FinRegLab

Advancing the Credit Ecosystem: Machine Learning & Cash Flow Data in Consumer Underwriting

Empirical White Paper

ONLINE METHODOLOGICAL APPENDIX
JULY 2025

About FinRegLab

FinRegLab is a nonprofit, nonpartisan innovation center that tests new technologies and data to increase access to responsible financial services that help drive long-term economic security for people and small businesses. With our research insights, we facilitate discourse across the financial ecosystem to inform market practices and policy solutions.

Acknowledgments

This empirical white paper expands on FinRegLab's prior quantitative research on the use of cash flow data for credit underwriting in both consumer and small business markets and managing explainability and fairness concerns in connection with machine learning underwriting models. These prior reports, along with related data science, market, and policy analyses, are available at https://finreglab.org/category/projects/.

Support for this publication was provided by JP Morgan Chase & Co. and Capital One. Detailed information can be found on the inside back cover.

We would like to thank members of our project advisory board: Jay Budzik, Fifth Third Bank; Vickey Chang, Equifax; and Conrod Robinson, FinRegLab Advisor. Keith Ernst, Federal Deposit Insurance Corporation, also participated as a regulatory observer.

Thanks also to the following individuals who provided feedback in connection with this project and the report:

Erin Allard and Brian Duke, Prism Data
Charlie Costello, Alicia Dagosta, and Scott Weber, Upstart
Marsha Courchane and Steli Stoianovici, Charles River Associates
Hunter Hao, Lydia Huo, and Mike Petkun, Nova Credit
Esther Kahng and Sean Kamkar, Zest AI
Scott Nelson, University of Chicago Booth School of Business
Nicholas Schmidt, Solas AI
Sanjana Shellikeri, BLDS
David Silberman, FinRegLab Advisor
Michael Umlauf, TransUnion

Lastly, we would like to acknowledge Zishun Zhao for conducting the empirical analysis, Zishun Zhao and Kelly Thompson Cochran for writing this report, and other members of the FinRegLab team who worked on various elements of this research project, including Ali Bagherpour, Drew Bluethmann, and Sarah Davies.

When viewed with an Adobe Acrobat reader, elements listed in the Table of Contents or in blue text are links to the referenced section or feature. Functionality may be limited in non-Adobe readers. Adobe's reader can be downloaded for free at get.adobe.com/reader.

APPENDIX C CONTENTS

APPENDIX C

Detailed Model Development Methodology

C.1 Conceptual framework and modeling methodologies

An essential task of credit underwriting is to predict credit risk, i.e., the likelihood of default, if a loan being requested by a prospective borrower is granted. At a broad level, lenders often describe this process as involving the "5 Cs of credit," which include character (history and reputation), capacity (ability to pay based on current income and debts), capital (savings, investments, and other reserves), collateral (security for the loan that could be used to recover losses in the event of default), and conditions (broader economic factors that could influence the borrower's repayment). More specifically, credit risk prediction models use mathematical calculations to estimate the probability that a borrower will default on their financial obligations within a specified time frame using metrics that are generally associated with the 5 Cs. This process is inherently probabilistic, as it seeks to model the uncertainty associated with future borrower behavior based on historical data. The goal is to provide lenders with a consistent and reliable measure of risk, enabling them to make informed decisions about loan approvals, pricing, and terms across a broad range of applicants.

At its core, credit risk prediction can be conceptualized as a function of borrower characteristics and financial behavior. The probability of default is modeled as:

$$Prob\ (Default) = f\ (Credit\ Bureau\ Data,\ Cash\ Flow\ Data,\ and/or\ Other\ Predictors)$$

Where:

- **Credit Bureau Data** includes traditional credit metrics such as payment history, credit utilization, and balances on different types of accounts. These variables provide a historical view of a borrower's past experiences with credit as well as their current utilization patterns. These variables are typically updated monthly, with information in a credit report reflecting activity as up to 45 days prior to when the credit report is pulled.[1]
- **Cash Flow Data** includes information drawn from deposit and transaction accounts such as patterns in balances, income, and expenses. This data offers a more frequent and granular view of a borrower's financial health, complementing the periodic updates of credit bureau data. By incorporating real-time or near-real-time transaction data, cash flow information can provide insights into a borrower's current financial stability as well as payments of recurring obligations such as rent, utilities, and telecommunication bills that are often not reported to credit bureaus.
- **Other Predictors** include other information relating to the 5 Cs of credit, such as information relating to income that the consumer reports on the loan application or information relating to collateral (e.g., loan-to-value ratios) or economic conditions.

The model aims to estimate the probability of default by identifying patterns and relationships between these predictors and the target variable, which involves various measures and definitions of loan delinquency or default depending on the lender's circumstances.

To operationalize the conceptual model, we used two different methodologies:

1. **Logistic Regression:** Logistic regression is a statistical method that has been used for decades to build underwriting models that predict the probability of a binary outcome (e.g., default/non-default) based on one or more input variables (features). It models the relationship between the input variables and the log-odds of the outcome, which is then transformed into a probability using a logistic function.[2] The log-odds are typically assumed to be a linear function of the input features, which makes it easier to evaluate the role that each feature is playing in the model. Developers also generally strive to select inputs that are not highly correlated to avoid uncertainty in attributing predictive power to individual variables.[3] As a result, logistic regression models are typically designed to be "parsimonious" by selecting input variables that add significant independent predictive power, while excluding those that may primarily be useful in specific cases or for specific subpopulations.

   To allow greater sensitivity to customer segments that may have substantially different financial situations and risk profiles (such as consumers who have filed for bankruptcy or who have relatively little traditional credit history, in contrast to consumers who have used credit extensively without incident), underwriting model developers often construct a separate "scorecard" or logistic regression model for each major segment within the applicant pool. Logistic regression serves as a strong baseline methodology for credit underwriting and is particularly effective when the relationships between predictors and the target variable are well understood.

2. **XGBoost:** XGBoost is a powerful machine learning algorithm that is particularly popular among underwriting model builders because of its speed, performance, and flexibility. XGBoost works by combining many simple models (usually regression trees) in a sequential way, where each new model focuses on correcting the mistakes of the previous ones.[4] Unlike logistic regression, XGBoost approximates the log-odds of the outcome as complex functions of the input features using sums of regression trees. This flexibility increases accuracy and allows the model to capture intricate patterns among large numbers of variables, but both the model architecture and large numbers of variables can make it harder to understand how individual features are influencing the model's predictions. To address this, lenders often constrain the model's complexity, use post hoc explainability tools such as Shapley values, or deploy a combination of both approaches to analyze how the model is making its predictions.[5]

The process of developing a credit risk prediction model involves several key steps:

1. **Data Collection:** Gathering historical data on borrower characteristics, credit behavior, and financial outcomes. This potentially includes both traditional credit bureau data and alternative data sources such as cash flow information.

2. **Feature Engineering:** Transforming raw data into meaningful predictors. This may involve creating derived variables (e.g., credit utilization ratio), handling missing data, and normalizing variables for model input.

3. **Model Development:** Winnowing down potential input features to the final list, training models, and finalizing their specifications. This process varies somewhat depending on the type of model. When developing a logistic regression model, typically a substantial portion

of the development effort is to find a small number of predictors and transformations to ensure that the final model is predictive and explainable. When developing machine learning models, a primary focus is typically on tuning hyperparameters to optimize the model's generalization performance. A detailed explanation of the model development process we employed is covered in Section C.4.

4. **Validation:** Assessing the model's predictive accuracy and fairness on unseen data. This includes evaluating metrics such as AUC-ROC (Area Under the Curve – Receiver Characteristic Curve), KS (Kolmogorov-Smirnov) Statistic, and fairness measures, as applicable for the specific use case.

C.2 Data

C.2.1 Data sources

The data used in this study is derived from a comprehensive anonymized dataset compiled by one of the three nationwide credit bureaus for use by third-party score developers. It combines historical credit bureau data with bank account information from a large data aggregator for substantial numbers of consumers who opened a new credit account between April 2018 and March 2019. To build the models, we used a traditional credit score or detailed historical records provided by the credit bureau (including traditional credit metrics such as payment behavior, credit utilization, and account types), bank account transaction and balance data from the aggregator (which offers insights into income, expenses, and balances that are not directly reflected in credit bureau records), or both sources, up to the month before the new credit account was opened. For the dependent variable and validation of the models, we used performance data on the newly opened account from the first 12 months after origination. These new tradelines provide a clear and objective measure of credit performance, such as the occurrence of serious delinquencies (90 days or more), charge-offs, or bankruptcies within the specified time frame.

While this dataset provided a robust foundation for our analysis, due to the way it was constructed it is not a nationally representative sample of consumers as discussed further below. The dataset was constructed starting on the data aggregator's side by focusing on consumers whose information included at least one bank account and one loan account as of April 2019. The loan account was used to link the aggregator data to the consumer's credit bureau profile using a hashing protocol to protect privacy. For simplicity, ambiguous matches—such as cases involving joint ownership of the loan used for linking—were excluded, leaving only unique one-to-one matches between the cash flow data and the credit bureau data. The final dataset included 750,266 consumers with new credit originations during the sample period, although we excluded a number of observations from the sample as described in Section C.2.2.

C.2.1.1 Credit bureau inputs

The credit report data provided by the credit bureau can be divided broadly into two categories: credit attributes/features that summarize a consumer's credit profile as of a snapshot date (Snapshot Attributes) and attributes/features that summarize variations across time for the same consumer (Trended Attributes) over a 24-month period up until the snapshot date. The combination of snapshot and trended attributes provides a comprehensive view of a consumer's credit history and behavior, and one or both types of attributes are widely used in traditional credit scoring and underwriting models.

CREDIT SUMMARY ATTRIBUTES AS OF THE SNAPSHOT DATE (SNAPSHOT DATA)

The package contains a suite of static attributes that summarize a consumer's credit file as of a snapshot date. These attributes are organized into 11 categories of credit information:

- **Inquiries:** Records of credit applications made by the consumer, including hard inquiries (resulting from credit applications) and soft inquiries (e.g., pre-approved offers).
- **Trades:** Detailed information on credit accounts, including open, closed, and historical accounts, such as credit cards, loans, and mortgages.
- **Past Due:** Records of late payments, including the number and severity of delinquencies (e.g., 30, 60, or 90+ days late).
- **Satisfactory Trades:** Accounts in good standing with no delinquencies or negative marks.
- **Worst Rate and Worst Trades:** Information on the highest interest rates and the most problematic accounts in the consumer's credit history, such as those with severe delinquencies or derogatory events.
- **Public Records:** Records of bankruptcies, foreclosures, tax liens, and civil judgments.
- **Collections:** Accounts that have been sent to collections due to non-payment.
- **Number of Trades Reported:** The total number of credit accounts reported to the credit bureau.
- **Percent of Trades:** The proportion of accounts in different statuses (e.g., open, closed, delinquent).
- **Past Due Major Derogatory Event:** Records of significant negative events, such as charge-offs or repossessions.
- **Trades with Major Derogatory Reported:** Number and proportions of accounts that have experienced major derogatory events.

TRENDED ATTRIBUTES

The trended attributes package provides dynamic attributes that summarize consumer behavior trends and credit characteristics over a 24-month period. These attributes leverage historical tradeline data to capture patterns in:

- **Spending Patterns:** Monthly credit card balances and payment behaviors, such as whether the consumer pays off their balance in full or carries a balance.
- **Credit Utilization:** Trends in the percentage of available credit used by the consumer over time, providing insights into their financial stability.
- **Wallet Share:** The distribution of credit usage across different types of accounts (e.g., credit cards, loans), reflecting the consumer's credit portfolio composition.

In addition to the credit attributes/features, we obtained a conventional credit score (referred to as the credit score or CS) for the consumers in the sample calculated using credit bureau inputs as of the month-end before the new origination. We used these scores as inputs for some hybrid models and as benchmarks for assessing the performance of the models we built for the study.

C.2.1.2 Bank account history

The data aggregator files include transaction and balance history for various types of bank accounts, including checking, savings, certificates of deposit, money market accounts, flexible spending accounts, health spending accounts, and prepaid cards, covering the period from April 2016 to February 2019. Not all consumers may have linked all of the accounts they owned. The account history files for the individual accounts provide the account type and periodically refreshed balances, along with the refresh dates. The transaction history files include the transaction type, date, amount, currency, and a description that is provided by the underlying financial institution (such as the transaction description that appears on a checking account statement).

The aggregator also provided its own categorization of the transactions, based on models that it has built based on the other data elements from millions of transactions. The aggregator's classification scheme includes 64 categories designed to provide insight into the source or use of the funds involved. For example, a debit transaction could be classified as a purchase, loan payment, rent, etc., depending on the description and parties involved. Similarly, a credit transaction could be classified as a deposit, salary, refund, etc. Such characterization of the sources and uses of funds sheds light on the nature and stability of a consumer's income and liabilities, which can be valuable for predicting credit risk.

C.2.1.3 Implementation of the dependent variable

Since our models are designed to predict the probability of default, it is critical to define and implement a default definition that captures all material credit losses. For this study, we defined a default event as one or more of the following occurrences on the new account (or one of the new accounts) within 12 months of origination:

- **90+ Days Delinquency:** The account has been delinquent for 90 days or more, indicating a significant delay in payment.
- **Charge-Off:** The lender has written off the account as a loss, typically after 180 days of delinquency.
- **Repossession:** The lender has repossessed collateral (e.g., a vehicle) due to non-payment of the account.
- **Foreclosure:** The lender has initiated foreclosure proceedings on a property due to a default on the mortgage account.
- **Bankruptcy:** The account has been included in the borrower's filing for bankruptcy, which may include Chapter 7, Chapter 11, or Chapter 13 bankruptcy.

These events are widely recognized as indicators of severe credit distress and are commonly used by lenders and credit bureaus to assess credit risk.[6]

With regard to choosing a performance window, model developers typically use between 12 months and 24 months. The selection is driven by the consideration of multiple factors, including.

1. **Whether most payment defaults emerge during the performance window.** Ideally, to allow a model to recognize signals for all defaults, one should wait until all defaults have emerged and then proceed with the model development effort. But given the other factors discussed below, in practice most model developers aim for a performance period in which 70% to 90% of default events have emerged for a particular product.
2. **Representativeness of development sample.** As the environment in which a model operates is constantly changing, it is unrealistic to expect a model developed using a snapshot from 10 years ago to be able to accurately predict credit risk in today's demographic and economic environments. As a rule, modelers always aim for the sample that is most representative of the environment in which the model is expected to produce predictions. Under normal circumstances, this means that using a more recent sample to develop the model is preferrable to using an older sample because it limits the amount of drift in the operating environment for the model.

3. **Time decay of predictive power of the input features.** A third consideration is the effect of time between the prediction and the default events on the predictive power of the model. An innate characteristic of an underwriting model is to use information available at the time of a loan application to predict the applicant's payment behavior in the future. The further into the future, the more likely that events that occurred after the application have prompted the observed default. A longer performance window therefore introduces additional noise to the target variable. This consideration leads to modeler's focus on finding the number of months on book when the instantaneous default rate peaks and use that as the performance window.

In practice, all three factors are considered when selecting a performance window. For this study, we chose 12 months because it allowed us to avoid the risk that COVID-19 impacted the performance data and because it is generally consistent with industry practice, which most commonly uses between 12 and 24 month performance periods for general purpose underwriting scorecards.

Another important consideration for the target variable is whether it is defined at the account level or the consumer level in situations in which a consumer opens more than one account in a given month.

- **Account-Level Model:** Predicts the probability of default for a specific credit account. This approach is useful for lenders evaluating individual loan applications, as it provides a risk assessment tailored to the specific loan being considered.
- **Consumer-Level Model:** Predicts the probability of default for a consumer across all their new credit accounts. This approach is more holistic and aligns with the goal of assessing overall borrower risk. It is particularly relevant for lenders managing portfolios of loans or evaluating borrowers with multiple credit relationships.

While the account-level model offers granularity, the consumer-level model provides a broader perspective on credit risk. In practice, the choice between these approaches depends on the lender's objectives and the context in which the model will be applied. In our data, the consumer-level definition is used with the added nuance that the same consumer could appear more than once in the sample if he/she opened one or more credit accounts in multiple months between April 2018 and March 2019. About 58% of the consumers used in the final sample as described further below appeared only once in the data. "Consumer" therefore designates an individual consumer at a specific time. While this method introduces some level of correlation between observations for the same consumer, it allows us to take advantage of changing circumstances of the same consumer over time to increase the sample size for model selection and estimation.

To ensure that the target variable is based only on performance data from relevant and actionable accounts, several exclusion criteria were applied at the account level before the consumer level rollup was created. These criteria removed performance data from accounts that did not represent genuine credit obligations or that were associated with special circumstances that could bias the model's performance assessment, including:

- **Authorized User Accounts:** Accounts where the consumer is an authorized user were excluded, as these do not represent the consumer's direct credit obligations.
- **Lost or Stolen Accounts:** Accounts reported by the lender as lost or stolen were excluded. They typically are replaced by another account with payment history ported over.

- **Disputed Accounts:** Accounts whose veracity or information accuracy is disputed by the consumer. These are typically excluded from being considered by an underwriter until the dispute is resolved.
- **Natural Disaster-Affected Accounts:** Accounts that have been flagged in the credit bureau records because the borrower was affected by a natural disaster.
- **Specific Account Types:** Certain account types were excluded due to their nature or the fact that they do not represent standard consumer credit obligations. These include:
  - Factoring Company Account (debt purchaser)
  - Business Loan (individual personally liable)
  - Family Support
  - Commercial Installment Loan (individual personally liable; company is guarantor)
  - Commercial Mortgage Loan (individual personally liable; company is guarantor)
  - Business Credit Card (individual has primary responsibility)
  - Deposit Related (overdrawn account)
  - Medical Debt
  - Child Support

These exclusions ensure that the target variable as defined is directly related to the consumer's credit behavior and is indicative of their ability to manage personal credit obligations.

C.2.1.4 Demographic data

For the purpose of analyzing model performance and its impact on underwriting outcomes across different demographic groups, we also obtained a separate dataset from the credit bureau that includes inferred information about race/ethnicity. The dataset, which the credit bureau provides for non-underwriting purposes, enabled us to evaluate the impact of the model on different subgroups.

The race/ethnicity field determines ethnicity by analyzing first and last names using ethno-linguistic rules, geographic data, and cultural patterns. It identifies unique naming conventions (e.g., prefixes like "Mc" for Irish or suffixes like "NEN" for Finnish) and combines them with geographic clustering (e.g., ZIP+4 codes). The process handles multi-ethnic names, hyphenated names, and exceptions, and the credit bureau calculates that it achieves 94% accuracy across over 130 groups through a rule-based, geo-centric approach. The credit bureau did not provide names or geographic information to FinRegLab.

C.2.2 Sampling methodology

In general, whether a consumer is included in the development sample for a credit underwriting model depends on three factors: (1) the availability of the dependent variable, (2) the availability of data for predictive features of interest, and (3) the representativeness of the environment in which the model will be used. For the dataset we were working with, these factors manifested as the following.

1. **Availability of the Dependent Variable:** The population was limited to consumers with tradelines in their credit reports that can be used to identify the target variable, i.e., newly originated tradelines between April 2018 and March 2019 with payment history within the first 12 months after origination.
2. **Availability of Bank Account History:** One of the core objectives to be fulfilled by the dataset is to support the development of cash flow-based underwriting models. As such, the population was limited to include consumers whose aggregator records included at least one bank account and at least one loan account that could be successfully matched to credit bureau data.
3. **Avoiding the Impact of COVID-19:** The economic interventions and relief measures instituted by governments in response to the COVID-19 pandemic could have created noisy signals. To ensure that the model is not influenced by these atypical conditions, the study focuses on data from before the pandemic.

From the baseline data set using this approach, we made a number of exclusions based on data quality, completeness, and other reliability considerations:

1. **Messy credit bureau matches:** Due to complicating factors in credit reporting—such as co-borrowing, personal loan guarantees, and authorized usage—the matching between the aggregator records and credit bureau files was not always one-to-one. There were instances of one-to-many, many-to-one, and many-to-many matches. These ambiguous matches could introduce noise and inaccuracies into the dataset, potentially affecting the reliability of the analysis. After careful consideration, we decided to remove all ambiguous cases and retain only unique matches. By focusing on unique matches, we (1) ensure that the credit attributes and cash flow data are accurately linked to the same consumer, enabling a fair comparison of their predictive power, and (2) prevent matching errors from becoming a source of noise for the dependent variable. The result dataset included 750,266 consumers with new originations from April 2018 to March 2019.[7]
2. **Consumer records without credit scores for benchmarking:** Approximately 2500 consumers in the sample did not have a credit score under the benchmark model. Because that population was too small to allow reliable study in any event, we excluded those consumers from the sample. This allowed us to benchmark probability of defaults and other metrics for all sample measures.
3. **Consumers whose bank account data did not meet basic quality standards:** As discussed in greater detail in Section C.3.1, using the bank account data required pre-processing of the transaction and balance information to construct a coherent history of the individual accounts. For about 90,000 consumers, the available information was too limited for us to build a reliable history for any of their bank accounts. We therefore excluded those consumers from the sample.
4. **Consumers whose bank account data did not include 6 months of history on a "primary checking account."** To ensure that the cash flow information included a transactional account that was being used relatively heavily on an ongoing basis, we defined criteria to identify what we called "primary checking accounts."[8] We excluded consumers who had less than 6 months of history on such accounts—about 230,000 consumers in all—to ensure that the findings concerning cash flow data are representative of a real-world underwriting scenario where bank account data is provided.

PRIMARY CHECKING ACCOUNT DEFINITION

After extensive analysis of checking account data in the sample, we defined primary checking accounts by using four criteria, all of which must be met to remain in the sample:

- **At least 6 months of continuous history stretching back from the end of the month prior to the new loan origination month:** This aligns with common practices in credit risk modeling, where a sufficient history of financial behavior is necessary to assess creditworthiness accurately. It was the most impactful criterion, accounting for the exclusion of 34% of checking accounts from being considered primary.
- **At least 2 deposits in every 31-day window within the last 6 months:** Regular deposits are a key indicator of an account's use as a primary transactional account. Deposits typically represent income sources, such as paychecks or other recurring inflows, which are essential for maintaining financial stability. Among checking accounts with 6 months of history, about 14% did not meet this criterion.[9]
- **At least 5 payments in every 31-day window within the last 6 months:** A primary checking account is expected to be used for frequent payments, such as bills, purchases, and other everyday expenses. Among checking accounts with 6 months of history, about 10% did not meet this criterion.
- **At least $500 in total payments in every 31-day window within the last 6 months:** The total amount of payments made from an account provides insight into the account holder's spending habits and financial obligations. Requiring at least $500 in total payments in every 31-day window ensures that the account reflects meaningful transactional activity, consistent with the use of a primary checking account. Among checking accounts with 6 months of history, about 10% did not meet this criterion.

For each consumer in the sample, we used credit bureau information and bank account information up to the month before the origination to derive the input features. This dataset serves as the foundation for model development, ensuring that the sample is representative of typical credit risk conditions and is suitable for evaluating the impact of cash flow data on credit risk prediction.

These exclusions had some impacts on the makeup of the final sample distribution. For example, the decision to focus on only consumer profiles with one-to-one matches between the aggregator and credit bureau data excluded some consumers with higher incomes and credit scores. The decision to focus on only consumers with accounts that met our definition of a "primary checking account" excluded some consumers with lower incomes. Such tradeoffs are inherent in model development, and our sensitivity analyses suggest that the impacts were minimal. Such residual risks are typically managed by industry model builders by monitoring model performance after deployment to determine how the model performs with regard to subgroups and to detect whether applicant populations are shifting over time.

C.2.2.1 Data partitioning

After applying these exclusions, we arrived at a final sample of 424,546 observations for model development and validation. We partitioned the data as follows:

- **Out-of-Time Validation Sample:** Approximately 25% of the data was set aside as a "stratified" out-of-time validation sample. This includes consumers with originations in July 2018, November 2018, and March 2019. This sample was not used in any part of the model development or validation process, and instead was used to calculate the performance and credit access metrics discussed in Section 4 as a measure of the extent to which the models are generalizable to unseen data.
- **Model Development and Validation Samples:** The remaining 75% of the data was randomly split into a model development sample and a validation sample in a 2:1 ratio. The model development sample was used for feature selection, hyperparameter tuning, and estimating the models, while the validation sample was used to validate and compare models before finalization.

TABLE C.1 SAMPLE WATERFALL

CATEGORY	# OF OBSERVATION	% POPULATION
TOTAL POPULATION	750,266	100.00%
NO CREDIT SCORE	2,588	0.34%
LOW QUALITY BANK ACCOUNT DATA	90,342	12.04%
INSUFFICIENT PRIMARY CHECKING ACCOUNT DATA	232,790	31.03%
SAMPLE FOR MODELS	424,546	56.59%

TABLE C.2 DATA PARTITIONING

SAMPLE	# OBSERVATIONS	# DEFAULTERS	PROB. OF DEFAULT
OUT-OF-TIME VALIDATION	107,789	2,468	2.29%
VALIDATION	105,520	2,398	2.27%
DEVELOPMENT	211,237	4,706	2.23%

While it is a common practice to validate model performance on samples that have not been involved in the model development or validation process, using an out-of-time sample for this purpose is typical among banks but may not be as universally used by other lenders or data scientists in other settings. The approach is designed as an extra check to ensure that results are robust and generalizable, although as discussed in the Section C.2.3 we faced certain limitations due to the onset of COVID-19 pandemic.

C.2.3 Data limitations

While the dataset used in this study provides a robust foundation for evaluating the impact of cash flow data and machine learning techniques on credit risk prediction, it is not a nationally representative sample and has limitations due to the nature of the data sources, the sample selection process, and the timing of the data collection. Below, we outline the key limitations and their implications for the study:

C.2.3.1 Limited representation of thin and no-file consumers

Because the sample selection process focused on consumers who were using the data aggregator's services and had both an existing bank account and credit account, the sample includes relatively few consumers whose credit bureau data is so limited that they may have a difficult time being scored by traditional models. These include three primary categories of consumers:

- **No file/credit invisible consumers:** Consumers who do not have a credit profile at all.
- **Consumers who are unscoreable under various third-party models:** The percentage of such consumers differs by model type. For example, FICO's general models will not generate scores for consumers with no trade activity older than 6 months or no trade activity in the past 6 months, while VantageScore will generate scores unless there are no trades, public records, or unpaid collections reported. General scores offered by individual credit bureaus and specialty scores may have other parameters.
- **Thin file:** Consumers with two or fewer credit accounts (often called tradelines) on their credit reports may be scoreable under one or more third-party models, but their likelihood of default is often more difficult to predict than consumers who have more extensive main-

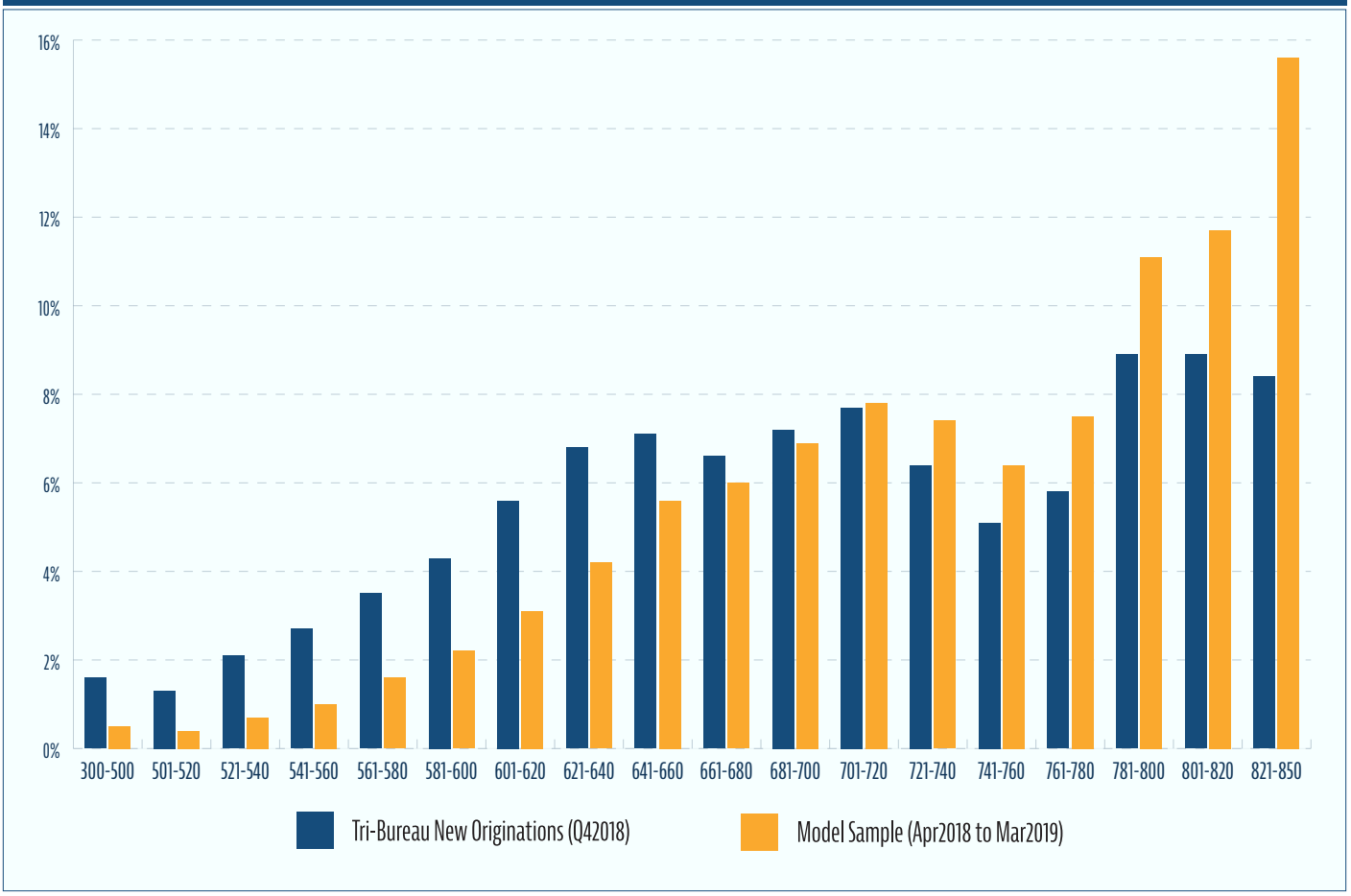
stream credit experience. Lenders may subject thin file consumers to special requirements or higher pricing due to risk and fraud concerns.

This limitation restricts our ability to assess the predictiveness and credit access impacts of our models on these specific subpopulations. In particular, it prevents precise estimates of the impacts on consumers who are likely to see the largest positive impacts from using cash flow data for credit underwriting. In this respect, our results may not fully account for the impacts on credit access for certain populations that frequently struggle to access mainstream credit products.

C.2.3.2 Skew toward prime borrowers

The sample selection process also skews the overall sample toward borrowers with prime credit scores (generally 660 to 680 depending on the particular score used) and higher incomes compared to the overall tri-bureau new originations population distribution from the same time period. In addition, the exclusion of consumers without a primary checking account slightly added to the skew. Although there is adequate representation across the entire credit spectrum as reflected in Figure 5 to support model development and analysis, the measurement of impacts on populations with higher scores and incomes may tend to be the most precise.

FIGURE 5 SAMPLE DISTRIBUTION BY CREDIT SCORE



C.2.3.3 Validation sample constraints

To ensure that the sample and performance data were unaffected by the economic disruptions caused by the COVID-19 pandemic, we restricted the sample period to originations before March 2019 so that we could observe 12 months of performance. When the sample was compiled, originations from after the end of COVID-19 restrictions had not accumulated 12 months of performance. We therefore did not obtain a separate stable "out of time" sample to use for validation, and instead held back data on consumers whose originations fell in every fourth month of the performance period as described in Section C.2.2. As a result, the validation sample is nearly contemporaneous with the model development sample, rather than being drawn from a later period as would be typical in a real-world setting.

This stratified out-of-time approach is commonly used by industry modelers faced with similar data constraints for an initial assessment, but nonetheless limits our ability to fully assess the model's performance under changing economic conditions. As noted in Section 1, the use of a stratified out-of-time sample tends to generate higher performance metrics. It may also impact comparisons to traditional credit scores that were developed in a different time period and applied to the study sample.

C.2.3.4 Absence of rejected applicants

Our sample consists only of consumers who opened a new credit account between April 2018 to March 2019 and does not include consumers who experienced rejections on all applications that they submitted during that same period. This introduces selection bias, since we lack evidence as to how those consumers would have performed on new accounts. Lack of data on rejected applicants is a well-documented limitation in observational credit risk modeling, absent experimental or reject-inference adjustments. Different industry actors take a variety of approaches to reject inference concerns, such as:

1. **The null approach** (using only approved applications): This assumes no systematic difference in predictive patterns between approved and rejected applicants in which case using approved applications only does not introduce model bias.
2. **Hard cutoff augmentation:** This classifies rejected applications as likely to have paid or not paid a hypothetical loan based on the predictions of an inference model. The inference model is typically estimated on approved applications using either all information available at the time of application or post-application information obtained from a credit bureau up until the end of the performance window.
3. **Fuzzy augmentation:** This duplicates records of rejected applications and assigns them different predicted probabilities of default using an inference model. The first record is assigned the predicted PD and the second is assigned 1 minus the predicted PD, which effectively creates a range of outcomes that reflects the degree of probabilistic uncertainty for particular applicants.

Because there is limited experimental evidence with regard to these approaches, the effectiveness of these methods is a subject of debate within the industry and study among academic researchers as discussed in Appendix A.

In our study, we did not attempt to augment our dataset to include consumers that could have been rejected during the sampling period. We believe the practical impact of the sample selection bias on our study is mitigated by two factors:

- **Consumer-level risk assessment:** By analyzing all new originations per consumer, we are able to incorporate performance history for consumers who were rejected by some lenders but approved by others. However, while this captures partial rejections, it does not account for consumers rejected by all lenders.
- **Broad credit spectrum coverage:** Our sample includes sufficient representation across all credit score bands, unlike the situations faced by some individual lenders that use a score cutoff that requires them to consider extrapolating into a risk tier for which they have no actual observations. However, while we have some visibility, it is possible that rejected consumers within the same score band might behave differently.

As discussed further below we also weighted the sample to adjust the ratio of cases that did and did not involve defaults and in performing simulations of the impact on credit access to account for the lack of a representative distribution.

C.3 Cash flow feature engineering

The integration of cash flow data into credit risk prediction models necessitates a robust framework for feature engineering that captures critical dimensions of consumer financial behavior while adhering to regulatory and practical constraints. Unlike credit bureau data, where predictive features (e.g., payment history, credit utilization) have been distilled into standardized metrics over time, bank statement data are provided as raw transactional records. To leverage this data in a predictive model, we first engaged in pre-processing to match transaction and balance information so that we could understand the account history, as discussed below. We then developed features representing meaningful aggregates of transaction and balance history, focusing on solvency, liquidity, and behavioral patterns.

We took some account of compliance considerations in this process as described below, although we have not conducted all of the analyses that lenders would typically perform for compliance purposes in preparing a new model for deployment. In addition, while we spent substantial time on feature engineering activities, our techniques and uses of the data may differ from other model developers and researchers. Given the relatively short amount of time that developers have been working with electronic cash flow information. Accordingly, practices vary as we and other model builders continue to refine tools and strategies for processing the data and distilling predictive insights.

C.3.1 Preprocessing of cash flow data

The cash flow data used in this study required significant preprocessing before it could be used to derive cash flow features. Due to the nature of how the data is captured and aggregated from multiple financial institutions—each with its own reporting practices—we focused on addressing two primary issues: (1) identifying gaps in account history and dividing it into coherent "episodes" where all transactions are reported, and (2) imputing a complete daily balance history for each episode. Below, we outline the key steps taken to preprocess the data.

C.3.1.1 Identifying account episodes

Data aggregators pull bank account information on behalf of other financial services providers who use the information for a wide range of purposes. The scope of these snapshots and the frequency with which they are refreshed depends on the authorization granted by the consumer for account access and the consumer's use of the end user's services, which may involve personal financial management tools, loan underwriting, payment services, or various other use cases. As a result, data aggregators' records related to individual consumers vary in their scope and completeness. To address this, we had to compare transaction and balance history in order to construct a structured picture of the account history. The first step was to develop logic to identify "episodes"—continuous time windows within which we are confident that all transactions are fully reported.

- **Data Refresh Mechanism:** The cash flow data vendor refreshes account data based on user activity. When a consumer first authorizes access to an account, the aggregator typically obtains at least 90 days of transaction history and one balance record. Refreshes may occur daily, biweekly, or weekly, depending on the consumer's agreement with the end user but will cease after 90 days if the end user does not actively pull additional data on behalf of the consumer. This mechanism can lead to gaps in the data, as not all transactions within the sampling period may be captured. It also means that we have more complete information for consumers who are using end users' services on a more continuous basis.
- **Episode Identification:** By leveraging the refresh dates captured in the data and confirming with the vendor that each refresh brings at least 90 days of history, we identified coherent episodes within each account's history. These episodes represent time windows where we think we have a complete picture of transactions.

C.3.1.2 Imputing daily balances

While the episode records provided a complete picture of transactions, they only list the account balance as of the date on which the information was refreshed rather than end-of-day (EOD) balances for each day within the snapshot. However, once episodes were identified, we developed a procedure to impute the missing information, ensuring a complete daily balance history for each episode. This process involved several steps:

- **Valid Inference Date Range:** We calculated the maximum difference between the file creation date and the transaction date at the account level. This allowed us to establish a valid inference date range, assuming that the account provider consistently supplies the same amount of transaction history upon each refresh.
- **Ambiguous Balance Records:** To mitigate ambiguity, we eliminated any dates associated with more than one balance record, leaving only balance records associated with a unique reporting date as our anchor for imputing the EOD balance for the dates with no reported balance.
- **Filling Missing Balances:**
  1. We created a Cartesian product of unique account IDs and dates to generate a complete timeline for each account.
  2. We identified reported EOD balance by treating a reported balance as the EOD balance if no transactions occurred on that date.

3. Missing EOD balances were filled by subtracting cumulative daily flows from the latest available EOD balance (backward path) or adding daily flows to the last available EOD balance (forward path), ensuring no gaps in reported transactions.

C.3.1.3 Validating imputed balances

To ensure the accuracy of the imputed balances, we implemented a validation process to determine whether an account's imputed balance history was consistent with the raw transaction and balance records. This process relied on two performance metrics:

- **Kendall's Tau-b:** This statistic measures whether imputation errors—the difference between an actual balance and the corresponding imputed balance—exhibit a time trend. A high absolute Tau-b value between the imputation errors and their timestamps indicates a strong time trend, which may suggest systematic inconsistencies between the transaction flows and the changes in balance for an account over time. Such inconsistencies could arise from unreported transactions or other data anomalies.
- **Mean Absolute Percent Error (MAPE):** This metric assesses the accuracy of the imputed balances where actual balance records are available. A low MAPE indicates high accuracy, meaning the imputed balances closely match the reported balances.

We accepted episodes based on the following criteria:

- **High Tau-b, Low MAPE:** For episodes with high absolute Tau-b values, we accepted the imputation only if the MAPE was low. While high Tau-b values suggest potential unreported transactions or systematic inconsistencies, accounts with low MAPE are typically Certificate of Deposit (CD) accounts without interest accrual transactions. In these cases, the low MAPE indicates that the imputed balances are still accurate despite the time trend in errors.
- **Low Tau-b, High MAPE:** For episodes with low Tau-b values, we tolerated higher MAPE values, as the absence of a time trend suggests random errors rather than systematic issues. Accounts with low Tau-b and high MAPE are typically transaction accounts with end-of-day (EOD) balance sweeps into a different account. The high MAPE in these cases is primarily due to timing mismatches between the imputed balance and the actual balance record, rather than systematic errors.

This validation process ensured that the imputed balance histories were both accurate and reliable, providing a solid foundation for deriving cash flow features. By distinguishing between systematic and random errors, we were able to retain episodes with acceptable levels of accuracy while excluding those with significant inconsistencies. Note that we only removed stretches of account history that were deemed incomplete and/or inaccurate based on the tests described here. High quality episodes for the same account or other accounts owned by the consumer were preserved. If a consumer had any high quality episode on any account, we are able to construct the cash flow features described in the rest of this section.

C.3.1.4 Handling other data quality issues

While the primary focus was on identifying episodes and imputing balances, we also addressed other data quality issues arising from idiosyncratic practices by individual financial institutions. To name a few examples:

- Some institutions adjust account balances to negative numbers to prevent automatic check clearing for accounts with suspicious activities.
- Some institutions report authorization transactions that do not affect account balances.
- Some institutions accrue interest in account balances without creating corresponding credit transactions.

These issues were not prevalent across all data providers and typically affected only a small number of accounts. Where possible, we corrected the data based on our understanding of the underlying cause. In cases where corrections were not feasible or the issues were immaterial, we removed the offending transactions.

The preprocessing steps outlined above were critical for ensuring the reliability of the cash flow features derived from the data. By identifying coherent account episodes and imputing complete daily balance histories, we established a robust foundation for feature engineering. These steps ensured that the cash flow features were based on (1) account periods with complete transaction histories and (2) accurate daily balance records, enabling more precise and reliable credit risk assessments.

C.3.2 Basic cash flow features

As described in Section C.2.1, the dataset included balance history, account types (e.g., checking, savings, certificates of deposit), and transactional records, such as amounts, directions (debit/credit), currencies, descriptions provided by the bank or financial institution where the account is held, and the aggregator's standardized categorizations. We initially developed a basic set of cash flow features using only structured data fields, without using the institutions' transaction descriptions or the aggregator's categorizations. The transaction descriptions are more complicated to process and interpret because their contents are less standardized, and some lenders may prefer not to use those fields for simplicity when generating adverse action disclosures for individual consumers. Three primary categories of features were created to assess both long-term stability measures and shorter-term liquidity.

C.3.2.1 Long-term stability features

Solvency metrics were designed to evaluate long-term financial stability. These included net worth (total balance divided by credit bureau-reported debt), total assets (sum of account balances), and debt-to-asset ratios. Income and expense proxies were derived from median monthly inflows and outflows, respectively, while savings rates were calculated as the ratio of median monthly net outflow to median inflow. Savings and investment metrics, such as balances in certificates of deposit and money market funds where reflected in the data, provided additional insights into financial reserves.

C.3.2.2 Short-term liquidity features

Short-term financial stability was assessed through liquidity indicators. The current ratio (balance divided by median monthly outflow) and expense-to-income ratio (median outflow divided by median inflow) quantified cash flow adequacy. Cash and liquid asset totals were aggregated across accounts to measure immediate resource availability.

C.3.2.3 Other financial features

Behavioral features captured dynamic financial habits and stress signals. Metrics such as time since the first account activity were computed by account type. Transactional activity was quantified through recency, frequency, and ticket size, while affluence was inferred from the number of large purchases. Financial stress indicators included counts of days with negative balances and deviations from historical balance trends. Trend analyses, such as current-to-average balance ratios and percentile rankings of recent balances, provided temporal insights into financial stability.

C.3.3 Advanced feature engineering

While the basic features were exclusively based on structured data fields, we spent additional time and effort to develop an advanced feature set using information derived from either the transaction descriptions or the categorizations provided by the aggregator to identify more nuanced patterns in regular and irregular income and expenses as well as liquidity events. This process involved extensive data analysis to inform reasoned judgments about how to define key concepts and classifications and extract information from less structured data fields with an eye toward ensuring that the features would be verifiable and justifiable, particularly in light of adverse action requirements.

To prevent internal transfers from inflating debt and income calculations, internal transfers were identified and removed by matching debit-credit pairs of identical amounts occurring within ±3 days across a consumer's accounts. This threshold balanced completeness (capturing weekend processing delays) and precision, minimizing over-exclusion.

C.3.3.1 Features related to recurring transactions

Recurring transactions in a consumer's bank statements are of particular interest because recurring credit transactions often represent regular income, while recurring debit transactions typically reflect debt payments or non-discretionary liabilities. This may provide a more nuanced picture of the consumer's capacity to take on additional obligations than consideration of only annual income and debts appearing on traditional credit reports.

Through extensive case reviews of cash flow data, we defined criteria for identifying transactions that were sufficiently similar that we classified them as recurring transactions. To start, we focused on transactions within the same account that occurred at regular intervals (weekly, biweekly, or monthly—common pay and billing cycles) with a minimum frequency (at least one three-month period with at least three similar transactions) and an amount exceeding $50. Three similarity definitions were applied:

- **By Transaction Amount:** Transactions with identical amounts. In addition, we created a separate version by excluding internal bank transfers and fees.
- **By Transaction Description:** Transactions descriptions that indicated they involved the same source, excluding bank transfers and fees.
- **Hybrid Criteria:** Transactions meeting either the amount- or description-based similarity criterion, with bank transfers and fees filtered out.

C.3.3.2 Classification of income and expenses based on aggregator transaction categorization

We also used the aggregator transaction categorizations as a second approach to distinguishing between the regularity of income sources and the degree of discretion over types of expenses. For

example, we constructed debit transaction definitions based on the aggregator categories that were structured to progress from core obligations (e.g., loans, rent) to expanded categories (e.g., utilities, insurance), and indirect obligations (e.g., checks, credit card payments) to all outflows. This approach moved from narrowly defined, highly predictive categories with limited transaction coverage to broader, more inclusive definitions that encompassed most transactions, albeit with some ambiguity. To improve signal-to-noise ratios, we applied a $50 threshold to exclude insignificant transfers and checks. Similarly, income proxies ranged from conventional sources (e.g., salaries, retirement benefits) to expanded, less regular categories (e.g., tax refunds), to total inflows.

C.3.3.3 Liquidity events

We also used the financial institutions' description fields to identify indications of overdrafts and insufficient funds (NSF) through keyword patterns (e.g., "NSF FEE"). Recognizing the widely different practices in the frequency of a check being re-presented and the amount of fees charged by different financial institutions, we only count the number of days an NSF event occurred and the time since the last NSF event to minimize idiosyncratic information content in these features. This information complemented the liquidity signals we extracted from the first stage of feature engineering by analyzing raw balances and other structured fields. The first stage information alone may not be sufficient to detect returned checks or overdrafts that are covered from linked accounts.

C.3.4 Final feature set

For this study, we ultimately decided to use both the simple and advanced features, computing them across account types (checking, savings, other) and various time windows (30–365 days). Key variables included net flow (credits minus debits), credit/debit sums, and ratios of recurring debt payments to income. Liquidity stress was further quantified through recency metrics, such as days since the last overdraft or sub-threshold balance. The feature engineering process yielded a comprehensive suite of 1,976 features.

Both sets of features were also made available for the cash only and hybrid logistic regression models to select from. Following the process described in Section C.4.1, those final models typically use 10 to 20 features covering areas such as low or negative balance events, stability of cash inflows, balance trend over time, number and amount of recent cash outflows, and debt-to-income proxies.

C.4 Model development

To address the research objectives, we developed eight distinct models, each employing different methodologies and input configurations. These configurations included: (1) cash flow data only, (2) cash flow data combined with a traditional credit score, (3) cash flow data combined with credit bureau data, and (4) credit bureau data only. For each configuration, we constructed both a logistic regression model—a traditional approach widely used in consumer loan underwriting—and an XGBoost model—a machine learning algorithm known for its ability to capture complex, non-linear relationships. and an XGBoost model. We used the traditional credit score as a general benchmark for comparative analysis across all of the models.

Following a common practice to balance the classes in the sample, we applied sample weights to adjust the ratio of "good" to "bad" cases to 3:1 during the model development process for both LR and ML models. This approach is standard in credit risk modeling, as it enhances the model's ability to distinguish between default and non-default cases while maintaining stability during training. The 3:1

ratio strikes a balance between emphasizing the minority class (defaults) and leveraging the majority class (non-defaults), aligning with the business objective of minimizing default-related losses.

C.4.1 Logistic regression models

A typical development process for logistic regression models can be characterized as an expert-guided dimension reduction process, where the model developer's knowledge of the target relationship heavily influences the final model specification. The process often begins with an exploration of whether the population should be segmented into subpopulations, ensuring that a linear function within each segment adequately captures the relationship between predictors and the target variable.

Segmentation may be informed by domain expertise, statistical methods, and exploratory data analysis. For each segment, the model developer then narrows down the predictive features to a small set, balancing automated techniques (e.g., regularization) with expert judgment to maximize generalization performance. The entire process is iterative, involving cycles of feature engineering, selection, and validation until an optimal tradeoff between predictive power and intuitive appeal is achieved. This balance is particularly important in underwriting, where model interpretability and regulatory compliance are critical. In this section, we discuss the process followed for developing the logistic regression models used in this study.

C.4.1.1 Segmentation

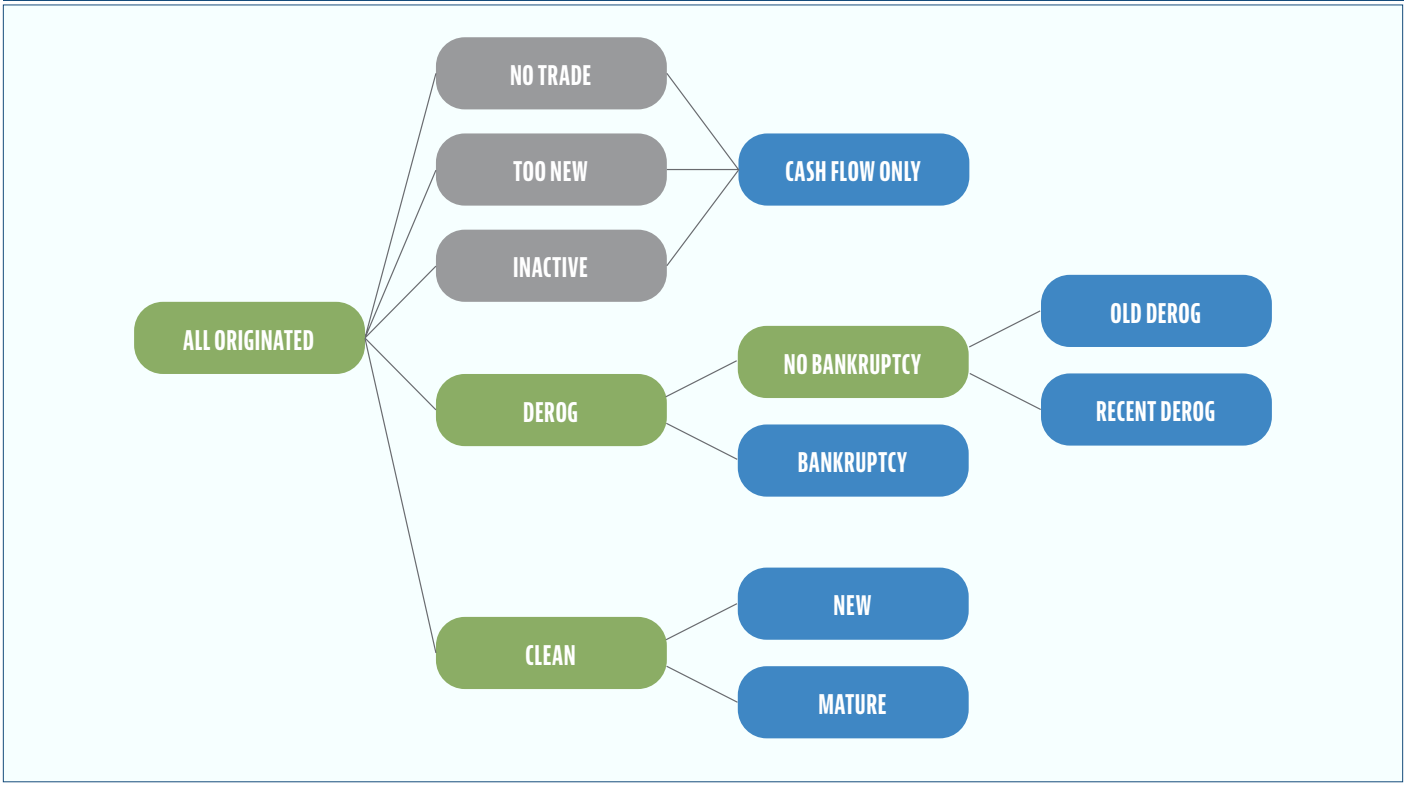
Underwriting risk models/scorecards are typically segmented into a few subpopulations, such that a separate regression model or scorecard is developed for each of the mutually exclusive segments. Segmentation is frequently used in building traditional underwriting models for a mix of statistical and operational considerations. Statistically, it accounts for the possibility that the same feature may have different effects across segments, reducing the need for complex interaction terms. By dividing a population into distinct subgroups (e.g., by product type, risk tier, or geography), models tailored for each segment can capture segment-specific behaviors—such as varying impacts of debt-to-income ratios on subprime versus prime borrowers. This avoids the pitfalls of a 'one-size-fits-all' model, where critical interactions may be missed or obscured by averaging effects. Operationally, segmentation can help simplify model specifications, particularly for subgroups where certain features are irrelevant. For example, a model for consumers with no prior delinquencies can exclude delinquency-related variables, simplifying interpretation.

The segmentation scheme is typically determined judgmentally by experienced developers, balancing the need for simplicity and accuracy while considering physical constraints such as sample size in a branch of the segmentation tree. Resources permitting, segmentation schemes can be tested empirically via a parent-child analysis, where the performance of a suite of "child" models (individual models for different segments within a branch of a segmentation tree) is compared against the "parent" model (one model estimated on the combined sample with all "child" segments) to ascertain whether it is desirable to have a suite of child models or just one single parent model at any node of the segmentation tree. Due to resource constraints, we did not perform this kind of segmentation analysis. Our segmentation schemes for the logistic regressions mainly borrowed from industry experience on credit risk models based on credit bureau inputs, focusing on simplicity, interpretability, and the validity of statistical inference (sample size).

For the two logistic regressions using credit bureau inputs, we adopted a segmentation scheme similar to those used by many third-party credit scores. This scheme categorizes consumers based

on credit history characteristics, such as the presence of derogatory events, credit utilization, and trade age, and is widely recognized as conducive to producing parsimonious scorecards within each segment and encapsulating the main behavior segments as differentiated by risk drivers. As an example, the scorecard for a consumer with a derogatory history is likely heavily reliant on the type, amount, frequency, and timing of the derogatory events, whereas a scorecard for a consumer without a derogatory history tends to be heavily influenced by credit utilization and proxies of the debt-to-income ratio. Figure 6 illustrates this segmentation scheme in detail. For both the credit only and credit + cash hybrid LR models, the population with sufficient credit bureau information were divided into 5 segments. The hybrid LR model has a sixth, cash only segment to cover the population with very limited credit bureau information. This approach ensures that the models are both interpretable and aligned with industry standards.

FIGURE 6 SEGMENTATION SCHEMES FOR LOGISTIC REGRESSION MODELS WITH CREDIT BUREAU DATA



One logistic regression model is developed for each of the blue terminal node. Segments with gray background do not have sufficiently populated data elements to support a model. Segments with green background are parent nodes that will be further divided. Cash Flow Only segment is only available for the configuration with combined cash flow and credit bureau data. ***No Trade:*** No bureau record or no valid trade on file; ***Too New:*** Age oldest trade <= 6 months; ***Inactive:*** No trade reported within 6 months; ***Derog:*** Ever 60+, bankrupt, Foreclosure, unpaid collection > $250; ***Recent Derog:*** Derog within 24 months; ***Mature:*** Age oldest trade >= 60 months.

We did not develop segmentation schemes based on cash flow data due to a range of resource and data considerations. First, since that all cash flow features are consistently densely populated for all consumers, there was no operational need to segment on this basis.[10] Second, the fact that the performance increases in moving from LR models to XGBoost models were generally similar for the credit only models (which used segmentation) and the cash only models (which did not),

suggested that segmentation would not produce large gains in performance of the LR models that incorporated cash flow data. Due to resource and experience constraints, we decided to focus on other priorities because developing segmentation schemes can be quite labor intensive. Consequently, the cash only LR model and the credit score + cash hybrid LR model may not perform as well as some of the commercially available "cash scores" developed by a team of data scientists over an extensive period of time. However, this issue is unlikely to affect other models in the study.

C.4.1.2 Exploratory Data Analysis (EDA)

EDA is a crucial initial step in the model development process. It involves a comprehensive examination of the dataset to assess the quality, propriety, relevance, and strength of candidate features from each data source. The primary goals of EDA in the context of underwriting scorecards include identifying and handling missing values, detecting outliers, understanding the distribution of each feature, and exploring relationships between features and the target variable. Through statistical summaries, EDA helps in selecting a set of features that are not only relevant and defensible but also contribute to the robustness and predictive power of the logistic regression model. At a high level, we dropped features that are populated for only a few dozen consumers, dropped redundant features (features with near linear dependencies with other feature(s)), replaced missing values with sample means and created an accompanying indicator (dummy) variable, and, where deemed necessary, capped feature values at the 99th percentile and floored feature values at the 1st percentile. A more detailed examination is employed in the model finalization step after the entire set of input features have been winnowed down to a few dozen of the most promising candidates.

C.4.1.3 Variable (feature) selection

The primary goal of this step is to select the "best" model specification from among all models that could be spanned by the entire set of features/variables. We used a two-step selection process: a forward selection step using the Least Absolute Shrinkage Statistical Operator (LASSO) to select an "optimal" model and a backward elimination process using P-value. Where necessary, the developer might add to or remove from the initial set of the backward elimination step features along the LASSO path based on the developer's evaluation of the resulting model specification's explainability and performance.

C.4.1.4 Model finalization

This step involves a manual process in which each feature in the final specification is evaluated in detail and, where necessary, transformed to ensure the final model specification is parsimonious, explainable, and robust. A feature might undergo further transformation to remove nonlinearity suggested by the residual plus component plot. The preliminary missing treatment in the backward elimination step might be replaced with more robust imputations based on the developer's assessment of the meaning and risk level associated with a particular missing code. For example, missing number of inquiries could be replaced either by 0, indicating the absence of a valid inquiry represents a natural zero, or a negative value commensurate with the risk level indicated by the missing indicator. A floor and/or a cap might be applied to restrict the domain of an input feature to prevent extrapolation deemed unsupported or unreliable. Although rare, model developers sometimes exercise judgment to add or remove features to/from the final model based on their experience.

C.4.1.5 Binning and weight of evidence: considerations and omissions

The model development process outlined above adheres to well-established practices for logistic regression in credit risk modeling. However, we recognize that alternative techniques, such as feature binning and weight of evidence (WoE) transformation, are frequently employed in traditional scorecard development. While these methods offer certain advantages, we ultimately chose not to incorporate them into our approach.

Binning is the process of breaking continuous variables up into discrete ranges. An example is to categorize a variable such as credit utilization rate into separate intervals such as 0 to 10%, 10.1% to 25%, etc. Such techniques are used for a variety of reasons, including reducing models' sensitivity to outliers and simplifying the process of finding the best feature transformations in the finalization step. This approach aligns well with rule-based underwriting frameworks, where thresholds naturally segment risk. Additionally, binning can approximate complex or nonlinear variable relationships without relying on higher-order terms or interactions. For example, if default risk tends to be higher among consumers with very low or very high credit utilization rates compared to consumers with a middle rate, using bins would be one way to detect the different patterns at different points along the range of possible utilization rates.

However, binning also introduces some limitations. For instance, because it discards granular information into broader categories, it can potentially weaken predictive power. The selection of bin boundaries often involves subjective judgment or arbitrary rules (such as equal-width or equal-frequency splits), which may not optimize model performance. Furthermore, additional steps are needed to ensure the relationship among coefficients for the different bins is explainable.

To mitigate some of these drawbacks, practitioners frequently apply weight of evidence (WoE) transformation, which replaces each bin with a numerical value representing its log-odds relationship with the target. WoE offers several benefits, including standardized feature scaling, enforcement of monotonic trends (a desirable property for regulatory compliance and model risk management), and built-in handling of missing data. Despite these advantages, WoE does not fully resolve the fundamental constraints of binning, as it still discards granular information and relies on arbitrary bin boundaries.

While binning continuous variables and applying WoE transformations represent common practices in traditional credit scoring, we elected not to use these techniques in our study. Our decision was primarily driven by the need to maximize the informational value from our relatively small sample. Binning inherently discards granularity within intervals and limits the model's ability to extrapolate beyond predefined categorical thresholds. In contrast, we deployed continuous transformations, carefully designed to maintain linear and monotonic relationships with the target variable, to enable more effective generalization across the full spectrum of feature values. This approach allowed us to preserve model simplicity and ensure robust performance despite our limited sample size. That said, we acknowledge that with substantially larger datasets, carefully optimized bin boundaries, and proper monotonicity constraints, binned approaches (with or without WoE) may become preferable alternatives.

C.4.2 Development process for XGBoost models

Unlike logistic regression, which relies heavily on manual feature engineering and model specification, XGBoost is designed to automatically capture complex, non-linear relationships and interactions between features. This makes it particularly well-suited for high-dimensional datasets. Since the learning algorithm is largely automated, the development process for XGBoost models shifts focus away from manual intervention during estimation and instead prioritizes three key areas: (1) ensuring

the relevance of input features, (2) optimizing hyperparameters to suit the specific learning task, and (3) implementing early stopping to prevent overfitting and safeguard generalization performance.

C.4.2.1 Input feature selection

Because most of the input features from either the credit bureau side or the cash flow side are relevant for predicting credit risk by design, we allowed the XGBoost models to select from as many variables as possible. The only features excluded from the candidate set were certain age-related features in the credit bureau data that can only be used under certain limited conditions under the Equal Credit Opportunity Act.[11] This approach, which includes all eligible features without prior filtering, is often used to explore the full predictive potential of machine learning models.

We note that some practitioners conduct feature selection to narrow down the set of candidate features on which an XGBoost model to be trained. For example, weak features—those with low correlation to the dependent variable—are often removed to reduce the risk of overfitting to noise in the data. Similarly, redundant features, which are highly correlated with one another or with a group of features, are excluded to create simpler and more interpretable models. However, our primary objective was to study the predictive potential of machine learning models, particularly in the context of integrating cash flow data with traditional credit bureau attributes. During experimentation, we observed that removing the least predictive features generally resulted in worse performance on unseen data. As a result, we decided to retain all eligible features as candidate inputs, allowing the XGBoost algorithm to select the most relevant features during training. While this approach may reduce interpretability, it aligns with our goal of maximizing predictive performance and ensuring that the models generalize well to unseen data.

C.4.2.2 Model training and hyperparameter optimization

The XGBoost algorithm was chosen for its ability to capture complex, non-linear relationships in high-dimensional datasets. To optimize model performance, we employed the Tree-structured Parzen Estimator (TPE) method for hyperparameter tuning.[12] TPE is a Bayesian optimization technique that iteratively refines hyperparameter selections by modeling the distribution of the objective function.

The hyperparameters optimized included:

- **Lambda:** The tree-level L2-regularization parameter, which penalizes large weights (differences between estimates from different branches) in the model to control overfitting.
- **Maximum Tree Depth:** The maximum depth of each tree, balancing model complexity and interpretability. Deeper trees can capture more intricate patterns but may increase the risk of overfitting.
- **Minimum Child Weight:** The minimum sum of instance weights required in a child node. This parameter regulates tree complexity by limiting the size of tree leaves, preventing the model from creating overly specific splits.
- **Column Sampling Rate:** The fraction of features randomly sampled for each tree. This promotes diversity among the regression trees within the boosting ensemble, reducing overfitting by ensuring that trees are not overly reliant on the same subset of features.
- **Subsampling Rate:** The fraction of observations randomly sampled for each tree. Similar to column sampling, this promotes diversity among trees and helps control overfitting by introducing variability in the training data used for each tree.

The optimization process involved 300 trials, with each trial evaluated using 4-fold cross-validation on the development sample. The primary performance metric was the ROC-AUC.

While learning rate is an equally important hyperparameter, we maintained it at a fixed 0.30 during the TPE search process before final tuning to limit computational costs. For model estimation, we then reduced the learning rate to 0.01-0.05 with early stopping (1,200 rounds without ROC-AUC improvement on test data), achieving better convergence while respecting our 12-hour training window constraint. In optimization algorithms like XGBoost, the learning rate determines how quickly or cautiously the model moves toward minimizing the loss function. A higher learning rate allows larger parameter updates, potentially speeding up convergence, while a lower learning rate results in smaller, more cautious steps that slows down convergence but reduces the risk of overfitting in the early stages of model training, a critical consideration for high-dimensional and low signal-to-noise ratio applications such as predicting the probability of default. In our experience, setting the learning rate to the lowest number allowed by the time limit to achieve convergence usually led to the best generalization performance for an underwriting model.

The optimal hyperparameters, shown in Table C.3 are generally consistent across configurations, with some variation in complexity and diversity parameters. The L2-regularization parameters (lambda) are similar in magnitude on a log scale, indicating a comparable level of regularization across models. Most configurations favor a relatively small number of complex trees (maximum depth = 7), while the cash only configuration benefits from a larger number of less complex trees (maximum depth = 3). Another notable difference is the column sampling rate, which varies significantly between configurations. Single-source configurations (cash only and credit only) perform best with a small column sampling rate, suggesting that limiting the number of features per tree helps mitigate overfitting when data sources are less diverse. In contrast, mixed configurations (cash + credit score and cash + credit bureau data) achieve higher accuracy with larger column sampling rates. This is likely because individual trees in these models benefit from a higher likelihood of incorporating features from both data sources, enabling them to capture complementary patterns in cash flow and credit behavior.

TABLE C.3 KEY HYPERPARAMETERS FOR THE FINAL MODELS

The search spaces for hyperparameters are as follows: Lambda: log-uniform (0, 4096); Max depth: discreet uniform (1,7); Min Child Weight: log-uniform (0, 4096); Column Sampling Rate: discreet uniform (0.01, 1.00); Subsampling Rate: discreet uniform (0.01, 1.00)

HYPERPARAMETER	CASH ONLY	CS + CASH	CREDIT + CASH	CREDIT ONLY
NUMBER OF TREES	8,539	456	1,704	1,414
LEARNING RATE	0.01	0.01	0.05	0.05
LAMBA	2,958	3,224	3,319	4,041
MAX DEPTH	3	7	7	7
MINIMUM CHILD WEIGHT	18	3	1	2
COLUMN SAMPLING RATE	0.16	0.77	0.50	0.06
SUBSAMPLING RATE	0.25	0.92	0.89	0.92

Endnotes

1 *See*, for example, TransUnion, "How Often Do Credit Scores and Reports Update?"

2 When dealing with probabilities, it is sometimes easier to convert the probability into odds, the ratio of the frequency of an event occurring to the frequency of the event not occurring. This transformation is particularly useful because odds, unlike probabilities, are unbounded and can represent extreme likelihoods more naturally—for instance, an event with a 99% probability has odds of 99:1, while a 1% probability translates to 1:99. Additionally, odds simplify multiplicative comparisons: doubling the odds (e.g., from 1:1 to 2:1) is more intuitive than describing the equivalent probability shift (50% to 66.7%). To further linearize relationships, we often take the logarithm of the odds (log-odds), which maps the odds scale to the entire real number line and changes the multiplicative comparisons into additive ones. This logit transformation is the foundation of logistic regression, where linear combinations of predictors model the log-odds of an outcome, ensuring predictions remain interpretable while accommodating the S-shaped relationship between predictors and probabilities.

3 Correlation between input variables means that as one variable shifts in value, the other tends to change as well. For example, negative payment history and amounts owed are the two most influential components of many credit scoring models, and may often tend to shift together as consumers who become over extended start to fall behind on their payments. Where features tend to change in tandem, coefficients in a traditional logistic regression model that measure the magnitude of each individual feature's impact will have greater uncertainty levels because it is unclear which feature is actually driving the change in the predicted outcome.

4 Tree-based models use a hierarchical structure of "if-then" nodes to generate predictions of the likelihood of default. For example, an initial node might separate consumers based on whether they had previously filed for bankruptcy, and then subsequent nodes on each branch would further separate the relevant group based on current balances or other criteria. XGBoost methods generate multiple tree-based models, each of which are based on the prediction error of the prior model, and then create a final prediction based on the weighted sum of the prior models. This is more complex than a single decision tree but leads to lower prediction error rates and better predictive power. Chen and Guestrin, "XGBoost: A Scalable Tree Boosting System."

5 For more background on these approaches, *see* FinRegLab, Explainability and Fairness in Machine Learning for Credit Underwriting: Policy Analysis.

6 Note that mortgage loans involved in a 'short sale' may be classified as a default or not, depending on the outcome. If the sale proceeds are insufficient to cover the remaining principal, the lender typically reports special codes to the credit bureau, such as 'settled' or 'account legally paid in full for less than the full balance,' which is then treated the same as a charge-off in our default definition. However, if the proceeds exceed the amount owed, the lender does not incur a credit loss and typically reports the account as 'paid off as agreed,' which would not be considered a default.

7 Note that the same consumer could appear in the dataset multiple times and be counted as separate observations if he/she had new originations in different months with the sampling period. "Consumer" therefore designates a consumer at a specific time.

8 The sample included very few prepaid accounts, so we focused on checking accounts as the primary vehicle.

9 We recognize that this treatment might raise concerns that consumers relying on a single regular monthly income source, such as those whose only income is a monthly Social Security benefit, could be excluded more frequently. However, in our data, the proportion of consumers incrementally excluded by increasing the threshold from 1 to 2 is slightly lower for those with transactions categorized as "Retirement Income" by the aggregator compared to the rest. This suggests that the criterion does not disproportionately affect retired consumers.

10 Unlike credit bureau features/attributes, cash flow features do not have apparent sparsity patterns in terms of being missing for a large segment of the consumers in the sample. One example of such a sparsity pattern on the credit bureau side is that the credit features/attributes related to the number or timing of delinquency events are all missing for consumers without a delinquency event. Having one separate model segment for these consumers obviate the need to include features related to delinquency in the equation. Cash flow features, on the other hand, tend to be densely populated for all consumers in the sample.

11 The separate dataset that includes certain demographic information for purposes of subgroup analysis was not used for model building either.

12 Watanabe, Shuhei. "Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance."

With support from:

JPMorgan Chase & Co.

JPMorgan Chase & Co. (NYSE: JPM) is a leading financial services firm based in the United States of America ("U.S."), with operations worldwide. JPMorgan Chase had $4.4 trillion in assets and $351 billion in stockholders' equity as of March 31, 2025. The Firm is a leader in investment banking, financial services for consumers and small businesses, commercial banking, financial transaction processing and asset management. Under the J.P. Morgan and Chase brands, the Firm serves millions of customers in the U.S., and many of the world's most prominent corporate, institutional and government clients globally. Information about JPMorgan Chase & Co. is available at www.jpmorganchase.com.

Capital One

Capital One Financial Corporation (www.capitalone.com) is a financial holding company which, along with its subsidiaries, had $367.5 billion in deposits and $493.6 billion in total assets as of March 31, 2025. Headquartered in McLean, Virginia, Capital One offers a broad spectrum of financial products and services to consumers, small businesses and commercial clients through a variety of channels. Capital One, N.A. has branches located primarily in New York, Louisiana, Texas, Maryland, Virginia, New Jersey and the District of Columbia. A Fortune 500 company, Capital One trades on the New York Stock Exchange under the symbol "COF" and is included in the S&P 100 index.

Digital version available at finreglab.org

Published by FinRegLab, Inc.

1701 K Street NW, Suite 1150
Washington, DC 20006
United States