The data sources used for the PDV-ML 2024Q4 model are appropriate, sufficient, and aligned with the model’s business objectives. The model relies on enterprise-managed sources including EPDI/PTD (property transactions), TRAX (foreclosure sales), EDL-EDW (loan attributes), CAP (AVM values), and Oasis Gold (BPO values). These sources are refreshed on a monthly or daily cadence and are governed under standard enterprise data SLAs. The data is reliable, complete, and representative of the REO population being modeled.

The documentation indicates that data for model development and scoring is fully aligned with the disposition date (or current scoring date, for ongoing prediction), ensuring temporal consistency. The data also undergoes filtering, deduplication, and transformation based on business logic and modeling needs. Feature derivation methods, such as AVM imputation, KNN-based local comps, and year-built proxies, are transparently documented. Data sufficiency is further demonstrated by consistent 100% hit rates across all validation cycles.

No standalone external datasets beyond enterprise-managed sources are used in model development. However, the model integrates third-party data (e.g., Freddie Mac and Black Knight transactions) through the EPDI/PTD and TRAX databases, which are centrally managed and incorporated into the enterprise data infrastructure. These external sources are subject to internal deduplication and quality checks to ensure consistency. The inclusion of these third-party records is justified for enhancing REO data volume and diversity, particularly in non-Fannie segments, and their use is clearly described and validated.

The model uses a curated data frame composed of over 50 variables spanning property characteristics, loan attributes, transaction details, and derived features. Source systems include EDPI, EDL-EDW, CAP, and TRAX. Feature selection is informed by correlation to the target (REO sale price), business judgment, and data completeness. Variables with high missingness or low predictive value (e.g., cash, tract\_id) are excluded.

Sampling spans three years of REO transactions for the final model, and five years for intermediate models (e.g., KNN and AVM imputation). This ensures balanced temporal coverage and mitigates seasonality and volume-related bias. Each monthly cycle includes ~10,000 transactions, and sampling excludes records with invalid or missing sale prices. Final model inputs are numeric and compatible with XGBoost, following preprocessing steps.

The dataset structure is well-documented, and upstream system linkages are stable and governed, with no licensing or access concerns identified.

There are no fine-tuning or customizations applicable under the definition of GenAI models. However, intermediate model components (KNN and AVM Imputation) are re-trained periodically using updated data and retuned hyperparameters. These models are treated as part of the broader deterministic modeling pipeline rather than standalone ML applications. The final model configuration is consistent with prior versions and enhancements in the 2024Q4 release are limited to expanded training windows, revised sampling logic, and recalibrated hyperparameters.

The model aggregates REO transactions from two primary sources: the EPDI/PTD database, which includes all historical property transactions (regular and foreclosure), and the TRAX database, which is restricted to foreclosure transactions but is updated more frequently. These data sources are aggregated using **street address and ZIP code** as the composite key. Where loan ID is available, it is used as the primary key; otherwise, address-based mapping is applied.

The aggregation process consolidates property, loan, AVM, tax, and BPO records for each REO transaction to create a unified and enriched sample. Data from third-party providers (Freddie Mac and Black Knight) is incorporated through EPDI/PTD and TRAX, and prioritized using predefined logic. The aggregation approach is appropriate and ensures that all required features are available for training and scoring, while minimizing redundancy and conflicts across systems.

A multi-step data cleaning protocol is applied to ensure sample integrity:

* **Duplicate Removal:** REO transactions that appear in both EPDI and TRAX are deduplicated using logic based on transaction date gaps (365-day threshold), price similarity, and a data provider hierarchy (Fannie > Freddie > Black Knight).
* **Invalid Records:** Transactions with zero or negative REO prices are filtered out. For properties with more than three records across providers or unresolved conflicts, all records are dropped.
* **Missing Data Handling:** Where critical fields like AVM or year-built are missing, proxy values are substituted—e.g., tax assessment year is used when loan year-built is unavailable.
* **Categorical Normalization:** Variables such as pool\_ind and occ\_stat are transformed to binary or numeric indicators. For pool indicator, ‘N’ and missing values are treated as no-pool; for no\_stors, rare fractional values (<0.7% of records) are retained and left for the model to interpret.

All assumptions and cleaning steps are consistent with the prior production version and are explicitly documented in the MDD.

The raw variables are processed and transformed into model-ready features through a combination of domain logic and statistical methods:

* **AVM Imputation:** Missing AVM values are imputed using an intermediate XGBoost model, trained on properties with valid AVM values. The model excludes cscore\_r per MRM guidance and is tuned using BayesSearchCV.
* **KNN Local Features:** Three engineered features—REO price per sqft, raw REO price, and haircut (REO/AVM)—are generated using KNN based on Cartesian coordinates. Neighbor counts are optimized (k=5 for comps, k=0 for haircut in final version) based on performance metrics.
* **Derived Dates:** Variables such as loan\_age, time\_since\_disp, and days\_REO\_to\_DISP are derived from original and disposition dates.
* **Final Formatting:** All features are transformed to numeric format, as required by XGBoost. Categorical fields are encoded as integers or dummy values where needed.

Transformations are appropriate, transparent, and support model estimation and downstream interpretability.

Comprehensive data quality checks were implemented during model development to ensure completeness, consistency, and relevance of the final dataset. All source systems (EPDI, TRAX, EDL-EDW, CAP) are enterprise-governed and refreshed regularly, with defined SLAs ensuring timely updates. Final datasets used for training and scoring achieved a 100% hit rate across monthly cycles, demonstrating full record resolution.

Business rules were applied to filter invalid records (e.g., non-positive REO sale prices), deduplicate overlapping transactions across data providers, and convert categorical indicators (e.g., pool, occupancy status) into numeric form for model ingestion. These transformations are clearly documented and aligned with modeling standards. Issues arising from third-party data dependencies (e.g., missing AVM or year-built) were addressed using proxy fields or imputation models, with rationales justified and tested.

Summary statistics reported in the MDD confirm data sufficiency and representativeness. Monthly training samples include approximately 10,000 transactions, covering a balanced distribution across time. The three-year window used for training ensures seasonal diversity, reducing market-cycle bias. AVM imputation is applied to fewer than 5% of records, and categorical encoding preserves data integrity.

Performance metrics such as PPE20 and PPE30 improved significantly over the prior version, reflecting data quality and relevance. No material issues were observed in the sample distribution, frequency of key attributes, or outlier behavior.

Several data proxies were used appropriately in the model:

* **AVM Values:** For records missing AVM (~4–5%), a dedicated XGBoost-based imputation model fills the gap. The model is trained on complete records and excludes confidence scores (cscore\_r) per MRM feedback.
* **Year Built:** If the year built from loan data is missing, it is substituted using the tax assessment year. This ensures consistency across older properties or missing originations.
* **Pool Indicator:** High missingness (~94%) is observed. All null and ‘N’ values are treated as no-pool (binary value 0), with ‘Y’ as pool-present (value 1), based on field expert guidance.

These proxies are well-justified and mitigate potential biases. They do not materially distort the model’s predictive behavior or fairness.

Data representativeness is preserved by using a recent 3-year window for final model training. This balances seasonality, avoids pandemic-era shocks, and maintains alignment with current market conditions. All training data is disposition-aligned and scoring data uses the accounting month as a proxy for disposition.

Correlations among variables were assessed to ensure relevance. Features with near-zero correlation to the target or redundant relationship with other variables (e.g., tract\_id, std\_use\_cd, cash) were excluded. Variable distributions reflect a realistic cross-section of Fannie Mae’s REO portfolio.

Partial dependence plots and exploratory statistics indicate that key features (e.g., AVM, orig\_val, sqft) drive model output without evidence of bias or instability across time or segment.

No image, video, textual, or high-dimensional unstructured data is used in this model. All input features are structured, tabular variables derived from loan and property datasets. Therefore, this section is not applicable (N/A).

The model employs **purposive sampling** of REO sales and disposition transactions to train on the most relevant and recent market behavior. For the 2024Q4 version, the training dataset includes a 36-month window (March 2021 to March 2024) of REO transactions, an extension from the prior 26-month window used in earlier versions. This adjustment was made to balance seasonal effects and ensure each calendar month is equally represented.

Intermediate models such as KNN and AVM imputation use a broader five-year window to capture neighborhood-level variability. Invalid transactions (e.g., zero or missing REO price) are excluded, and duplicated transactions across data providers are resolved using predefined logic.

This approach is appropriate for the model’s objective of estimating current distressed property values under changing market conditions.

A comparative analysis was conducted to assess representativeness of the sampled data against the broader REO population. Features showing weak correlation to the target or redundancy with stronger variables were excluded (e.g., tract\_id, cash). Visualizations and statistics indicate that the sample is broadly representative in terms of property type, location, and loan characteristics.

The 36-month training window reduces recency bias while avoiding distortions from pandemic-era transactions. Seasonal balance is achieved by selecting three full years rather than most recent months only.

### **3.4.3 Sampling Methods Impact on Model Performance**

The extended training window improved model performance substantially. Out-of-sample testing shows:

* **PPE20** (accuracy) improved to 75–80% vs. 44–52% in the prior version.
* **PPE30** (tail error) reduced to 12–16% vs. 34–42% previously.
* **WAPE and WPPE** metrics are closer to zero, indicating more stable and precise predictions.

In-sample 100-fold cross-validation confirms similar gains. The new sampling logic contributes directly to performance consistency and robustness across cycles.

Two primary data limitations were identified and appropriately mitigated:

* **Missing AVM Values:** AVM, the most predictive feature, is missing in ~4–5% of records. A separate imputation model using XGBoost fills these gaps effectively. The model is tuned using non-missing data and excludes confidence score variables per MRM guidance.
* **High Missingness in Pool Indicator:** Around 94% of pool values are missing or marked as 'N'. Based on business logic and expert input, all missing values and 'N' are grouped as no-pool (value = 0), and 'Y' as pool-present (value = 1).

Other missing features like year-built are addressed using tax proxy data. Categorical variables are transformed into numeric formats to comply with XGBoost input constraints. These mitigations are reasonable and do not materially impact the model's predictive capability.

The model data used in the PDV-ML 2024Q4 version is deemed appropriate, sufficient, and well-aligned with the model’s objectives. All source systems are enterprise-managed with adequate update frequencies and documented transformations. Sampling methodology is well-justified, and feature engineering steps—such as AVM imputation and KNN-based comps—are documented and validated.

Mitigations for data limitations are appropriate and consistently applied. Performance improvements in both in-sample and out-of-sample testing support the adequacy of the data. No significant concerns are noted regarding representativeness, accuracy, or stability.

**MRM concludes that the model data is reasonable for use in model development and scoring for the current cycle.**