



# Unlocking the Value of Multiple Data Sources Through Ensemble Models



January 19-20, 2026

Bengaluru, India

# US Market Situation



Financial institutions of all shapes and sizes aim to grow profitably, control risk and promote financial inclusion



Alternative data has proven effective at finding pockets of creditworthy consumers who are less visible in traditional sources



Availability and regulatory acceptance have been growing steadily — with more recent advances in rental history, cashflow and buy now, pay later (BNPL)



Integrating alternative data with traditional scoring methods can be challenging, and modelers face a complex set of options



*In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.*

**Here, we're building multiple, individual models employing different data sources —and using those model outputs to develop a more powerful model.**

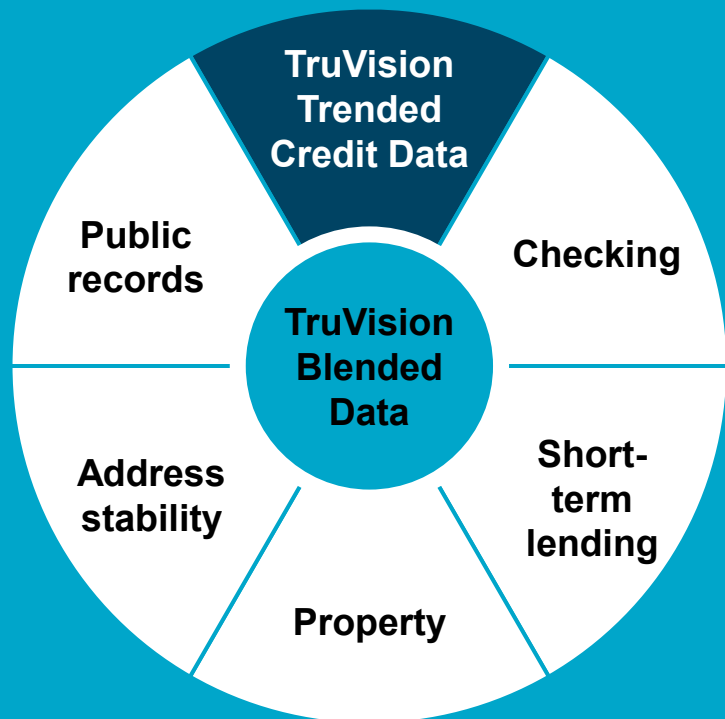


**Unlocking Value through Ensemble Modeling**

# Data Sources

# TransUnion aims to provide the most insightful and comprehensive view of consumer behavior

## TruVision™ Blended Data



### Trended consumer credit data

- Trending 30 months of historical data on a consumer provides a time-series view, helping uncover patterns invisible in point-in-time views
- **New:** Conditional algorithms focusing on consumer behavior after an occurrence of a specific condition like moving, delinquency or new account; expanded suite of magnitude algorithms\*

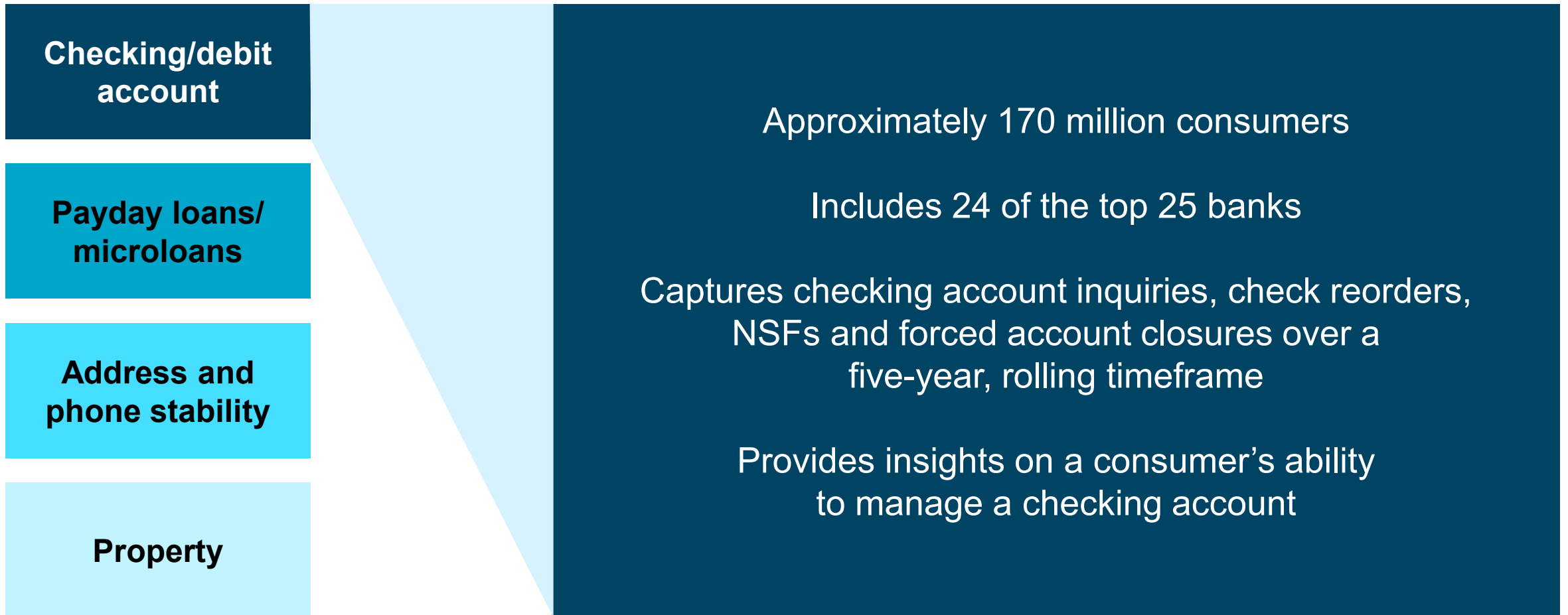
### Alternative data

Over 3 billion non-traditional data records collected on over 260 million Americans (95% of US households), including:

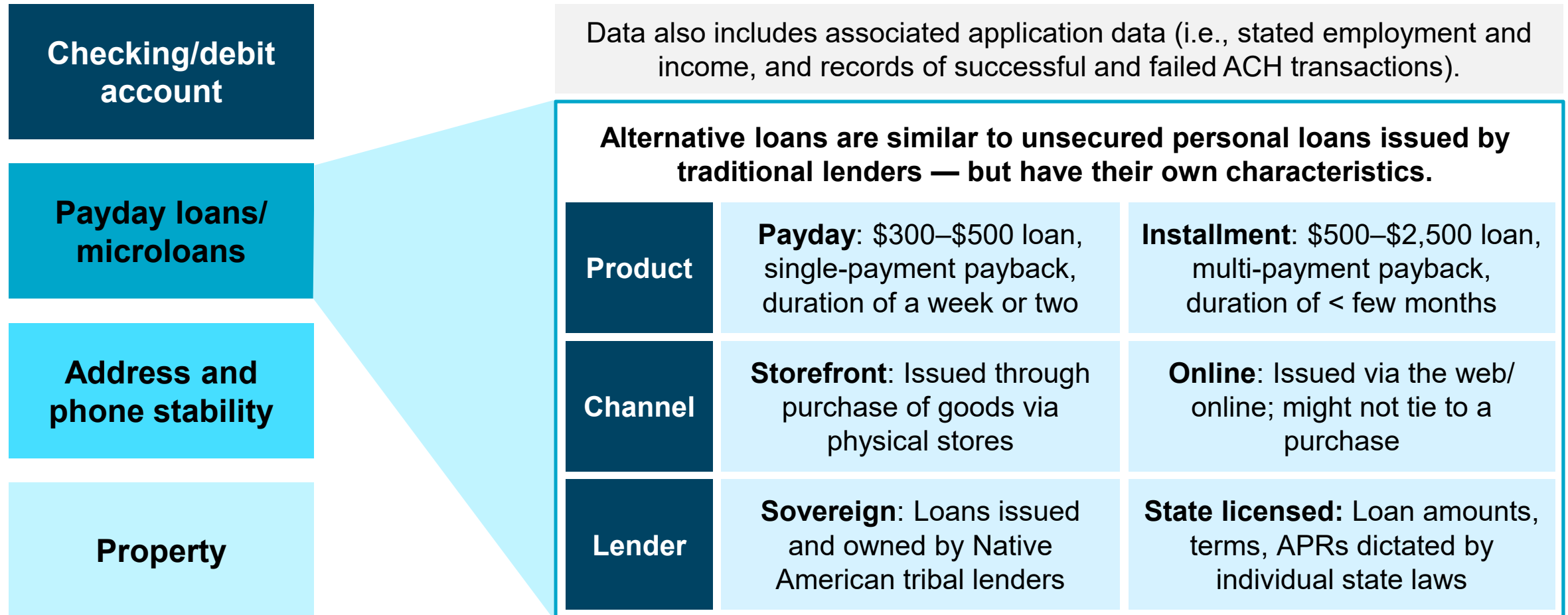
- Checking/debit account
- Payday lending/microloan activity (a.k.a., short-term lending or STL)
- Public records and real estate
- Address and phone stability
- Foreclosures and bankruptcies

\* Magnitude algorithms give greater weight to recent months' directional changes for a given credit measurement. Expansion included utilization, minimum payment due, past due, aggregate excess payment by product

# Checking data captures banking behavior that is unrelated to lending and hence unavailable in the core credit file



# FactorTrust data represent ~62 million consumers with over 650 million detailed STL and RTO transactions over a two-year period



# Address and phone stability attributes have equivalent coverage to TruVision attributes and can be incorporated as such

**Checking/debit  
account**

**Payday loans/  
microloans**

**Address and  
phone stability**

**Property**

The database contains more than 250 million consumers — with up to 20 addresses and phone numbers on each consumer.

Address attributes include number, type of addresses over given lengths of time, and length of time at current residence, as well as changes in state and ZIP. Similarly, phone attributes include number of phone numbers over time and changes in area code.

# Our property data has ~ 90% residential property coverage with 2/3 of updates occurring within a month of transaction

**Checking/debit  
account**

**Payday loans/  
microloans**

**Address and  
phone stability**

**Property**

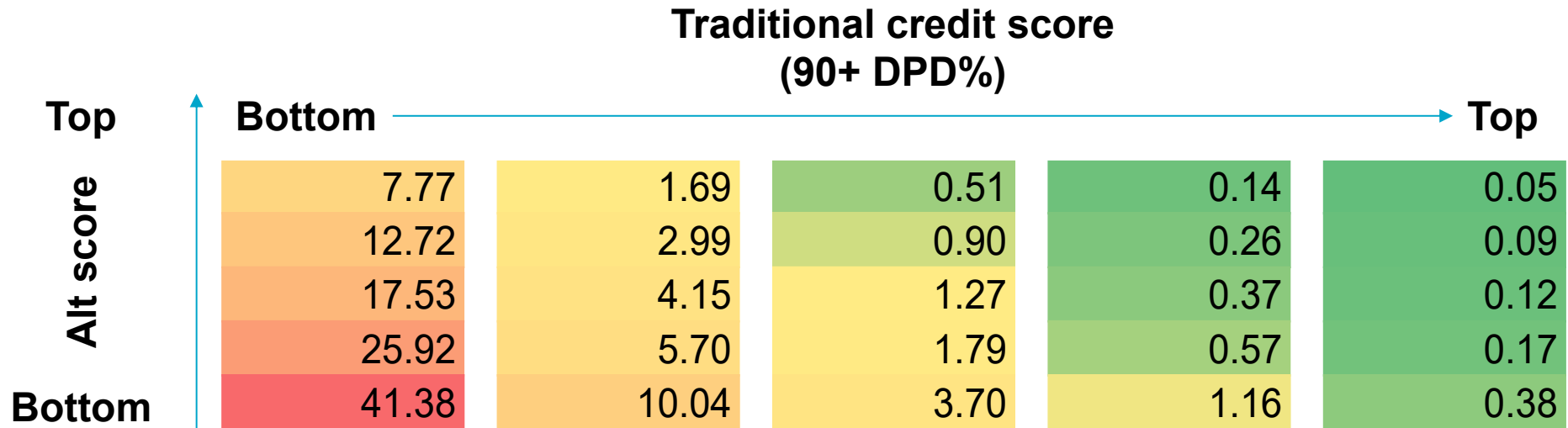
This data comes directly from the county tax and deed commissioners' offices and represents over 96% of the US counties. The data provides annual and/or transaction (property sale) triggered assessed property values, sales prices, tax amounts and land size — and conventional and FHA counters and interest rates.

Property values can be combined with mortgage amounts on the bureau to create a combined loan-to-value (CLTV) ratio.



# Orthogonal data provides independent and uncorrelated information, enhancing overall accuracy and comprehensiveness of risk assessment

In the following example of a dual score matrix, we split the population first on a traditional credit score (e.g., 20% equal bands) — and within each band, we equally split based on an alternative credit data score (~4% per node).



<b>Column total</b>	21.12%	4.96%	1.65%	0.5%	0.16%	<b>Overall</b>
<b>Spread</b>	5.33x	5.94x	7.25x	8.29x	7.6x	5.7%



**Unlocking Value through Ensemble Modeling**

# Methodology

# There are multiple ways to incorporate alternative data

Directly include existing alternative credit score as an attribute

Cascading when there is little to no traditional credit data

Create dual score matrix by overlaying alternative credit data score and traditional data score

Build custom model in a modular fashion and treat alternative data attributes differently from traditional attributes

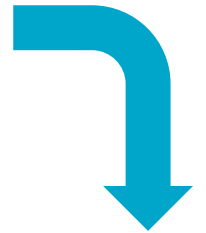
**Within this presentation, we will discuss best practices in custom model build**

# ... but combining alternative and traditional attributes in the same scorecard generally doesn't yield the best possible result

Traditional credit attributes have **higher coverage** (hit rate) than alternative credit attributes.



Lower coverage means there is **lower variance** within the overall population, which leads to little or low predictive power overall.

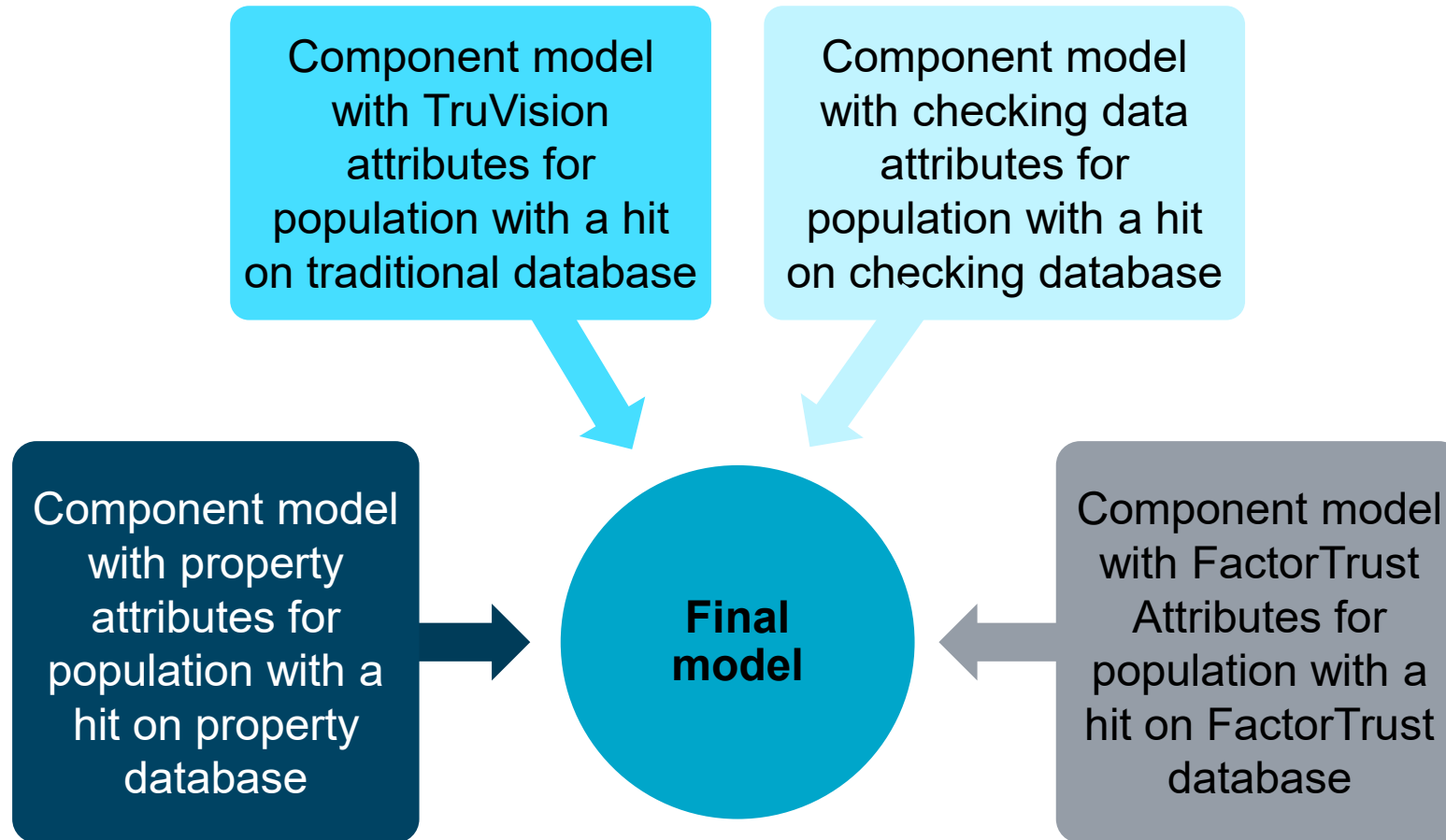


Using alternative data in a **modular fashion** allows lenders to exploit these data to their full advantage and capture the predictive power.



Simply using alternative and traditional credit attributes together in a **"kitchen sink"** fashion leads to loss of potentially valuable information from additional dimensionality.

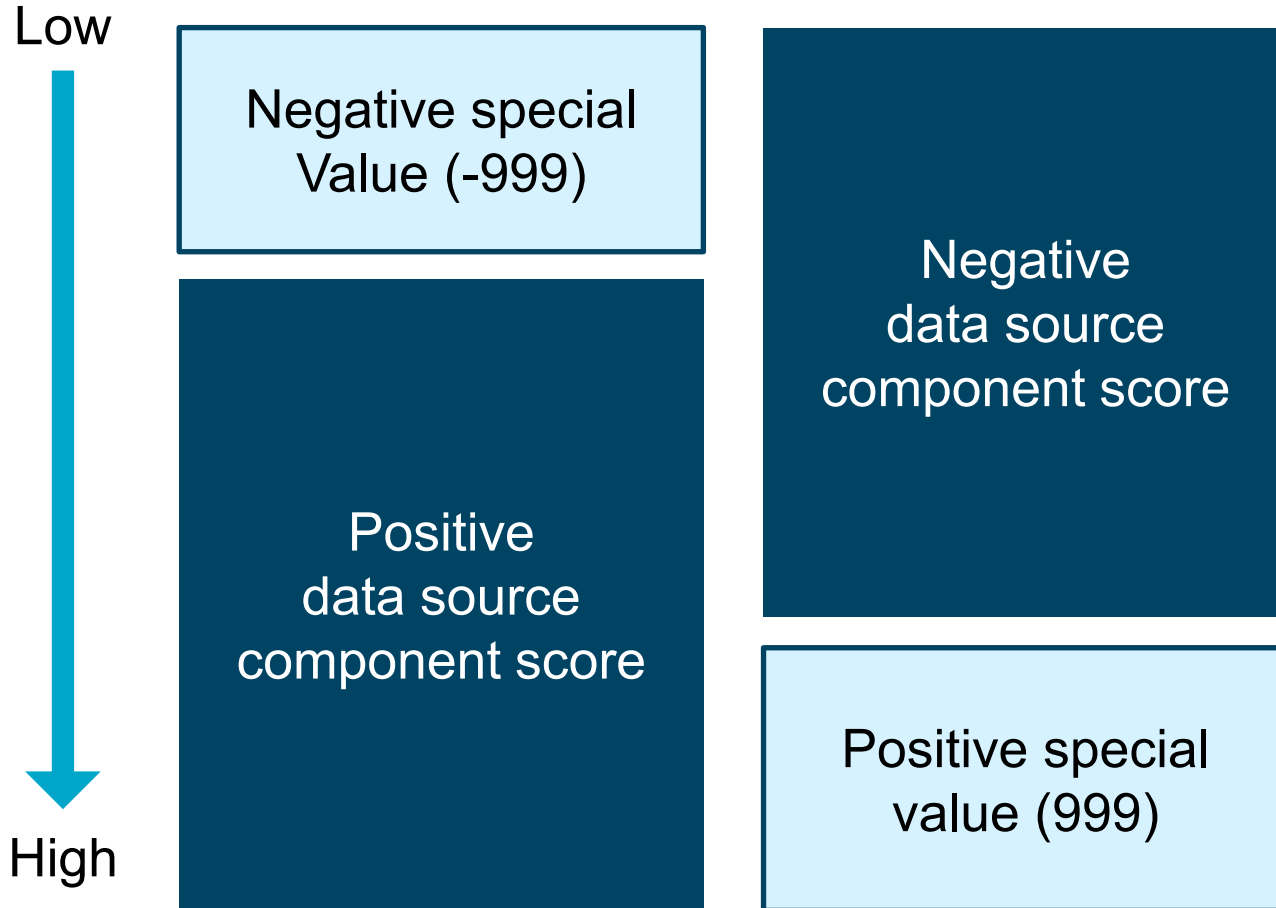
# Component models are trained on the distinct data sources, with component scores used as inputs to the ensemble model



## Model iterations

- Depending on data coverage, nature of product and population, not all the component models are suitable to be used in ensemble model
- Different combinations of component models should be created to select the optimal ensemble

# An alternative data “hit” could have positive or negative connotations depending on its nature



Alternative Data Source	Hit vs No-Hit
FactorTrust (STL) Data	—
Checking data	—
Property data	+
Traditional credit data	+

# Segmentation depends on the method used to construct component models

## XGBoost

Segmentation is typically not required for traditional credit data unless there are business reasons. XGBoost is able to capture interaction between attributes. It iterates through all features' potential split points to identify the one that maximizes gain and effectively partitions the data into segments based on the most predictive features and values.

## Constrained logistic regression model

Traditional component model(s) follows the same rules and conventions as non-ensemble models. Segments can be defined by data analysis or business rules.

- Segmentation is typically only done on the TruVision data source or other sources with a high hit rate and sufficient volumes
- If segmentation is carried out, traditional component models should be equalized before ensembling
- Once traditional component models are equalized, they can be scored and treated as one variable in the ensemble



**Unlocking Value through Ensemble Modeling**

**Use Case: Prime Credit Card**

# Our study used a bureau extract of prime card originations in 2023 to compare outcomes from an ensemble versus a kitchen sink model

- Sampled **approx. 201K** originations within 2023, deduplicating on one trade per consumer per quarter for consumers with **VantageScore® 4.0 > 660** a quarter prior to originations.
- Used performance definition of **90+ DPD ever** for a **~2.4%** bad rate.
- Built **high-depth XGBoost** models with **monotonic constraints**. Used **hyperparameter tuning** to select the desirable parameters. Conducted **correlation checks** to identify any overly correlated attributes and provide additional model stability. All attributes for both approaches received the same treatments/recodes if applicable.

Data source	Hit rate
Traditional credit data	100%*
FactorTrust data (Link F)	8.59%
Property data (Link B)	64.10%
Checking data (Link A)	45.81%

Population	%	Bad %
Train	60%	2.4%
Test	20%	2.3%
Validation	20%	2.5%

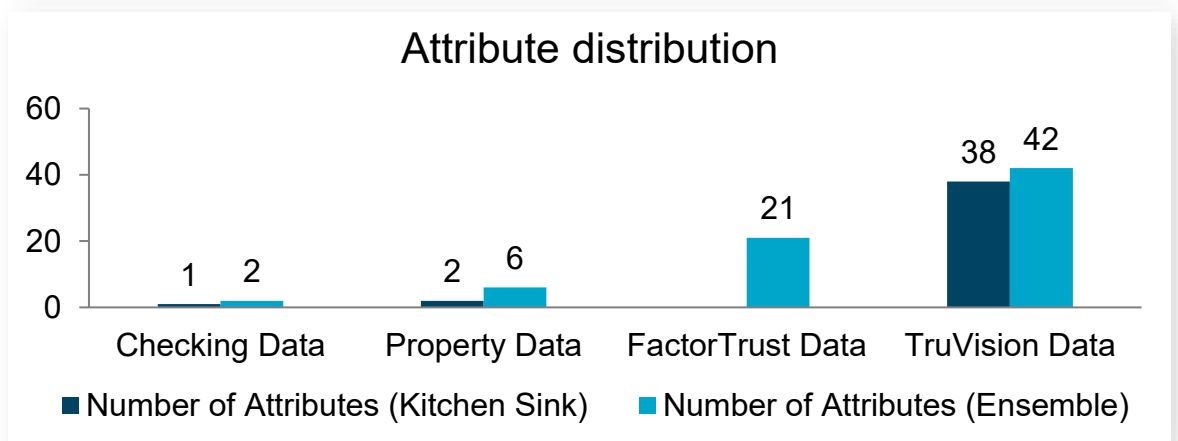
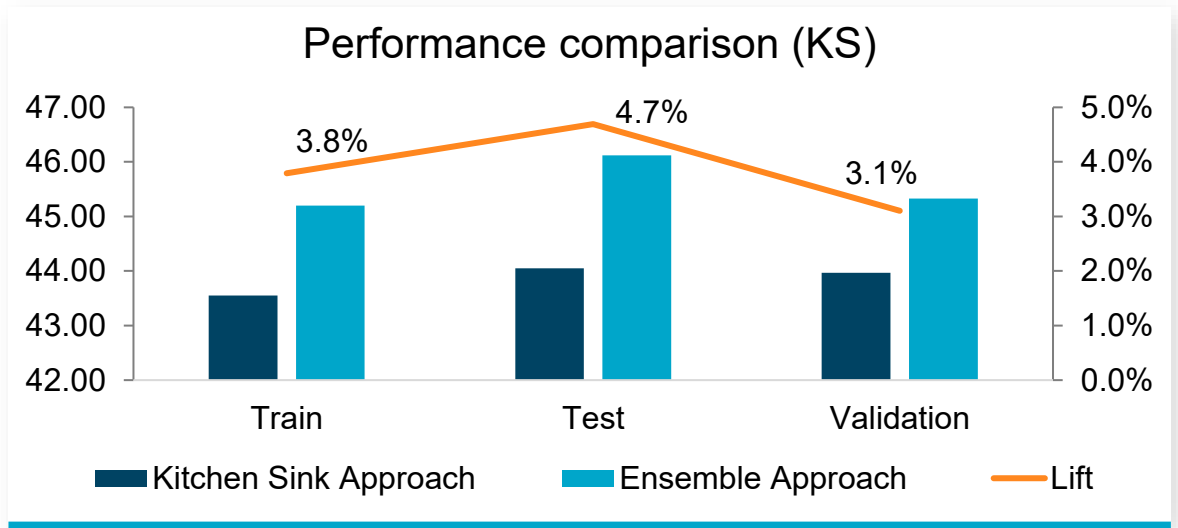
\* Due to bureau extract, hit rate for TU database is 100%. If client input data were the sample source, there could be consumers with an alternative data hit, but no hit on the traditional credit database.



# Ensembling not only outperforms the kitchen sink approach, it also allows more alternative attributes to be included

Capture rate				
	Test		Validation	
Score range	Kitchen sink	Ensemble	Kitchen sink	Ensemble
Top 10th	100.0%	100.0%	100.0%	100.0%
90th	99.6%	99.8%	99.6%	99.4%
80th	98.9%	98.9%	98.2%	98.4%
70th	96.8%	97.4%	95.8%	96.9%
60th	93.8%	93.6%	93.2%	94.0%
50th	89.2%	89.3%	89.3%	88.9%
40th	82.5%	83.4%	82.0%	84.1%
30th	72.4%	77.7%	71.4%	74.3%
20th	59.2%	60.9%	58.6%	62.1%
Bottom 10th	39.1%	40.7%	39.5%	41.4%
Lift		7.3%		4.1%

Gain	TruVision	FactorTrust	Checking	Property
Ensemble	88.3%	6.2%	2.8%	2.6%
Kitchen sink	94.7%	NA	1.9%	3.4%



# Ensemble methods allow for modular flexibility in the management of multisource models

## Compliance

Changes in **government regulatory practices** and standards or changes in **internal compliance requirements** could necessitate changes to the model.



## Data source

Changes in **lenders' reporting practices**, or addition, adjustment or removal of product offerings by lenders which changes the **data availability** could impact coverage of the database. Additional data sources could also be introduced.



## Cost

Changes in **internal budget** or **cost of data** based on a multitude of factors could result in a need to adjust a model built with multiple data sources.



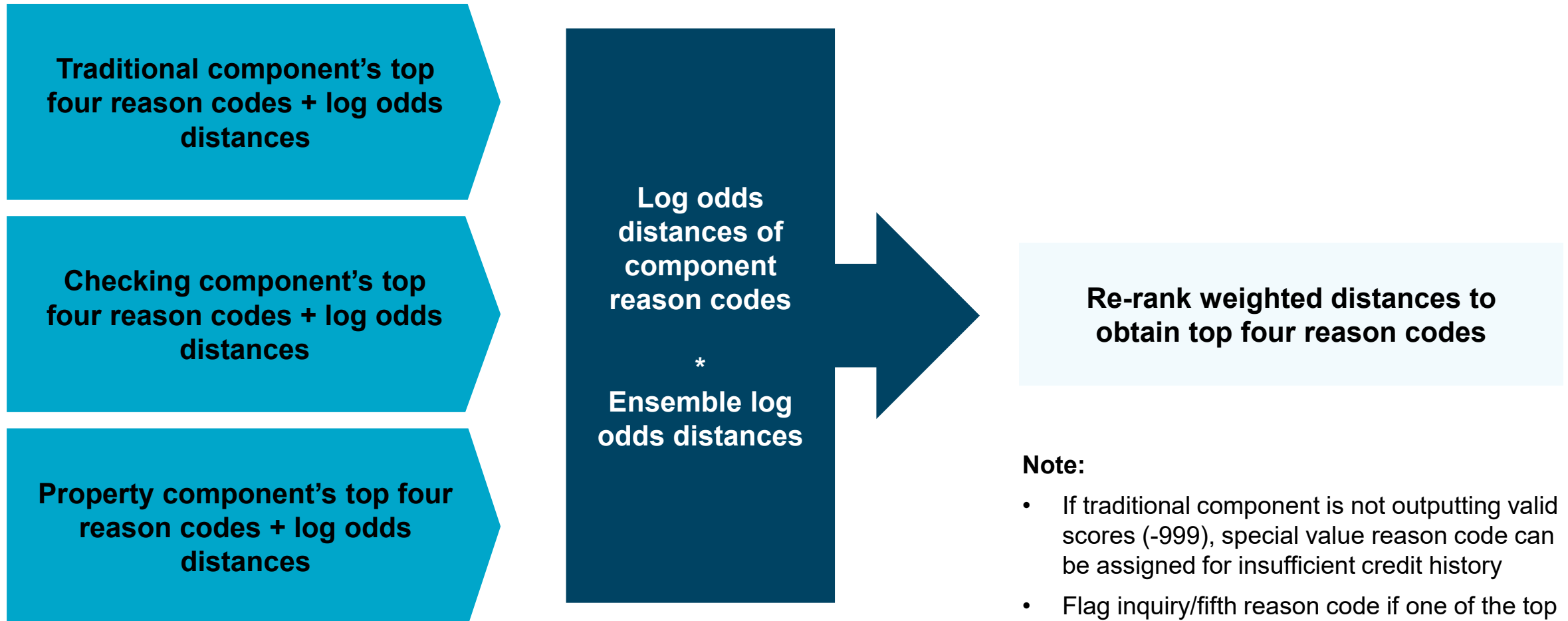
Building ensemble model in a modular fashion would provide flexibility for lenders to quickly adapt to changes in the business environment and requirements without significant overhead.



**Unlocking Value through Ensemble Modeling**

# Adverse Action

# Adverse actioning (AA) requires weighted distances that account for the components *and* the ensemble model

**Note:**

- If traditional component is not outputting valid scores (-999), special value reason code can be assigned for insufficient credit history
- Flag inquiry/fifth reason code if one of the top four AA codes are not related to inquiries

# Consider this example of a two-component ensemble model applied to Consumer X

Hypothetical ensemble model				Component model returns			Top	Wgt dist [B] x [C]	AARC	Source
Consumer X	[A] Model beta	[B] dist Max	[C] lo dist	Reason Code	[D] = [B] x [C] wgt dist					
750	TruVision Trended Score						<b>1</b>	0.489016	954	Checking
	(-Inf,574.5]	-2.4520	4.1698	0.404931	150	0.232025				
	(574.5,602.5]	-1.6611	3.3789	0.210507	144	0.120621	<b>2</b>	0.270387	951	Checking
	(602.5,653.5]	-1.1931	2.9109	0.132709	157	0.076042	<b>3</b>	0.232025	150	TruVision
	.	.	.	0.000000		0.000000				
	.	.	.				<b>4</b>	0.120621	144	TruVision
	(744.5,756.5]	1.1448	0.5730				<b>5</b>	0.076042	157	TruVision
613	Checking Score						<b>6</b>	0.059394	950	Checking
	(-Inf,611.5]	-0.5354	0.6614	1.431127	954	0.489016				
	(611.5,631.5]	-0.2157	0.3417	0.791300	951	0.270387	<b>7</b>	0.001245	953	Checking
	.	.	.	0.173820	950	0.059394				
	.	.	.	0.003645	953	0.001245	<b>8</b>	0.000000	NA	TruVision
	(689.5,690]	0.1250	0.0010							
(690,Inf]	0.1260	0.0000								

# XGBoost uses SHAP values, not log-odds distances, for selecting top four reason codes — but leverages the same logic

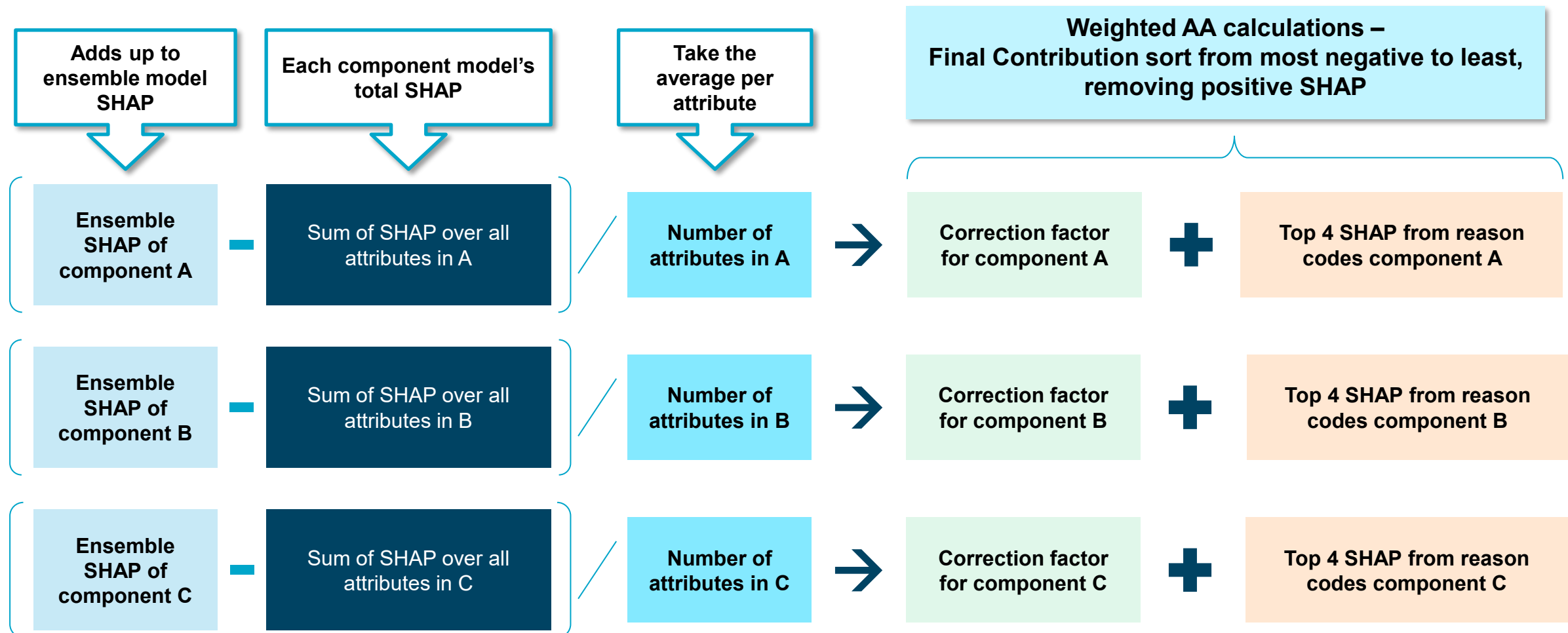
## A brief explanation of SHAP

- **Shapley values** are a game theory-based method for fairly assigning payouts to players depending on their contributions to the total payout in a cooperative game.
- **SHAP (SHapley Additive exPlanations)** is a framework for interpreting model predictions through the **Additive Feature Attribution Method**.
- **SHAP value** is a unified measure of feature importance from the framework. It's practical for implementation using approximations of Shapley values by sampling instead of calculating each possible combination (a model with  $n$  features contains  $2^n$  possible combinations).

## SHAP properties

1. **Additivity**
  - Contribution of each feature to the final prediction can be computed independently and summed up
2. **Local accuracy**
  - Sum of all features' SHAP values equals the difference between the model's actual prediction for a particular instance and a defined baseline prediction
3. **Consistency**
  - SHAP values do not change when the model changes unless contribution of a feature changes
4. **Missingness**
  - A missing feature gets an attribution of zero

# We can calculate the equivalent of AA distances and generate adverse action reason codes using properties of SHAP values






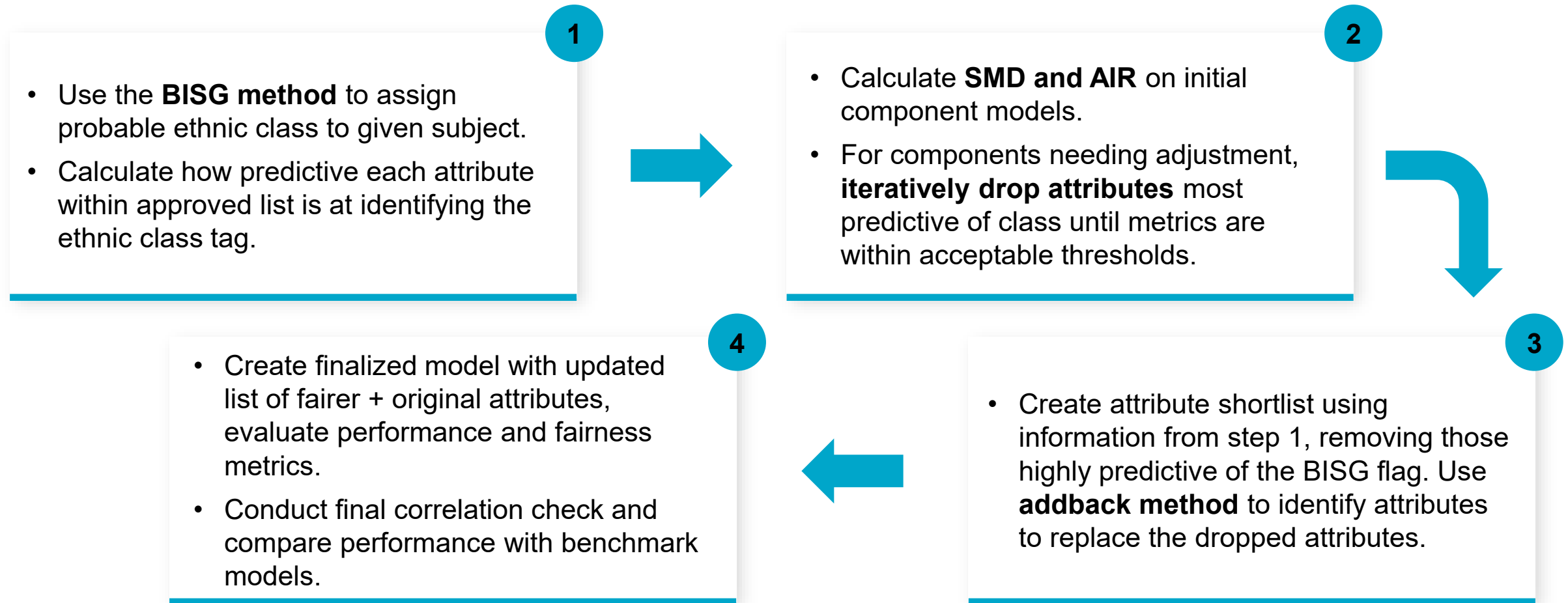
**Unlocking Value through Ensemble Modeling**

# Evaluating Fairness

# We can follow similar industry best practices to maintain fairness while incorporating alternative data in a custom model

Exclude variables	Correlation check	Model build	Fairness testing 
<ul style="list-style-type: none"><li>• Exclude variables that are perceived as functional proxies for protected class.</li><li>• Different compliance/legal teams might have different standards, and the exclusion list should be discussed prior to model build.</li></ul>	<ul style="list-style-type: none"><li>• Simply dropping some variables might not be sufficient as correlated variables might simply take their places.</li></ul>	<ul style="list-style-type: none"><li>• Use machine learning to train less discriminatory alternatives. Complex techniques generalize well to new data compared to simple modeling methods.</li><li>• Adopt adversarial debiasing technique by creating primary/adversary models.</li></ul>	<ul style="list-style-type: none"><li>• Conduct less discriminatory alternative (LDA) search to remove or substitute individual attributes predictive of protected class — while monitoring performance loss and satisfying legitimate business justification.</li></ul>

# The goal of fairness testing is to seek a less discriminatory alternative that still satisfies the legitimate business justification



**Note:** BISG = Bayesian Improved Surname Geocoding, SMD = Standard Mean Difference, AIR = Adverse Impact Ratio

# Observations from recent fair lending analyses on ensemble models

- Fewer fairness concerns have surfaced within **alternative data components** as alternative data coverage is more focused on either positive or negative dimensions of consumer behavior compared to wide-ranging core credit data.
- The core credit component's fairness metrics influence ensemble fairness metrics — strongly due to its **heavy contribution to ensemble model**.
- When ensembling all components, fairness metrics may show slightly more concern when combining all the different dimensions of data compared to each component individually.
  - The ease of adding or removing components allows for quantifying each component's effect on fairness through changes in fairness metrics.

# Unlocking the Value of Multiple Data Sources Through Ensemble Models

- Alternative data provide **orthogonality** on top of traditional credit data. They introduce new dimensions of credit behavior for lenders to assess consumer risk. Building a custom risk model in a modular fashion is one effective way to adopt alternative data.
- Ensemble models are assembled from **component models** with **data source-specific attributes and a data source-specific population**. Modelers evaluate which permutation of data sources provides the best result.
- There are differences in the model build process and adverse action code assignment depending on the chosen **modeling methodology**.
- Modular methods offer potential **cost, compliance and data availability** advantages while still delivering strong performance.