Deep Science Publishing
https://doi.org/10.70593/978-93-49910-91-1_5

Chapter 5: Building smart credit scoring engines using alternative data, behavioral patterns, and predictive modeling

5.1. Introduction

In recent years, there have been drastic changes in consumer sentiment toward financial services. Many customers who opened accounts with traditional banks have transferred service relationships to Fintech companies. There are myriad reasons for this phenomenon, such as uncompetitively low interest rates, lack of convenience in branch modernizing, customer disappointment with the risk of bailing out, complete substitution of financial intermediation by market-based financing through shadow banks, increased management overheads, and expansion of campaigns for non-customers from alternative financial service contractors. Fintech companies often do not have their banks. However, they provide loan products through their partner banks, who must not only comply with costly regulation but also take on huge risks with respect to those loans (Puschmann, 2017; Narula & Van der Straaten, 2020; Ng, 2020).

To promote lending and other financial services, alternative data sources to credit bureaus are becoming more important in making credit decisions. A positive credit score is a necessary condition to access a loan product from banks, credit cards from credit card issuers, and at any monetary transaction point from numerous other lenders. However, nearly 50 million adult consumers do not have a credit history with the major bureau aggregators. Furthermore, many consumers with an insufficient credit history or with a "thin file" cannot receive traditional lending access. Consumers within these voids, who are turned down by banks but confident about the repayment ability, have a strong desire to get loans. Banks and other financial institutions have begun to use models built with alternative data for credit scoring decisions associated with those consumers (Schindler, 2017; Ryll et al., 2020).

5.1.1. Background and Significance

The credit scored development, deployment, and utilization impact the daily lives of most individuals by influencing access to and cost for fundamental needs such as housing, transportation, and education. The outcomes of decision-making systems based on credit scores, such as higher housing costs or loss of eligibility for public assistance programs, can have a lasting impact on behavioral and/or health-related outcomes. Default patterns associated with credit scores can reflect a wide range of behavioral patterns, including increased risk for psychiatric disorders, generalized substance use disorder risk, increased risk for specific alcohol or marijuana use disorders, increased risk for health risk-related behaviors, increased risk for incarceration during pregnancy, increased risk for comorbid psychiatric disorders, greater number of emergency department visits, greater number of psychiatric emergency department visits, and lower antibody response following hepatitis B vaccination, specifically among adults with HIV. Predictive modeling can thus be used to discover novel associations between credit scores and other clinical-significant behavioral or health-related outcomes.

Fig 5.1: Building Smart Credit Scoring Engines Using Alternative Data, Behavioral Patterns, and Predictive Modeling.

While alternative data credit scores are typically built on substrate national databases and include disparate variables, behavioral anchors based on predictive modeling are built on localized project datasets that contain homogeneous variables. Alternative data and other area-specific variables contained in both score types would then reflect transactional information related to potential scoring discrepancies. A known criticism of credit scoring models is their use of historical payment behavior and insufficient consideration of recent life events that influence chance for account default. Despite this, alternative data credit scoring has been adopted by multiple financial institutions in jurisdictions worldwide, including Southeast Asia, Europe, and the United Kingdom. Adoption is due to increased availability of digital, predictive, and alternative data as well as need for increased credit access, lower underwriting costs, and augmented loss performance.

5.2. Understanding Credit Scoring

Credit scoring is defined as utilizing data modeling techniques to derive a numeric measure that depicts the likelihood of a discrete event happening or not happening that has substantive implications for the measure-user. In our context, the measure would be a numeric value derived from disparate forms of data about an individual applicant, and the event would be default or non-default for a loan provided to that individual by either a regulated traditional bank or a non-traditional lender. The numeric predictive measure is employed as a deterministic decision rule - such that a prospective applicant for credit is either rejected if the value is less than some cut-off point, or accepted if it is greater than or equal to that value. The decision by a bank or other lender to employ credit scoring as a method of predicting the chance of default on credit which it is offering is made in a cost-benefit framework so that employing the strategy is expected to be profitable for the lender, and there are specific provisions which require consumer credit lenders to disclose the fact that they use these methods, and provide certain information as to their application and results.

Credit scoring has become a common and popular part of granting consumer credit in the US and many other countries, so that a substantially large percentage of consumer credit lending and associated institutions routinely utilize credit scoring. The oldest known use of numerical scoring of credit risk dates back to an early use of a logic model for mail order credit, and has recently come to increasingly include Bayesian probability models. The principles underlying credit scoring have some fairly deep social science and quantitative modeling foundations, and a diverse academic literature has explored several different prediction methodologies, sample sizes, and alternative data source specification combinations.

5.2.1. Research Design

This study aims to describe the state of the art of credit scoring predictive performance measures, to understand stakeholder's perception about credit scoring, and to propose novel performance measures inspired from the characteristics described by the stakeholders. In the first stage, a thorough and systematic literature review on predictive performance measures was conducted. The second stage relied on semi-structured interviews with a valid group of experts from different areas of credit scoring. Discussions with stakeholders allowed for the identification of different dimensions or characteristics that can affect a credit scoring predictive model. Finally, the analysis of stakeholders' responses led to the proposition of several new predictive performance measures based on stakeholders' perception. A design science approach was embraced because it supports the iterative nature of our study and it contains all the necessary building blocks for our study. The contributions of this study lie within the area of performance analysis in credit risk management by managing the trade-offs between profitability and loss given default.

In this study, a design science approach is adopted to blend both existing knowledge and novel knowledge by proposing novel credit scoring model predictive performance measures based on expert's opinions. The goal of this paper is to address how stakeholder's opinions should influence the performance measures to be calculated and the way they should be calculated. The aim of these proposed measures goes further than mere performance evaluation. In a practical sense, we want to improve the process of credit scoring model deployment in production, to help modelers get quantitative feedback from stakeholders in a more informal and intuitive way.

5.3. The Role of Alternative Data

Traditional credit scoring models rely extensively on Financial Credit Bureaus, putting the vast majority of the world's population at the periphery. Alternative Data is data associated with borrowers that is used to complement the scant and sparse information traditionally available within the applicant's own credit history used for lending decisions and enhance the process of scoring individuals lacking sufficient historical Financial data. Consumers often lack sufficient or meaningful credit inquiries, derogatory marks, or accounts opened to be scored by traditional means. The output of our Smart Credit Scoring Engines is a Risk Score – risk scores are numeric representations of the probability of circumstance over a predefined time horizon. By modeling Individuals without credit data alongside individuals with credit data, models can learn from the predictive relationships established between alternative attributes and financial credit events. Alternative Data comes from digital footprint information as consumers increasingly rely on digital devices for everyday transactions. This

information increasingly provides insight into consumer behavior and presents itself in the form of transaction behavior such account balances and funding patterns, propensity to save or spend, non-credit based payments like rent and utilities, ability to sustain financial shock, geo-location patterns like business establishments visited, social circle information and relationships or associations with family members and friends, and engagement pattern with digital channels like mobile apps or websites of financial institutions. Using such behavioral characterizations lead to the identification of segments of consumers embracing, or transitioning through the various stages of life cycles, like credit invisible, near-prime, and sub-prime.

5.3.1. Types of Alternative Data

Another data is a segment of data available and used in the domain of business analytics, as well as in other domains outside business analytics. There exist two broad classes of alternative data depending on its structure, namely structured data and unstructured data. Structured data can further be divided into two types based on its expertise, namely domain expert verified data and non-domain expert verified data. The former includes cost segment data, consumer market statistics, employment statistics, earnings data, economic data, financial statement data, international trade data, and household income and expenditure data, while the latter includes news data, insider trading data, corporate and executive connections data, social media data, and patent data. Furthermore, unstructured data can also be classified into two segments, textual data and visual data. Examples of the former include news data, social media data, and patent data, while the latter consists of images. Textual as well as visual alternative data are generally non-domain expert verified unstructured data.

The primary motivation for fintech companies to explore both alternative data is the ease of access at a relatively low cost and the richness of information contained therein. With a significant growth of mobile technology and social media in the last decades, fintech companies can access consumers' online footprints at a relatively low cost. Furthermore, the advent of big data analytics, natural language processing, and internet of things has made the analysis of both textual and visual alternative data less complicated and time-consuming. In particular, the growth in mobile payment and credit card transaction volume has enormously increased the transaction data available to fintech companies as users of both services capture their transactions on a digital channel.

5.3.2. Sources of Alternative Data

Besides the important issue of what types of alternative data to use in your credit scoring models, another key question is where to find that alternative data. In this critical section,

we describe the main sources that have been used and/or are currently used in practice to collect alternative data.

Already today, the largest existing source of alternative data is the internet, which has been increasingly populated over the last 25 years with the credit-invisible and credit-thin files populations. Social media and applying digital footprints such as browsing behavior, location, and email usage are also increasingly used sources of alternative data. Indeed, a report explains how social media companies scrape data from users who have submitted credit applications or have provided personal data.

Crowd-based data is also currently being extensively used. This data is produced by borrowers interacting, and it creates great value if the borrower has performance data over their credit obligations and provides some identity verification information. In addition, open infrastructural data, such as real-time payments data, allows banks and fintechs to better screen and assess underserved customers. Engagement data helps assess the patterns of borrowers, such as their responsiveness to messages or campaigns. There also exist anonymous and aggregated spending data from credit card transactions or bank account transactions that credit bureaus and companies collect in order to enhance or create alternative scores.

5.4. Behavioral Patterns in Credit Scoring

Behavioral patterns are specific data that help understand why you are using the services the way you are and how that can be mapped to your chances of default. These patterns span across a varied spectrum of data – credit and non-credit. Behavioral patterns in credit scoring can be defined as the patterns identified using a life's worth of data (both credit-related and non-credit related) across any one or a combination of prediction points. A person is a sum of their behavioral patterns, and that information is what the score engines will use to determine creditworthiness. This involves using a defined set of data to identify probable behavioral patterns and quantify their impact on creditworthiness.

A common query that we face is what is the difference between behavioral modeling and predictive modeling. A fair answer is that behavioral modeling is building behavioral patterns and mapping them to prediction points, while predictive modeling is modeling the patterns that would impact the chosen target variable. Root cause explains the creditworthiness reasons, while predictor describes the reaction attributed to the reasons. For example, where the root cause could be late payments, it involves modeling the patterns that could result in early defaults, while predicting variables such as the reason for late payments that affect future creditworthiness or defaults. Behavioral modeling also goes beyond predictive modeling, as it creates these patterns to explain

why a probability model has identified a person to be a poor credit risk and the inverse query – why it has assigned a score in the lower percentile for this person and why such a poor score would lead to taking such action.

5.4.1. Identifying Behavioral Patterns

Behavioral patterns are actions that are repeated over time and are correlated. Identifying detectable consumer's behavioral patterns from their transaction history has been used in numerous applications outside the scope of general credit scoring. Examples of such applications include cross-selling, forecasting market trends, anomaly detections, and customer loyalty maintenance. In financial services, there are multiple use-cases in which entities generate similar behavior patterns but have varying credit confidence. We use the term "behavioral patterns" to describe actions or events that are correlated over a certain time span and can identify the target group performing this type of behavior.

Examples of behavioral groups include early expenses for luxury retail spending correlated with delinquency prediction or late expenses for essential purchases that are correlated positively with the target. When associations are detected in a structured way, it can be possible for a person to have great festive spending at the end of each year but still not pay on time in early years, as would be expected based on the group's average behavior. Other examples range from crypto transactions being correlated with new unmanageable personal loans to fintechs offering buy-now-pay-later services.

5.4.2. Impact of Behavioral Patterns on Creditworthiness

Using behavioral patterns, in addition to static features representing customer attributes such as demographics, in credit scoring enables increased explanatory and predictive power. Financial behaviors are typically represented in relational credit files in two forms: transaction-based dynamic features summarizing specific numerical attributes of how and when particular transactions occur over a duration; and sequence-based categorical features, identifying the sequence of events for behaviors such as the merchant type using which purchases were made or the type of payment processed on a given day. Such behavioral features capture unique and concrete spending and payment characteristics that general static features cannot adequately explain. Behavioral features enhance credit scoring performance by reflecting actual historical patterns of how living expenses are incurred on a day-to-day basis. In fact, more than static features, these detailed patterns comprising information on when customers spend, how much they spend with what kind of merchants, and how customers are able to manage their spending can provide a clearer picture of the customer's current creditworthiness. These

features can even make predictions for thin-file customers who might not have expansive past credit histories.

Behavioral features can also directly impact scoring results. The presence of certain patterns, such as spending absorption spikes or late payments during a month leading up to a scheduled big expense, could increase credit risk, while patterns showing ability to defer certain purchases related to the anticipated upcoming expense may reduce risk. Having detailed insights into key behavioral patterns that directly impact the ability to service debt during critical periods of anticipated excessive spending could help financial institutions in helping thin-file customers utilize credit responsibly. Such insights aid lenders in assisting customers whose cash flow predictions are deteriorating.

5.5. Predictive Modeling Techniques

Different predictive modeling techniques fall broadly in two categories: traditional statistical methods and machine learning algorithms. Prediction models are built using

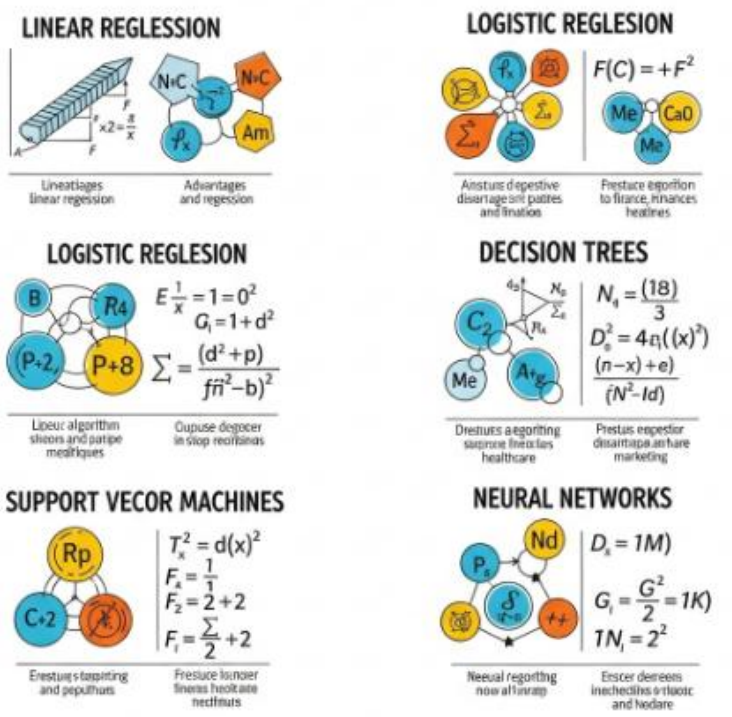

Fig 5.2 : Predictive Modeling Techniques.

either or both of these methods or algorithms. Statistical methods primarily model the dependent variable based on a functional form of the predictor variables and their

parameters are estimated using the entire sample. Machine learning algorithms in contrast learn from a sample of predictors and their relationships with the response variable, to build a “black-box” and use it to predict future values. The learning is typically done using a portion of data called the training data and the performance is evaluated on the remaining sample called the test data. Predictions may be made using both approaches, and the accuracy of predictions for specific applications should dictate which of the methods is chosen for the task.

Predictive analysis focuses on three different types of analytical functions. Predictive functions for credit risk modeling predict the probability of default, severity of loss due to defaults, and the exposure at default. When available, data and behavioral attributes are used to build predictive functions that estimate one or more of these probabilities. Typically, data variables are used to build predictive functions that estimate the probability of default, attributes describing borrower behavior are used to assess the loss given default, and measures associated with contract duration are used to estimate the exposure at default. The creation of segments among the borrowers is the final task for creating risk management strategies for firms.

5.5.1. Overview of Predictive Modeling

Predictive modeling is a statistical and analytic technique that deals with the prediction of a future outcome based on historical information and is often referred to as forecasting. Looking at past data, an analyst tries to identify the trends and patterns that are associated with the variable that is to be predicted and then builds up a mathematical equation, or function that calculates the predicted result based on the other input variables. This function is then applied for predicting the values of the output variable for the future data when it becomes available. The area of predictive modeling as part of statistical analysis has been around for quite some time and has many diverse applications such as predicting sales, industrial output, and electricity demand. The major reasons for the recent large growth in predictive modeling in commercial applications and as part of operational data mining are the availability of large amounts of business-related data in data warehouses, the results of analytical campaigns being built into ongoing business activities such as risk assessment, credit scoring, and fraud detection.

Predictive modeling techniques are used across various verticals and for addressing diverse business problems. The techniques are used by organizations across diverse industries such as retail, banking, insurance, healthcare, and telecommunications for a number of задач, including risk assessment, financial forecasting and predictions, attrition of valuable customers, targeted marketing, promotions and offer recommendations, loyalty program improvement, new product pricing, demand and

sales forecasting, and customer and market segmentation. While these techniques assume specialized meanings when applied within a specific discipline, as is usually the case with most analytical tools, they are extensions of the same basic technique that is employed over and over again.

5.5.2. Machine Learning Algorithms

In financial risk assessment, the term predictive modeling is typically used to refer to scoring risk models and risk segmentations that assign evaluations to each subject in a population. Predictive modeling means developing a relationship between past results for a sample of subjects and the associated features of each subject, whether behavioral, demographic, or both. Machine learning broadly refers to algorithms that heuristically search for models to describe data. While traditional statistical modeling is geared at testing hypotheses and attempting to determine the functional form of the underlying model, the goal of machine learning is to provide a model that predicts the most accurate probabilities of the events of interest for new subjects, often at the expense of interpretability. In many marketing cases, the purpose of the model is not to explain variation of scores among individual subjects at all, but rather to selectively apply the model to those subjects in the population that are the most likely candidates for a marketing action, such as an offer of credit or some type of upsell or cross-sell.

Machine learning algorithms fall into several types. First are traditional predictive algorithms such as classification trees, neural nets, logistic regression, and regressions using generalized additive models. These are traditional statistical regression algorithms that attempt to find the additive contribution of each predictor variable. The second class is known as modern boosting or ensembling techniques. These models contain many predictive algorithms applied multiple times, each working to fix the inadequacies in probability prediction or scores of prior models. The need to parallel-process sophisticated statistical predictions ensures that these models are only as interpretable as the original implementing algorithms; in the cases of boosted classification trees and boosted neural nets, they are less interpretable, while the case of boosted regression is often considered the most interpretable due to it employing a traditional statistical model.

5.5.3. Statistical Methods

Common statistical methods used for predictive modeling include Logistic Regression, Generalized Linear Models, Generalized Additive Models, Linear Discriminant Analysis, and Survival Analysis. Regression models (Generalized Linear Models, Logistic Regression, Generalized Additive Models) are standard methods used when creating predictive models for classification type outputs. A linear regression model is

used when the target variable is continuous. In the case of probability scores, linear regression has limitations, since the predicted target score is equal to or greater than zero. Probabilistic classification models like Logistic regression map predicted probabilities while satisfying the aforementioned constraint. The advantage of the Logistic regression is the simplicity and the relative small number of parameters estimated.

Generalized Linear Model and Generalized Additive Models generalize the Logistic regression in that they allow the use of alternative link functions which can address cases where the relationship between features and the target output is not well represented by the Sigmoid link function used in the Logistic regression formulation. Linear Discriminant Analysis can be seen as a constrained version of the Logistic Regression, in that it assumes normality of distributions of the target classes and that both classes share the same covariance matrix. When the number of observations for one target class is limited, the normal distribution assumption of that class can help with the construction of a predictive model and thus reduce overfitting. The limitation of LDA, however, is the normality assumptions do not hold for every use-case. Survival Analysis models duration type target variables. The model can be used to estimate time-to-failure probabilities and is suitable for prediction where the output is however restricted to estimation of time-to-event.

5.6. Data Collection and Preprocessing

When building credit scoring models, data collection and preprocessing is the foundation step for successful predictive modeling. It contributes to how predictive the final models will be. Different types of data are available for a fintech startup conducting credit scoring. The first place to be considered for data collection is the applicant and his/her social network and the easiest way is asking for permission to access this information. Everything that identifies the applicant will be stored as identifiers. For each applicant, characteristics embedded in photo, video or text (or a combination of them) can be specified. For example, the applicant can choose an advertisement to test how its embedding is behaving. The category of the advertisement can also be used as an identifier of what the ad represents. Clustering will identify the probabilities of generating different advertisements.

Behavioral patterns capture the life of the applicant. Application of deep learning on the behavioral data would discover improved feature engineering of the applicant's different stages. Behavioral pattern-based features can significantly enhance credit scoring predictive modeling results. The data will be collected according to the behavioral usage patterns that are specified to be analyzed. Deep learning can enable input of a sample of an incomplete set of behavioral data and predict filling the time series to be stored with

shorter sample time series artifacts to accurately predict the rest of the time series to be used for predictive modeling purposes.

5.6.1. Data Collection Methods

Data has been collected through multiple channels to build smart credit scoring models. The final collection includes a mix of primary and secondary data collections followed by a vetting process to ensure quality and accuracy. A mix of data sources such as registries and bureaus to scrape and retrieve information on borrowers who went through a default. Internally, behavior data is captured through the platforms of each product and shared for scoring on a regular basis. For instance, for Property, the partners have regular updates in terms of interest, views, follows, etc. that help us track engagement. For travel, the internal data reflects the user engagement in terms of bought travel, planned travel, and past travel. For payments, transaction monitoring points to the bill payments missed and those paid along with the frequencies. Such data heterogeneity helps in ISR segmentation, which is used to ascertain the quote shared for the customer early on in the interaction.

The data sources used for the project include support and network clusters developed from contact databases. Apart from the data collected from existing datasets, the information used from repositories. The partner data for Property came from partners, the Travel service providers include others. The partners for Payment are among others. These data sources have been further vetted in terms of validation and de-duplication. New Customer data are mostly populated from their fresh interactions with the products across these partners. With location being the main aggregator tool used for audience development.

5.6.2. Data Cleaning Techniques

Data Cleaning Techniques. Data preprocessing serves multiple purposes. The first step is to transfer objects originally in different formats to a unique format, called normalization, which is often accomplished by a simple validation function. The data can then be loaded into an internal representation, which is the process of constructing a data structure appropriate for processing the object. The different steps of the data collection step prepare objects for the construction of an internal representation. Data normalization is often required for heterogeneous data, including biometrics, and it is a critical step since errors faced on data streams have a cascading effect on the internal representation.

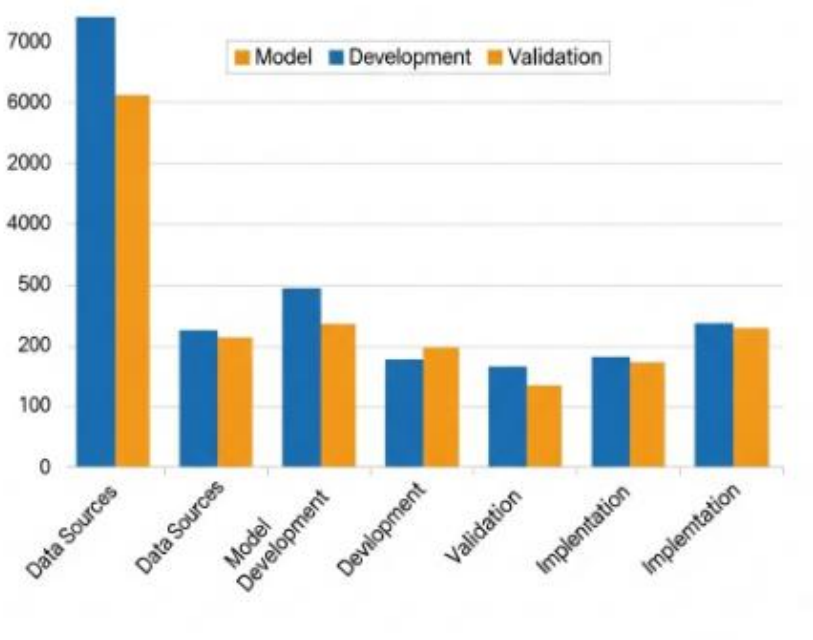

Fig : Building Smart Credit Scoring Engines Using Alternative Data, Behavioral Patterns, and Predictive Modeling.

Two possible alternatives are the use of confidence maps that assign units of the internal representation a confidence level, or module-based internal representations in which a particular mapping exists for specific object subsets. Both of these techniques attempt to recognize objects with specialized techniques used only when high confidence levels are reached or the representation lies in the proper module. The internal representation is a critical step in any supervised learning problem since this representation will ultimately be used to create the reliable predictor for the supervised task. Finally, the data collection step requires that the collected data be converted to a volume that will not hamper computer system performance.A simpler technique is to construct confidence level entails some form of memorization, which is simplified by the use of a hierarchical model, and associate a confidence with a portion of the model. All of these data preprocessing techniques are preferred when possible since their reliance on actual labeled data limits their general applicability. Furthermore, the methods can only be applied when actual labeled data is available or the data can be re-collected.

5.7. Conclusion

The emergence of pervasive digital technologies has redefined the way the modern economy works – often termed platform economy, which is increasingly becoming interconnected. The sharing economy enabled businesses and individuals to share

services of products for micro-utilization that were previously owned and live dormant and non-utilized for extended periods. Therefore, previously owned but non-utilized assets have gained prominence and played a pivotal role in the economy. The hyper-connectivity powered by internet connectivity and mobile digital devices has ushered in change. Everything around us is continuously sending and receiving messages, enabling faster and smarter decision-making. This inter-connectivity has led to the growth of alternate data sources – coming from either private repositories or social media platforms. With the advancement in AI and machine learning techniques to mine and discover fresh, meaningful, and deep insights from these data, financial institutions are increasingly becoming analytics-driven to support their core business processes. Modeling propensity to consume and propensity to default and other behavioral-related conduct by using alternate sources of data having those characteristics can extend the credit to those who were earlier considered under banked, unbanked, or risky propositions. Large pools of data made available by the new platforms along with advanced analytics are increasingly being used in predictive modeling. This would be the future of credit scoring.

5.7.1. Future Trends

Alternative data has been gaining traction in the credit scoring engine domain. Recent changes in data privacy laws could decrease the available credit history data as those will allow individuals to delete historical data. Adding to the reduced supply of credit data, increased labor market and immigration fluxes may intensify the underbanked problem. All of the above will cause a higher demand for alternative data. Nowadays, there is a market for alternative data to score first-time borrowers. Companies are starting to build credit scores for individuals who for political reasons in their home country would not be able to own a bank account, receive credit, send or receive money transfers, or transfer property. Innovators have created a system in which foreigners can use their digital “identity” to access product funnels in other countries without company-specific middleware. For emerging-market immigrants in host countries or individuals with no credit history in distant commons large enough to house their debts, this would be a transformative step.While small- and microbusinesses struggle to obtain credit due to the lack of financial history, the application of advanced data science can be a market-making tool that enables banks to expand products that will make bank identification card offerings and credit-delivery products even more seamless. Earlier revolutions in business credit risk assessment had negative economic consequences for consumers due to a broader economic fallout from under-investment in credit-rationed conditions. As the innovative destruction of credit markets gains speed, the business risks associated with credit-nondoing credit vis-a-vis their customers will raise the potential for

asymmetric information, creating increasing needs for banks to be conduits for identifying both parties in transactions where values are determined over the longer term.

References

Narula, R., & Van der Straaten, K. (2020). A comment on the multifaceted relationship between multinational enterprises and sustainable development. Transnational Corporations, 27(1), 137–148.

Ng, A. (2020). Machine learning yearning: Technical strategy for AI engineers. deeplearning.ai. Puschmann, T. (2017). Fintech. Business & Information Systems Engineering, 59(1), 69–76.

Ryll, L., Seidens, D., & Schmid, J. (2020). AI in finance: What are the ethical implications? Journal of Risk and Financial Management, 13(12), 309.

Schindler, J. W. (2017). FinTech and financial innovation: Drivers and depth. Finance and Economics Discussion Series 2017-081. Board of Governors of the Federal Reserve System.