Case Study

Creating a True Picture of Income Using Bank Transaction Data

How SoFi leveraged Nova Credit to more efficiently verify customer income

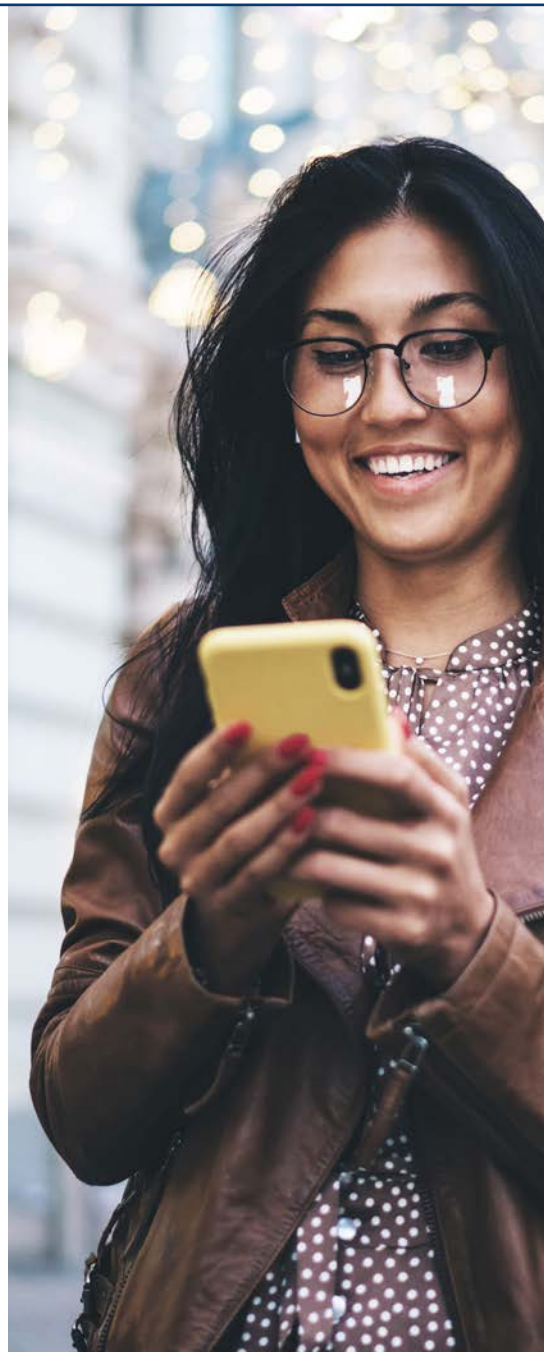
NOVA CREDIT

Table of Contents

Executive Summary

Nova Credit, a consumer-first data platform, and SoFi, a premier online personal finance company, have partnered together to develop and launch a suite of cutting-edge income insight tools to more accurately and efficiently verify income for more applicants using consumer-permissioned bank transaction data.

This whitepaper will discuss the challenges of quickly and accurately verifying consumer income in today's quickly changing macro-environment and will explore how bank transaction data has emerged as a compelling solution. The paper will then walk through the unique verification solution that Nova Credit and SoFi developed to interpret this transaction data, including details of the use case, model design and performance, and resulting business outcomes. Finally, it will discuss early learnings from the iterative partnership and offer thoughts on the broader potential of using new sources of consumer-permissioned data to inform lending decisions.



The Challenge

Consumer income is an increasingly important element of the credit risk underwriting process. In order to determine whether to approve or reject a loan application for high-value loans, it is not only important to understand a consumer's previous behavior of taking on financial responsibility and willingness to repay additional debt (through their credit history), but also their capacity to pay that debt by verifying that their stated income is accurate and up-to-date.

And yet, while verifying income became even more pressing last year as issuers navigated the waves of unemployment caused by the coronavirus pandemic, it became apparent that existing solutions each had their challenges. Some databases, such as Equifax's TheWorkNumber (TWN), enable lenders to verify income quickly through backend checks, but are largely limited to employees who have traditional income from traditional employers.

Alternatively, lenders may also rely on payroll data, which requires an issuer to integrate with several providers across the fragmented payroll market. Even once available, payroll data only captures a consumer's principal employment but overlooks secondary sources.

Without a full picture of the borrower, and at the exclusion of a large segment of consumers such as retirees, self-employed, and gig economy employees, these solutions result in costly, burdensome manual review processes for many applicants. Since there is often significant applicant drop-off as a result, this ends up capping issuers' potential upside and ultimately limiting many consumers' ability to get the loans they need.

The Solution

Bank transaction data, by contrast, can provide a more comprehensive understanding of income for a broader set of consumers. It not only captures secondary sources of income, such as investment returns, rents, and retirement income, but can also theoretically cover anyone who gets these various income streams deposited in their bank account and opts in to sharing that information.

And while consumers are becoming increasingly comfortable sharing data from their bank account with lenders, underwriting teams face challenges accurately interpreting and using this data in their decision-making processes.

Existing solutions based on bank data offer messy transaction-level data that is often misclassified or overestimated, particularly for non-salaried workers. Without being able to identify genuine income from amongst the many deposit streams into a bank account, underwriters cannot verifiably understand an accurate estimate of an applicant's ability to pay.

> **"In theory, bank transaction data gives you tremendous visibility into a consumer's ability to pay, but in practicality that data is extremely complicated and not easy to make sense of."**

– Ratinder Bedi

Chief Credit Officer, SoFi

In 2020, SoFi & Nova Credit formed a partnership to develop an income identification model using bank transaction data in order to address this specific challenge.

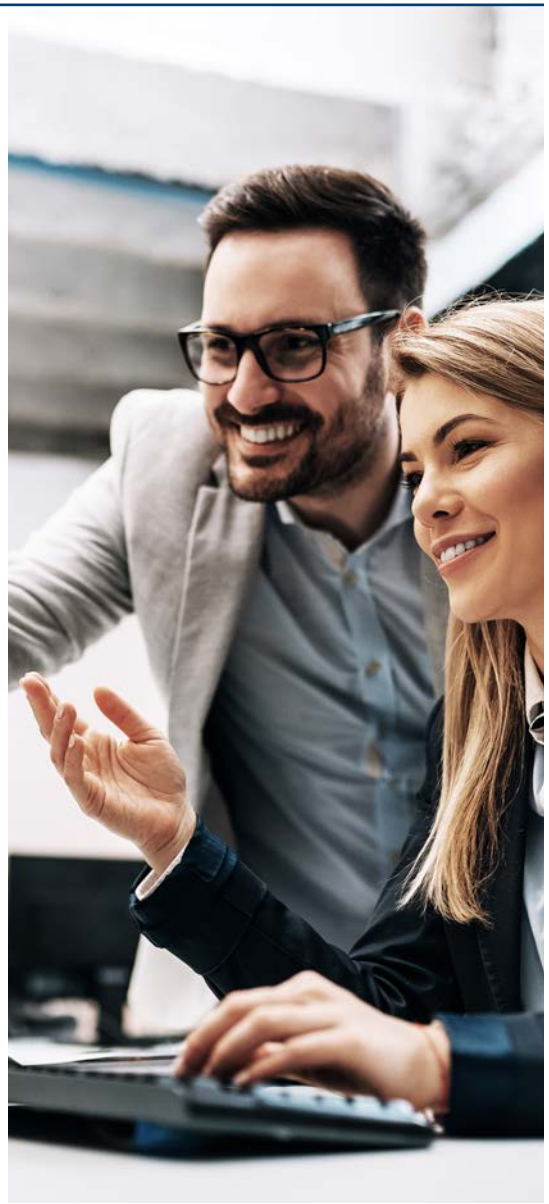
The objective of the partnership was to develop a tool that would benefit SoFi by reducing operational underwriting costs through automation efficiency and—over the long term—by reducing loss rates given more accurate income assessment.

Business Context

Loan application processes typically ask for stated income to provide insight into a consumer's capacity to afford an incremental loan payment. In the majority of cases, the underwriting strategy requires that the stated income is verified.

If a lender is not able to verify the income through backend-only checks like TheWorkNumber (TWN) because no data on the consumer exists there, they will often ask the applicant to upload pay stubs or tax forms—which they will review either through an automated process using OCR technology or, more commonly, through time-intensive and costly manual review processes.

In SoFi's workflow, applicants who are not able to be verified by automatic verification checks are given the option to verify their income by linking their bank account using Nova Credit or by uploading a pay stub.



For those who opt into linking a bank account via Nova Credit, the following income verification process happens in near real-time:

Once an account is linked, Nova Credit's income model reviews all deposit streams and identifies those that are likely to be income.

The model then aggregates and annualizes all identified income streams to determine net income.

Finally, the model estimates gross income and compares it to the stated income provided by the consumer.

If the model verifies an income that is consistent within a range of the stated amount, the loan application proceeds to the next step in the approval process.

Model Design

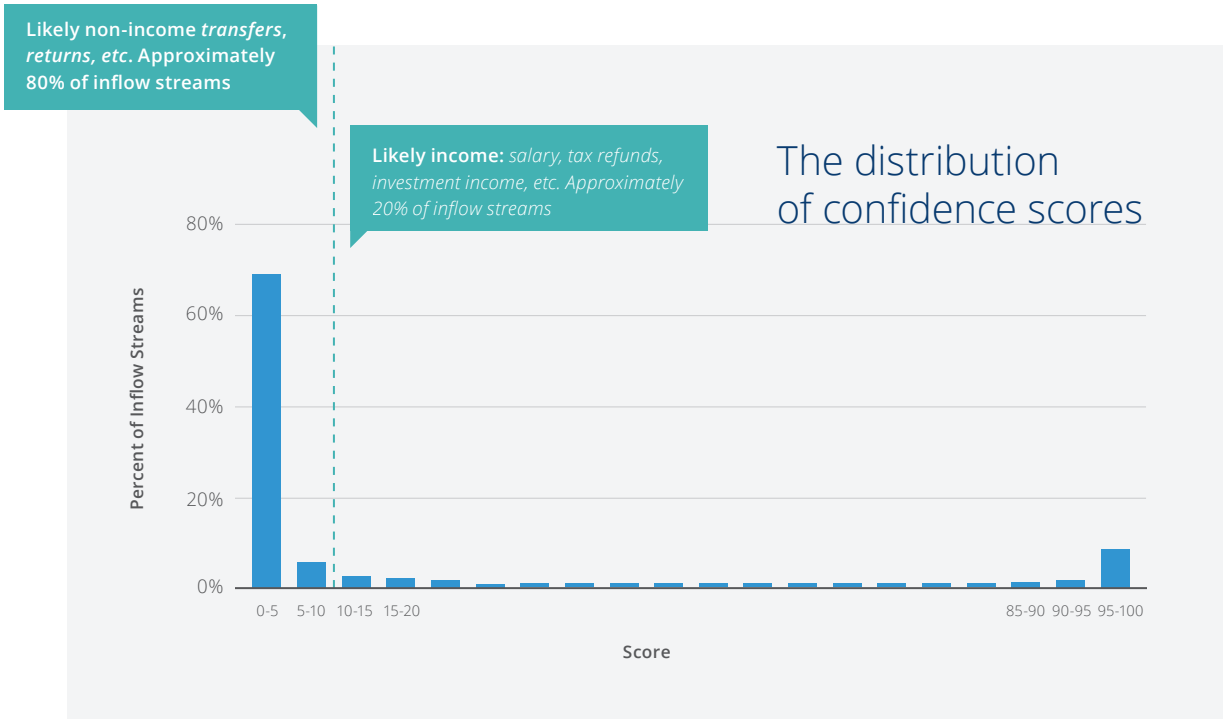
The purpose of Nova Credit's income model is to identify which deposit streams in an applicant's bank transaction data represent "income" (as opposed to random deposits, such as merchandise returns or money transfers). Essentially, the model evaluates the many deposit streams that might appear in the consumer's bank accounts and disqualifies those that fail to display income deposit characteristics, leaving only those deposit streams that most likely represent income. In reality, a typical consumer has eighteen deposit streams but just 2-3 income streams[1].

Using a combination of statistical classification (i.e. logistic regression) and fixed rules to classify income streams, the model identifies three classes of income: conventional, non-conventional, and non-wage income. Conventional income derives from full-time employment receiving a regular paycheck. Non-conventional income is income from any other form of employment, including contract labor and self-employment. Non-wage income is income that is not tied to on-going employment, such as interest income, retirement income, and tax refunds.

In reality, a typical consumer has eighteen deposit streams, but just 2-3 income streams.

[1]NovaCredit Analysis, October 2020

Each deposit stream is scored on a 0-100 range and a score cut-off is established at a particular threshold to disqualify non-income deposit streams. A threshold score of 10 is applied to maximally separate non-income streams from income streams: most non-income streams score in the 0-10 range, while the majority of income streams score above 10. This cut-off reflects the imbalance between income and non-income streams: mislabeling income streams is expensive due to their scarcity, while mislabeling non-income streams is cheap due to their abundance.



Finally, having identified a set of income streams, the model determines annual and current quarter net income, and then estimates gross income via a net-to-gross multiplier that accounts for average pre-tax deductions (healthcare, retirement, and dependent care) and anticipated state and federal taxes. The model returns net and gross income amounts on an annual and current quarter basis for use in the income verification process.

Model Performance

The model is able to distinguish income streams from non-income deposit streams with incredibly high precision. For conventional income, the classification model accurately classifies more than 85% of income streams in terms of dollar amount, while mislabeling less than 15% of non-income deposit streams. For non-conventional income, which is more challenging due to its variability, the model correctly labels more than 95% of income streams in terms of dollar amount, and mislabels under 5% of non-income streams in terms of dollar amount.

Performance of income stream classification models

Table. Performance of income stream classification models. KS is the Kolmogorov-Smirnov statistic. AUC is the area under the receiver-operator (ROC) curve. TPR and FPR are the true and false positive rates, respectively.

Very strong KS scores across both conventional and non-conventional models

	Conventional Model		Non-Conventional Model	
Metric	**Train**	**Test**	**Train**	**Test**
KS	0.87	0.87	0.62	0.58
AUC	0.97	0.97	0.87	0.86
TPR	0.96	0.96	0.27	0.30
FPR	0.09	0.09	0.03	0.03

When the model does make an error, its impact on the overall income estimate is generally small. The figure below compares the income calculated by model versus an estimate of income calculated using manually labeled income streams. For approximately 85% of applicants, the relative absolute error between the model-based and manual estimate is less than 25%.

Finally, the verified income amounts tend to support applicants' stated income. Here, the model's current annual gross income estimate is compared with applicants' stated annual income. In 60% of cases, the estimate falls within a range considered acceptable to verify stated income.

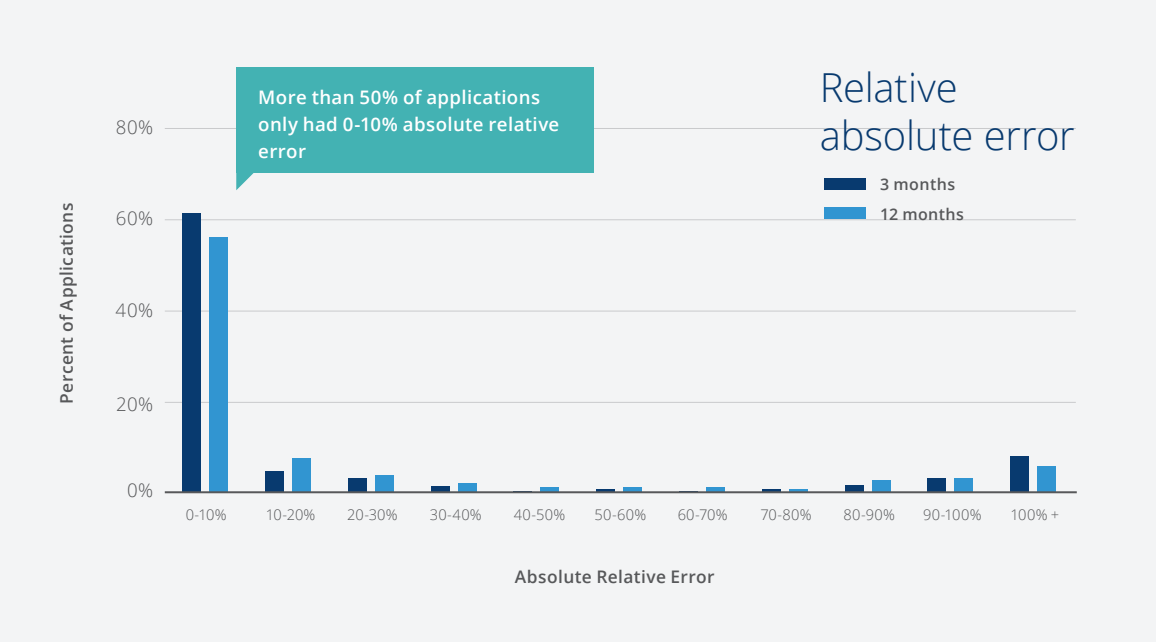


Figure. Relative absolute error of estimated gross annual income versus manually verified gross annual income for estimates using the most recent three months (3m) or twelve months (12m) of available transaction data.

"Nova Credit's income verification tool has delivered excellent, tangible results for our business – higher loan conversion, a shorter time to fund, and ultimately a great member experience because the process is so seamless."

– Ratinder Bedi

Chief Credit Officer, SoFi

For SoFi, this has led to a meaningful increase in automation, saving time and money from otherwise manual verification processes, and has done so with a higher degree of accuracy than other bank account verification solutions. As a result, many consumers have been able to have their income automatically verified, removing friction and stress from the application process.

Conclusion

The unprecedented macro environment in 2020 pushed lenders to embrace an inevitable reality: in order to accurately understand an applicant's risk profile, especially in uncertain times, one must look beyond credit data. An applicant's income is a critical piece of this expanded picture, but until now has been difficult to verify in a way that does not rely on manual processes, limited-coverage solutions based on payroll data, or messy transaction-level data from consumers' bank accounts.

Using SoFi's firsthand understanding of this problem and Nova Credit's in-house data and analytics expertise, the two companies developed and launched this proprietary income model to parse through bank account data shared by a consumer, intelligently identify and weight income streams across income types, and deliver standardized income insights to SoFi's risk team for real-time decisioning.

This partnership marks an excellent example of quick execution followed by ongoing iterative development. Product, engineering, and data science teams from both companies meet regularly to review data and align on areas for continued optimization. The two companies developed an exciting analytical innovation atop an emerging data source, which can now be applied to different credit use cases across lenders to solve a broader income verification market need.

NOVA CREDIT

Learn more online at NovaCredit.com

facebook.com/novacredit

https://www.linkedin.com/company/nova-credit-inc.

@nova_credit

Version 1