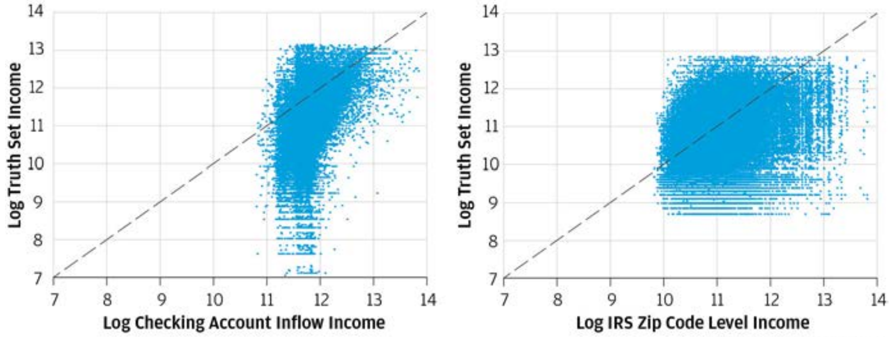
Support vector regressors are machine learning models constructed by partitioning the training data feature space into maximally-separate regions by a set of hyperplanes. The algorithm has flexibility to partition with linear or non-linear hyperplanes, but is prone to the same disadvantages of GBM and random forests—lack of interpretability and overfit.

Machine Learning Algorithms

We trained models using a variety of algorithms and hyperparameter settings to find the best option for our purposes. We considered the following methods: linear regression, gradient boosting machines, random forests, and support vector regressors. Box 2 briefly outlines each method, including strengths and potential weaknesses. Methodological References lists in-depth sources. We discuss model training details and results in subsequent sections.

Figure 6 - Benchmark income measures

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



Source: JPMorgan Chase Institute

Model Training

Data usage: Model training and optimization used a sample of 250,000 customers stratified on income quintile, as described in the Truth Set and Additional Filtering section. In order to train a model that is generalizable to different populations of our research universe, we separated the 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 overfitting
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

As described in the Data Sources section, the feature set and truth set were constructed on a yearly basis from 2013 to 2017. As such, we constructed separate training, validation, and testing datasets for each of these years and trained the models accordingly. We took this yearly approach to mitigate the performance degradation we observed on out-of-year predictions. In preliminary analysis, model error increased when scoring the model on samples outside of the model's training timeframe (Figure 7). Since this estimate will be used to predict income on customers for whom we have no income data, rather than to predict future income, having a consistent out-of-year prediction was less necessary than in other machine learning exercises. Thus, we were comfortable with the yearly training approach.

Figure 7 - MAE degradation on out-of-year data

For models developed on one year and then scored on another

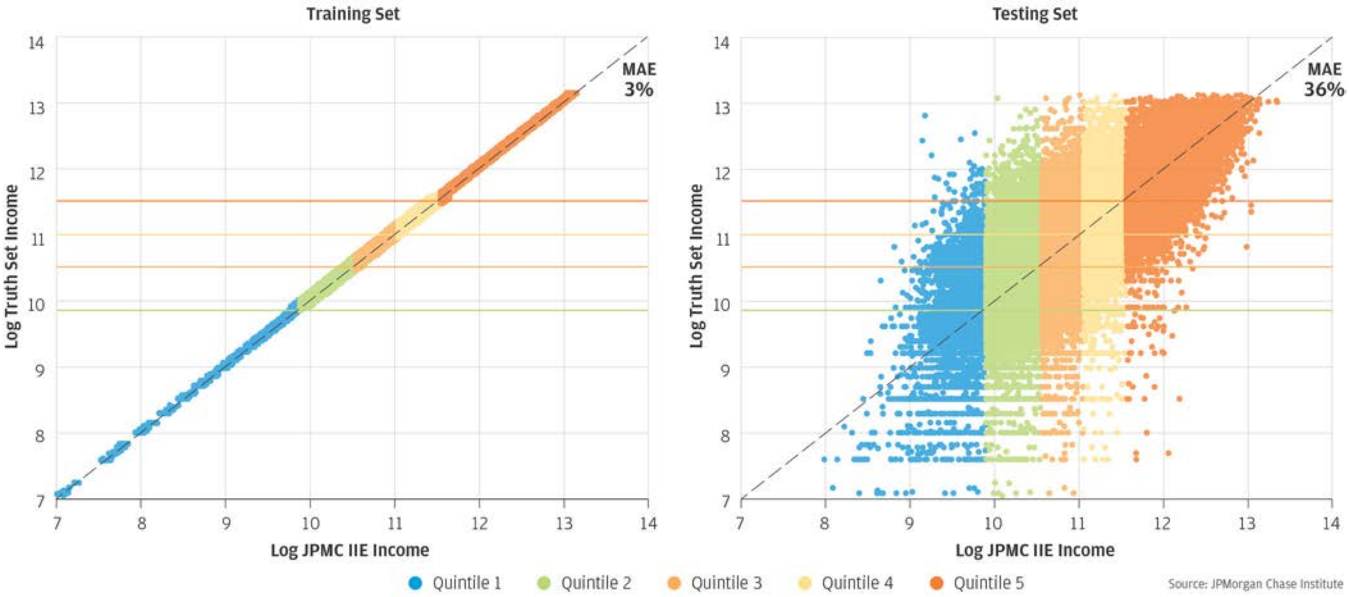
Data Model	2014	2015	2016
2014	37%	39%	40%
2015		36%	54%
2016			54%

Source: JPMorgan Chase Institute

Model selection to prevent overfitting: Overfitting occurs when a model performs extremely well during training and optimization but very poorly out of the training sample. One of the models we trained—a GBM with learning rate of 0.05 and maximum depth of 20—exemplifies this. The model performed very well on the training data (MAE of 3 percent) but was unable to generalize well beyond it (testing set MAE of 36 percent).

Figure 8 – Overfit candidate model, trained and assessed on 2017 data

Estimated income vs. truth set income for an example gradient boosting machine (GBM) with large performance degradation on the testing set



Source: JPMorgan Chase Institute

This highlights the importance of our data approach, and the necessity of assessing models beyond their training data. As model specification becomes more complex, model error will decrease on the training set, fitting ever more closely to the specific observations in the training data. In contrast, the error on the validation set will begin to increase when the model is over-specified to the training data, indicating that the model is no longer generalizable outside of the training set. Overfitting is always a risk because the model is tuned to minimize the error for the training set; assessing performance on a validation set is necessary to guard against that risk.

Throughout the model training process, performance (MAE) on the validation set guided our decisions regarding hyperparameter settings and model selection.

Hyperparameter tuning approach: For each year, model training iterated over 105 different models built from four different estimators: gradient boosting machines, random forests, elastic net linear regression, and support vector regressors. Each of these models required hyperparameter tuning in order to determine the best specification.

Hyperparameter tuning is useful to prevent over-fitting through regularization. In simple terms, regularization is tuning or selecting the preferred level of model complexity to ensure model generalizability. Without this step, models may be too complex and overfit or too simple and underfit, giving poor predictions in either case.[10]

Specifically, we focused on the following hyperparameters[11] to prevent overfitting:

Linear regression

- **Lasso & ridge regularization[12]:** The most common types of regularization. These update the general cost function by adding another term known as the regularization term.
- **Cost function** = Loss (say, binary cross entropy) + Regularization term Reducing feature set

Gradient Boosting and Random Forest

- **Number of estimators:** Adding more trees to the model can help improve performance, and high-performing GBMs often have hundreds of trees. Though at some point in the training process, adding more trees leads us to overestimate model complexity and thus overfit to the training data.
- **Maximum depth:** Deeper trees are more complex and shallower trees are preferred. Generally, better results are seen with 5-10 levels.

Support Vector Regressors

- **Kernel type:** The kernel introduces a similarity function that allows dropping assumptions of linearity. We considered two kernel options: linear and radial basis function (RBF).
- **Margin:** The margin alters the decision boundary to consider data by adding a penalty to the error term. We considered values ranging from 0.0001 to 1.

In this estimation we used a mixed search approach, combining grid search and random search techniques.[13] This enabled a more informed selection of parameters that didn't require the grid search approach of assessing all possible combinations, which can add substantial computational and timing costs for the project. Without the full grid search, we risk missing the optimal hyperparameter settings, as the mixed approach does not assess every combination of options. We accepted this trade-off, and decreased the risk by exploring several iterations of search, with multiple hyperparameter settings per iteration.

For each hyperparameter, the mixed grid search approach selects a few values from a wide range of possible values and then observes the results on both the training and the validation sets. In a second iteration, we increase the range around those hyperparameter values that show the most promise in model performance, as measured by MAE on the validation set. For instance, in gradient boosting, the first tuning iteration for maximum depth assessed three possible values: 5, 10, and 25. We observed that a maximum depth of 5 minimized the validation set MAE, and so focused the second iteration on a nearby range, assessing values of 2, 5, 6, and 7.

The choice of a mixed approach enabled the hyperparameter tuning and model performance benefits of grid search without exceeding timing and computational constraints.

Results

Final Model

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 aimed to minimize the mean absolute error (MAE) of the point estimates and also to predict the correct income quintile accurately, both overall and within each income quintile. Here we present the results of version 1.0 of JPMC IIE.

For each year of data, our mixed grid search iterated through many versions of each of our four candidate methods: gradient boosting machines, random forests, elastic net linear regression, and support vector regressors. For our particular data and performance criteria, the best-performing model from each year was a gradient boosting machine (GBM). Given our focus on performance optimization, rather than the functional relationship between truth set income and each feature input, we were not surprised to see the tree-based ensemble methods (GBM, random forest) outperform linear regression. The specified hyperparameters of the selected GBM models are shown in Figure 9.

Figure 9 - Hyperparameter values for version 1.0 of JPMC IIE

HYPERPARAMETER	VALUES
Learning Rate	.05 for all years
Loss Function	Least Square for all years
Max Depth	5 for 2013, 2016 & 2017, 6 for 2014 & 2015
Min. Samples Split	10 for 2013, 2014 and 5 for 2015, 2016, 2017
Number of Estimators	500 for all years

Figure 10 - Performance metrics for version 1.0 of JPMC IIE

YEAR OF TRAINING	TRAINING SET MAE	TESTING SET MAE	QUINTILE PREDICTION ACCURACY
2013	38%	40%	55%
2014	39%	41%	54%
2015	39%	40%	54%
2016	38%	42%	53%
2017	37%	41%	54%

Characteristics of the Final Model

Using the 2017 model, we will outline key characteristics of model performance. Results are shown only for 2017, but patterns and conclusions hold across each year's model.

As most of JPMCI's research uses income on the ACS quintile basis, we prioritized consistent accuracy across those quintiles. Figure 11 shows classification across truth set income quintiles, within each predicted income quintile. We observe that the accuracy rate is consistent across predicted quintiles, as desired. In addition, misclassifications tend to be concentrated in the adjacent quintiles. For instance, 0.8 percent of the individuals in the predicted first quintile actually belong to the fifth quintile, and 32 percent (the majority of mispredictions) actually belong to the second quintile. Across all income quintiles, roughly 90 percent or more of the observations were classified in the correct or an adjacent income quintile.

Figure 10 shows results for MAE and quintile prediction accuracy across years. Quintile prediction accuracy refers to 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, yielding on average across years an 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. This means that the models are robust to different underlying distributions and are not overfitting to the training sample.

Inputs to the Final Model

Across each modeled year, the features show some patterns in importance. For instance, credit card limit and total liquid assets across Chase deposit accounts are consistently two of the most predictive fields, along with socio-demographic variables such as age and longitude of ZIP code centroid. Of the remaining model features, the vast majority of predictive fields are related to checking account inflow amount and average, as well as checking account and credit card balances.

Figure 11 - ACS quintile accuracy by predicted quintiles

Consistent across predicted income quintiles, correct classification rate into corresponding truth set quintile is 53%-57% (shown for 2017 version 1.0 of JPMC IIE)

	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

The consistent accuracy across quintiles is the result of stratifying our training sample by ACS income quintile (described in the Truth Set and Additional Filtering section). 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 12). 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 13 shows that low predicted income values are skewed slightly toward underpredicting their corresponding truth set income.

Figure 12 - ACS quintile accuracy by stratification options

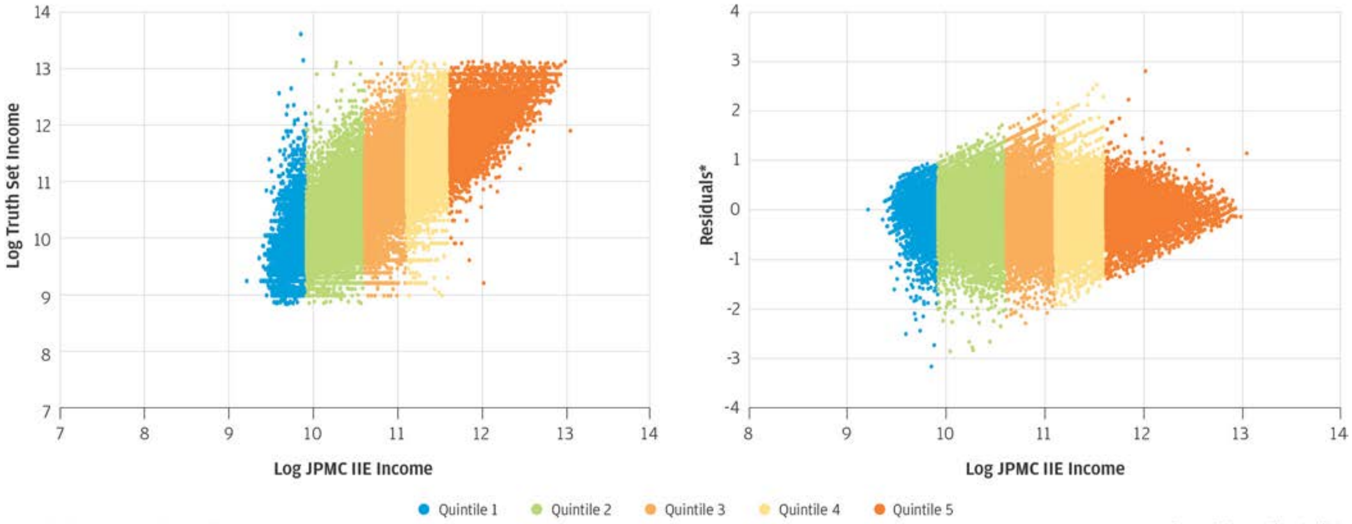
Comparison of final version 1.0 of JPMC IIE vs. test version trained on the un-stratified sample

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 13 – Estimated income vs. truth set income, and corresponding residuals

2017 version 1.0 of JPMC IIE; points are color-coded by predicted income quintile



*Residuals represent truth set minus estimate

Source: JPMorgan Chase Institute

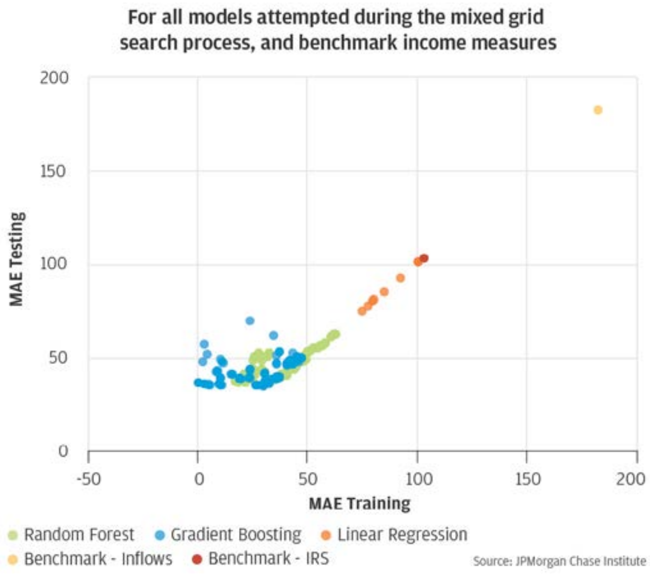
With broader scope of analysis, we would ideally perform residual analysis to gain a deeper understanding of the above patterns. We note this as an area for future model exploration (see Discussion section for additional commentary).

Cross-Model Comparisons and Trade-offs

The MAE on the training and testing sets is presented in Figure 14 for all candidate models attempted during the mixed grid search process. MAE values for our two benchmark estimates are included as well, showing that the models perform better than the benchmarks. In addition, broadly, we observe two categories:

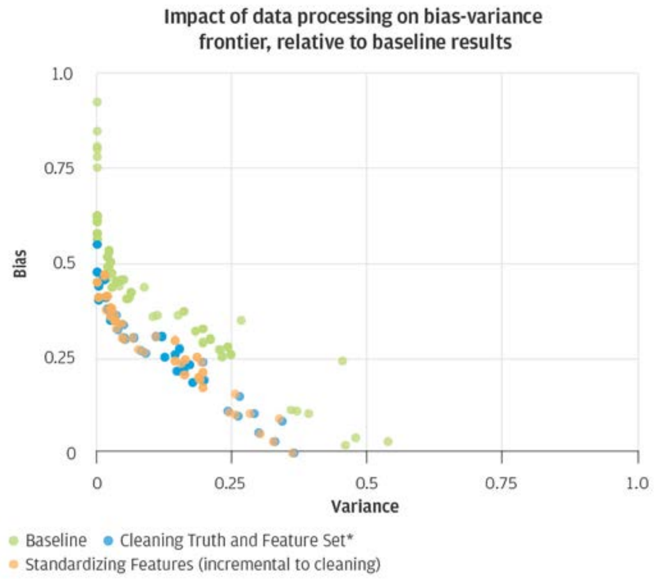
- Models that had low difference between training and testing MAE, but high error rates. These models have relatively low predictive power, including the attempted logistic regression runs and some of the GBM runs.
- Models with low training MAE but significantly higher testing MAE, indicating overfit. Several GBM and random forest runs exhibit this behavior.

Figure 14 - Testing set MAE vs. Training set MAE



Source: JPMorgan Chase Institute

Figure 15 – Model training bias-variance trade-offs



*For details, refer to the steps discussed under "Truth Set and Additional Filtering" and "Pre-processing numeric features"; note that Baseline uses the stratified sample and includes interaction terms, and remaining steps were applied here

Source: JPMorgan Chase Institute

The previous two characteristics are referred to as bias and variance, where bias is the predictive power of the model, and variance is the generalizability to different underlying distributions. The trade-off between these two characteristics is one of the most studied problems in machine learning.

In our analysis, we found that hyperparameter tuning results in modeling outcomes that form a frontier in which there is a clear trade-off between bias and variance. Additional updates to hyperparameters can move results along this frontier, trading off between the two without the ability to improve both simultaneously. (See the frontier formed by the Baseline points in Figure 15.) In order to improve both —to move the bias-variance frontier inward—thorough data exploration and treatment is critical. For example, completion of the data pre-processing steps discussed in the Feature Set section shifted the frontier inwards, allowing us to reduce both the bias and variance of our candidate models.

Case Study: Incorporation of Final Model into Institute Research

In a recent Institute publication, On the Rise: Out-of-pocket Healthcare Spending in 2017, we tested JPMC IIE performance in reweighting the sample population for the JPMorgan Chase Institute Healthcare Out-of-pocket Spending Panel (HOSP).

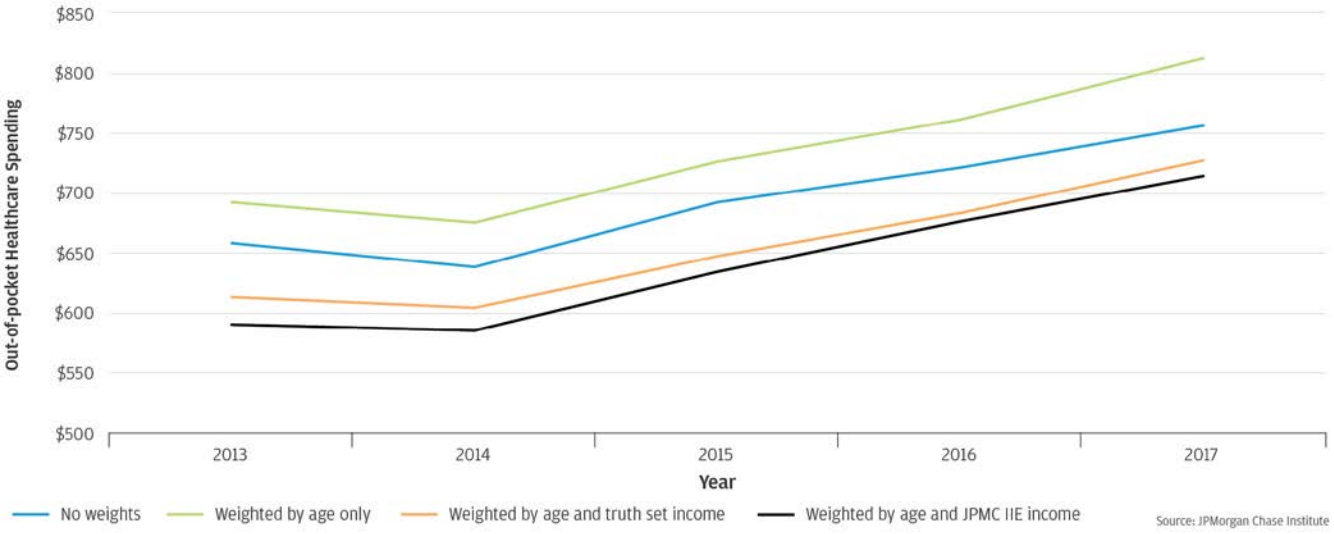
Figure 16 compares out-of-pocket health spending levels when we use version 1.0 of JMPC IIE income as an input to reweight the HOSP sample to match each state's joint age and income distribution. For the sample of families for whom we have truth set income, the average out-of-pocket health spending levels were similar when we weighted the sample using JPMC IIE and age (blue line) compared to when we weighted the sample using age and truth set income (black line)—the "gold standard" for comparison. Applying JPMC IIE and age yielded estimated spend levels much closer to this gold standard than either an unweighted sample or weighting the sample exclusively by age, another variable that is arguably less prone to measurement error. We concluded that use of version 1.0 of JPMC IIE is valuable in weighting the HOSP sample to more closely represent each state by age and income.

Box 3: Background on the Healthcare Out-of-pocket Spending Panel (HOSP) data and analysis

Leveraging financial transaction data, the JPMorgan Chase Institute provides a unique cash flow view of families' healthcare out-of-pocket spending and financial burden. In 2017 we released the first estimates of out-of-pocket healthcare spending levels and burden at the state and county level from 2013 to 2016, from our JPMorgan Chase Institute Healthcare Out-of-pocket Spending Panel (JPMCI HOSP) data asset. In On the Rise: Out-of-Pocket Healthcare Spending in 2017, we describe enhancements to, and key findings from, the JPMCI HOSP data asset that includes the first available estimates of 2017 healthcare out-of-pocket spending trends, as well as a first-ever look at year-over-year trends at the state and county level and for different demographic groups.

The HOSP sample draws on families who reside within the 23 states with a Chase branch footprint. In order to make each state sample more representative of the general population within that state, we reweight each state's sample to match the joint age and gross income distribution within that state according to the American Community Survey (ACS) for each year from 2014 through 2017. Our unit of analysis is the primary account holder, whom we refer to as a family. When reweighting our sample, we match the joint age and income distribution of our primary account holders to the heads of family in the ACS. Our weighting approach leverages the JPMorgan Chase Institute Income Estimate (JPMC IIE), which is an estimate of gross family income developed using machine learning techniques to generate an income estimate based on checking account and credit card attributes.

Figure 16 - Out-of-pocket health spending across years, by different weighting schemes



Source: JPMorgan Chase Institute

Discussion

JPMC IIE represents the JPMorgan Chase Institute's first attempt to leverage administrative banking data and machine learning approaches to estimate gross family income. In doing so we gleaned several important insights regarding both methodology and the value of administrative data.

1. **Administrative banking data offer powerful insights in enhancing income estimates, but alone are insufficient to predict income.** High-frequency deposit account information, in combination with other data features and a machine learning approach, yielded a significantly more accurate prediction of income than public estimates of income based on ZIP code (the IRS benchmark), reducing the mean absolute error by two thirds, from 103 percent to 41 percent.

   That said, deposit account inflows alone are insufficient to develop an accurate estimate of gross family income: the Inflow Benchmark had a higher error rate than even the IRS benchmark (an MAE of 162 percent compared to 103 percent). There could be several reasons why deposit account inflows alone are insufficient to approximate gross family income. First, to the extent that families spread their income sources across multiple financial institutions, any single account or set of accounts with a single institution may provide an incomplete picture of total income. Second, with respect to inflows, it can be difficult to discern the economic purpose of all inflows and specifically distinguish between incoming transfers and true income. Finally, inflows represent take-home income, after taxes and payroll deductions for retirement and other savings and charitable giving that can be facilitated by the employer. Appropriately accounting for all of those deductions can be difficult when attempting to scale take-home income up to an estimate of gross family income.

2. **Individual-level income information creates opportunities to study more granular income patterns, forming a valuable supplement to other available data.** Large scale, publicly-available income data is aggregated, often at the ZIP code or Census tract level. While this provides valuable information for analysis, it obscures information at the tails of the distribution, which is of interest to researchers. Another key source of information for analyzing consumer financial patterns, credit bureau data are available at the individual level. However, credit bureau information focuses on the debt side of consumers' finances. This creates a gap, with no clear sources of individual level information on the asset side of consumers' finances or their income statement. JPMC IIE shows promise as a means of filling this gap.

3. **Ascertaining and improving the veracity and representativeness of the truth set is critical when leveraging administrative data for prediction tasks.** On the veracity side, we cleaned the truth set to remove customers whose ground truth income was less than income inflows into their checking account. We saw that this shifted the bias-variance frontier inward, allowing us to reduce the bias and variance of our estimate simultaneously. For representativeness, our large sample sizes afforded us the opportunity to stratify our truth set to obtain more accurate predictions in underweighted parts of the income distribution. Stratifying the truth set yielded a 28 percentage point improvement in the quintile prediction for families in the lowest income quintile. In order to study the distributional impacts of economic trends and public policies, it is critical to measure income equally well across the income distribution.

Even still, there is room for improvement in balancing estimation bias across the income distribution. Version 1.0 of JPMC IIE tends to skew low predicted income values toward underpredicting their corresponding truth set income. Below we describe a few ideas to continue to tackle this thorny estimation challenge and ensure accuracy across the JPMC IIE spectrum.

LImitations and Future Direction

With a validated, working version 1.0 of JPMC IIE in place, upcoming efforts will focus on enhancement and expansion of scope. Due to the proof-of-concept nature of version 1.0, we prioritized performance and use-case assessments to confirm viability of an income model for use in research. As such, certain avenues of exploration were put on hold for future iterations. Broadly, planned future enhancements fall into three categories: data expansion, feature refinement, and insight exploration.

Data expansion: As discussed in detail, version 1.0 of JPMC IIE relies on the de-identified Chase checking account universe as its base population. This means that the model 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. For version 2.0, we will obtain credit bureau data for a sample of de-identified Chase customers, regardless of their presence in the checking account universe. This will accomplish two critical goals: (1) expand the population of customers for whom we can predict income by including credit-only Chase customers, and (2) enhance the predictive power of the model for the existing checking account population by adding new features to our modeling data.

We will also consider the addition of new features from both administrative account level data and other sources, such as expanding the set of features pulled from publicly-available sources. Potential sources include Zillow, Census and County Business Patterns, and IRS. In contrast to the bureau data, these features are at the aggregate instead of individual level (e.g., Census tract or ZIP code level). However, they represent different facets of a consumer's overall profile, and may still be powerful in conjunction with individual level features.

Feature refinement: Despite the pre-processing steps discussed in the Methods section of this paper, we spent minimal time on feature engineering for version 1.0. As noted in the discussion on bias-variance trade-off, thorough data exploration and treatment is critical for establishing a well-performing model.

Avenues for further exploration include creation of features based on additional geographic characteristics beyond longitude and latitude, trends over time (at both customer and geographical levels of aggregation), administrative data from additional deposit banking products, and account attributes. Thoughtful assessment of feature aggregation may also yield powerful predictors that the gradient boosting algorithm cannot easily approximate, such as ratios or other functions of multiple features.

Insight exploration: For version 1.0, we have not yet performed in-depth exploration of the insights underlying the relationships between input features and gross family income. Future assessments will focus on deeper understanding of the relationships captured by the model, and how individual features impact income predictions.

Beyond understanding feature relationships, we are also eager to explore the model holistically to gain perspective on areas of caution. We will explore two approaches: demographic monitoring and residual monitoring. These assessments will begin with version 1.0, prior to more granular use in research, and continue throughout development and validation of version 2.0.

- **Demographic monitoring** will assess whether modeled income reflects demographic biases. If, for example, predicted income is more skewed on the basis of age or gender than truth set income, this might indicate that the model is detecting and then exacerbating income biases present in the truth set. This is a known issue when machine learning algorithms are trained on biased data.[14]
- **Residual monitoring** will assess the model for systematic weaknesses. Are there segments of the population on which the model performs poorly, relative to overall performance? We will compare across segments to understand differences in predictive accuracy, in hopes of identifying improvements to address lower-quintile accuracy of the estimate. This exercise will also shed valuable light on whether JPMC IIE performs particularly well (or poorly) within certain geographies or demographic groups, providing important caveats for use in sample reweighting.

We are looking forward to deeper exploration on all of these fronts as we expand to JPMC IIE version 2.0. We are confident that coming iterations of analysis will yield insights of value to the broader academic, policy, and data science communities and look forward to sharing more in future.

References

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning.* New York, NY: Springer.

G. C. Cawley and N. L. C. Talbot, Over-fitting in model selection and subsequent selection bias in performance evaluation, Journal of Machine Learning Research, 2010. Research, vol. 11, pp. 2079-2107, July 2010.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*
New York, NY: Springer.

Murphy, K.P. (2012) *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: MIT Press.

Li, Ting, Bingyi Jing, Ningchen Ying, and Xianshi Yu. "Adaptive Scaling." (2017). Available at: https://arxiv.org/pdf/1709.00566.pdf

Li, Wei and Laurie Goodman. "Americans' Debt Styles by Age and over Time." (2015). Available at: https://www.urban.org/sites/default/files/publication/72976/2000514-Americans-Debt-Styles-by-Age-and-over-Time.pdf

O'Neil, C. (2018, March 9). Big Data Alone Can't Fix a Broken Bail System. Bloomberg. Retrieved from https://www.bloomberg.com/view/articles/2018-03-09/big-data-alone-can-t-fix-a-broken-bail-system

Glossary

Administrative banking data Data derived from the operation of administrative banking systems, including information from application processes, account characteristics, and customer transactions.

Channel The delivery channel by which money flows in or out of an account. Outflow channels include debit card purchase, ACH—debit, check withdrawal and ATM cash withdrawal. Inflow channels include ACH—credit, ATM cash deposit, ATM check deposit, and teller deposit.

Checking account universe The full set of checking account data compiled by the JPMorgan Chase Institute, with monthly summaries and daily transaction data for JPMorgan Chase checking account primary account holders. We restrict this to customer-months that have at least five checking account outflows.

Credit card income Gross family income provided by JPMC credit card holders at application stage or subsequent updates to their credit card information.

Feature A measurable property or characteristic of the modeling unit. In our case, a feature represents measurable information about a primary account holder in our data universe, or her accounts. Examples include credit limit, number of deposits this month, etc.

Feature set The full set of features used for model training, observed and recorded for each customer in our modeling data.

Hyperparameter A parameter whose value is set before model training begins, to control key elements of the training process. For example, with a gradient boosting machine (GBM) the modeler specifies the number of estimators (trees) to train, and the maximum depth each tree may take; in linear regression with elastic net, the modeler specifies the loss function and alpha values.

Inflow A credit transaction to an account holder's checking account.

Mean absolute error (MAE) A measure of the difference between ground truth income and estimated income based on the absolute value of that difference for each customer in the modeling data. We chose to optimize on mean absolute error rather than mean squared error to avoid penalizing errors in a quadratic functional form, thus increasing the penalty for larger errors.

Mortgage verified income Gross family income provided by JPMC mortgage applicants, and verified during the application process.

One-hot encoding Process for transforming a categorical feature into a set of binary variables representing each level of the original feature.

Outflow Occurs when a model is fit closely to the specific observations in the training set and is no longer generalizable, performing very poorly on observations outside of that sample.

Overfit Resident Consumers: Consumers that live inside of the CBSA in question.

Primary account holder The signatory legally responsible for the account. In the JPMorgan Chase data asset, all account activity is reflected under the person listed as the primary account holder. When there is more than one primary account holder, the account activity is reflected under the person listed first on the account.

Testing set Modeling data is separated into training, validation, and testing sets. The testing set is the portion of the modeling data used to assess the predictive power of the final model, on observations not used for training or hyperparameter tuning.

Training set Modeling data is separated into training, validation, and testing sets. The training set is the portion of the modeling data used to fit the models in order to determine the form of the relationship between truth set income and the feature set. Conventionally, the training set is the largest of the three, generally between 50 and 70 percent of the full modeling data.

Transaction A single deposit or withdrawal of funds by any transaction channel.

Truth set The set of customer-level ground truth gross family income values used for model training and validation; synonymous with dependent variable or target variable.

Validation set Modeling data is separated into training, validation, and testing sets. The validation set is the portion of the modeling data used in parallel with the training set, to tune hyperparameters and guard against overfit.

Endnotes

[1] In future versions of IIE we plan to update this analysis to a rolling 12-month window, corresponding to the 12 months prior to customer providing truth set income. This will avoid inconsistencies in the feature set across customers who provided income at different points in the calendar year.

[2] Using the American Community Survey for each year in our sample (2013 through 2017), we obtained quintile thresholds by calculating the 20th, 40th, 60th, and 80th percentiles of family income reported from the survey. These thresholds were then applied to our data to proxy national gross family income distribution, and are referred to throughout this paper as “ACS quintiles.” ACS income information from IPUMS-CPS, University of Minnesota, www.ipums.org

[3] For example, see Wei and Goodman (2015).

[4] Here income inflow represents inflow transactions into checking account that we categorized as income. It does not include transfers from other financial institutions or inflows that we could not categorize.

[5] During model training, we assessed the impact of including the top and bottom percentiles of truth set income in the modeling data. Reintroducing that population to the model training resulted in an MAE of 43 percent on the testing set, an increase relative to the model trained without exposure to these extreme observations. For comparison, the testing set MAE of the final model is 37 percent.

[6] See Ting et al (2017).

[7] Generally, multicollinearity poses a problem in its creation of: unstable coefficient estimates in regression, or feature importances in tree-based methods that are difficult to interpret. As this modeling exercise is focused on performance rather than interpretability of individual feature relationships, we were comfortable removing this step once the lack of impact on performance was established.

[8] To build this benchmark we aggregated the inflows categorized as income from customers in 2017 and then adjusted by the tax bracket to estimate gross earnings. We used the following tax brackets:

2017 Federal Tax Bracket	
TAX RATE	TAXABLE INCOME
10%	$0 - $18,650
15%	$18,651 – $75,900
25%	$75,901 – $153,100
28%	$153,101 – $233,350
33%	$233,351 – $416,700
35%	$416,701 – $470,700
39.60%	$470,701+

[9] All mean absolute error (MAE) values reported in this paper were calculated on the basis of log-transformed income, for consistency with model training

[10] See Cawley and Talbot (2010)

[11] The following hyperparameter ranges were used during our mixed grid search:

Random Forest

PARAMETER	DESCRIPTION	FIRST ITERATION	SECOND ITERATION
Number of Estimators	Number of trees generated	50,100, 500	500, 1000
Maximum Depth	Maximum number of levels in each tree	5, 10, 25	2, 5, 6, 7
Maximum Features	Function to calculate the number of features to consider when looking for the best split	sqrt, log2	sqrt, log2
Minimum Samples Split	Minimum sample size required to add a further split in a tree	2, 10	2, 5, 10

Gradient Boosting

PARAMETER	DESCRIPTION	FIRST ITERATION	SECOND ITERATION
Number of Estimators	Number of boosting stages to perform	50, 100, 500	500, 1000
Learning Rate	A shrinkage factor that controls the weighting of new trees added to the model	0.05, .01, 0.5	.01, .05, .1 and .5
Subsample	Fraction of observations to be used for fitting the individual base learners	0.1, 0.5, 1.0	0.1, 0.5, 1.0
Maximum Depth	Maximum number of levels in each tree	5, 10, 25	2, 5, 6, 7

Elastic Net Linear Regression

PARAMETER	DESCRIPTION	FIRST ITERATION	SECOND ITERATION
Alpha	Constant that multiplies the penalty terms; alpha = 0 is equivalent to OLS	0.5, 1.0, 1.5	0.5, 1.0, 1.5
L1 Ratio	Elastic Net mixing parameter; 0 for L2 penalty and 1 for L1 penalty	0.25, 0.5, 0.75	0.25, 0.5, 0.75
Intercept	Whether the intercept should be estimated or not	True, False	True, False

[12] Ridge regression adds a squared transformation of the coefficients in the model as penalty term to the loss function. Lasso regression adds the absolute value of coefficient as penalty term to the loss function. The key difference between these techniques is that Lasso shrinks the less important feature's coefficient to zero thus, removing some features altogether. This is particularly helpful for feature selection when we have a large number of features or are worried about multicollinearity.

[13] Hyperparameter optimization is a common problem in machine learning. Machine learning algorithms, from logistic regression to neural networks, depend on well-tuned hyperparameters to reach maximum effectiveness. Different hyperparameter optimization strategies have varied performance and cost. For this project, two methods were considered:

Grid Search suggests parameter configurations deterministically, by laying down a grid of all possible configurations inside your parameter space. To optimize, one evaluates the function at every point on this grid. One caveat is that the number of function evaluations required for this strategy increases exponentially with each additional parameter.

Random Search suggests configurations randomly from your parameter space. To optimize, one evaluates the function at some number of random configurations in the parameter space. One caveat is that it may be unclear how to determine the number of function evaluations required for your particular problem.

[14] See O'Neil (2018) for a discussion of the consequences of training an algorithm on a potentially biased truth set.

Suggested Citation

JPMorgan Chase Institute. "Estimating Family Income from Administrative Banking Data: A Machine Learning Approach" JPMorgan Chase Institute, 2018.

This material is a product of JPMorgan Chase Institute and is provided to you solely for general information purposes. Unless otherwise specifically stated, any views or opinions expressed herein are solely those of the authors listed, and may differ from the views and opinions expressed by J.P. Morgan Securities LLC (JPMS) Research Department or other departments or divisions of JPMorgan Chase & Co. or its affiliates. This material is not a product of the Research Department of JPMS. Information has been obtained from sources believed to be reliable, but JPMorgan Chase & Co. or its affiliates and/or subsidiaries (collectively J.P. Morgan) do not warrant its completeness or accuracy. Opinions and estimates constitute our judgment as of the date of this material and are subject to change without notice. The data relied on for this report are based on past transactions and may not be indicative of future results. The opinion herein should not be construed as an individual recommendation for any particular client and is not intended as recommendations of particular securities, financial instruments, or strategies for a particular client. This material does not constitute a solicitation or offer in any jurisdiction where such a solicitation is unlawful.