ISBN Lot Optimizer

Machine Learning System

Comprehensive Technical Documentation

Report Date: November 2025

Version: 1.0

Executive Summary

The ISBN Lot Optimizer is a sophisticated machine learning system designed to predict book prices across multiple online marketplaces. This document provides a comprehensive snapshot of the system as of November 2025, including data sources, model architecture, training procedures, and current performance metrics.

Metric	Value
Total Training ISBNs	19,384
ISBNs with eBay Sold Comps	11,134 (57.4%)
eBay Model MAPE	11.9%
eBay Model R^2	0.788
AbeBooks Model MAPE	17.7%
Total Features	87
Active Specialist Models	6

1. Data Sources

The system aggregates book pricing data from multiple platforms to build comprehensive market intelligence. Each platform provides unique insights into different market segments.

1.1 Primary Data Platforms

Platform	Type	Data Focus	Update Frequency
eBay	Auction/Marketplace	Sold & active listings	Real-time
AbeBooks	Specialty Books	Collectible/rare books	Daily
Amazon FBM	Marketplace	Third-party sellers	Daily
Biblio	Independent Sellers	Used/rare books	Daily
Zvab	European Market	German-language market	Daily
Alibris	Used Books	General used market	Daily
BookScouter	Buyback	Wholesale buyback prices	On-demand

1.2 Metadata Sources

Book metadata (titles, authors, publication details, ratings) is sourced from:

• Google Books API - Primary metadata source with comprehensive coverage

• Open Library - Fallback for missing data and additional edition information

• Internal caching system - Reduces API calls and improves response times

2. Data Structure

2.1 Database Architecture

The system uses three SQLite databases for different purposes:

Database	Location	Purpose	Key Tables
catalog.db	~/.isbn_lot_optimizer/	Market data storage	ebay_sold_listings, ebay_active_listings, sold_comps
books.db	~/.isbn_lot_optimizer/	Book metadata	metadata, series_lot_comps, edition_data
app.db	isbn_web/	Web app state	user_data, sessions

2.2 Core Data Models

The system uses typed Python dataclasses for data validation and type safety:

BookMetadata - Comprehensive book information

```
isbn, title, subtitle, authors, publisher, publish_date, page_count, language, cover_type,
description, categories, rating, ratings_count, series_name, series_position, signed,
printing, edition, isbn10, isbn13, raw
```

EbayMarketStats - eBay market intelligence

```
sold_count, sold_min, sold_median, sold_max, sold_mean, sold_price_spread, active_count,
active_min, active_median, active_max, active_vs_sold_ratio
```

BookScouterResult - Buyback pricing

```
isbn, top_buyback_price, avg_buyback_price, num_vendors, fetch_date
```

3. Feature Engineering

The ML models use 87 features across 10 categories. Features are designed to capture market dynamics, book attributes, and demand signals.

3.1 Feature Categories

Category	Count	Examples
eBay Market Signals	11	ebay_sold_median, ebay_sold_price_spread, ebay_active_vs_sold_ratio
AbeBooks Market	7	abebooks_min_price, abebooks_avg_price, abebooks_count
Amazon FBM Market	8	amazon_fbm_min_price, amazon_fbm_count, amazon_fbm_price_variance
Biblio Market	7	biblio_min_price, biblio_avg_price, biblio_count
Alibris Market	7	alibris_min_price, alibris_avg_price, alibris_count
Zvab Market	7	zvab_min_price, zvab_avg_price, zvab_count
Book Attributes	7	page_count, age_years, log_ratings, rating, is_textbook, is_fiction
Condition	7	is_new, is_like_new, is_very_good, is_good, is_acceptable, is_poor
Format	3	is_hardcover, is_paperback, is_mass_market
Special Attributes	5	is_signed, is_first_edition, demand_score, edition_score
Series Data	11	series_lot_min, series_lot_median, series_lot_max, lot_demand_score

3.2 Feature Engineering Techniques

- **Log Transformations:** Applied to skewed features like ratings_count (log_ratings = ln(ratings_count + 1))

- **Ratio Features:** Capture market dynamics (ebay_active_vs_sold_ratio = active_median / sold_median)

- **Spread Features:** Measure price variance (ebay_sold_price_spread = sold_max - sold_min)

- **Temporal Features:** Book age calculated from publish_date (age_years = 2025 - publish_year)

- **Demand Scores:** Composite metrics combining multiple signals (demand_score from ratings + sales velocity)

- **One-Hot Encoding:** Categorical features converted to binary indicators (condition, format, etc.)

4. Model Architecture

4.1 Stacking Ensemble Design

The system uses a two-tier stacking ensemble architecture where specialist models make initial predictions, and a meta-model combines their outputs for final predictions.

4.2 Specialist Models (Tier 1)

Model	Features	Training Data	MAPE	R^2
eBay Model	24 features	11,134 ISBNs	11.9%	0.788
AbeBooks Model	~30 features	~8,000 ISBNs	17.7%	N/A
Amazon FBM Model	~32 features	~6,000 ISBNs	N/A	N/A
Biblio Model	~30 features	~5,000 ISBNs	N/A	N/A
Alibris Model	~30 features	~4,500 ISBNs	N/A	N/A
Zvab Model	~30 features	~3,500 ISBNs	N/A	N/A

4.3 Meta-Model (Tier 2)

The meta-model combines predictions from all specialist models to produce the final estimate. It learns optimal weighting strategies based on each specialist's performance across different book types and market conditions.

4.4 XGBoost Configuration

All models use XGBoost (Extreme Gradient Boosting) with the following configuration:

- **n_estimators:** 300 trees
- **max_depth:** 6 (prevents overfitting)
- **learning_rate:** 0.05 (conservative for stability)
- **subsample:** 0.8 (row sampling)
- **colsample_bytree:** 0.8 (column sampling)
- **min_child_weight:** 3 (regularization)
- **objective:** reg:squarederror (regression)
- **tree_method:** hist (efficient for large datasets)

5. Training Process

5.1 Data Preparation Pipeline

Step 1: Data Collection - Aggregate data from all platforms (eBay, AbeBooks, etc.)

Step 2: Data Cleaning - Remove duplicates, handle missing values, validate price ranges

Step 3: Feature Extraction - Generate all 87 features from raw data

Step 4: Target Transform - Apply log1p(price) transformation to stabilize variance

Step 5: Sample Weighting - Apply temporal decay (365-day half-life) and price type weighting

Step 6: Train/Test Split - GroupKFold by ISBN to prevent data leakage

5.2 Sample Weighting Strategy

The system uses sophisticated sample weighting to prioritize recent and sold data:

- **Temporal Decay:** weight = exp(-days_old / 365) [365-day half-life]

- **Price Type Weight:** Sold prices: 3.0x, Active prices: 1.0x

- **Final Weight:** final_weight = temporal_weight × price_type_weight

5.3 Cross-Validation

GroupKFold cross-validation ensures that all instances of the same ISBN appear in either the training or test set, never both. This prevents data leakage and provides realistic performance estimates for predicting prices of unseen books.

Configuration:

- n_splits: 5 folds

- Grouping: By ISBN

- Metrics: MAE, RMSE, R^2, MAPE

- Feature importance: SHAP values calculated post-training

6. Feature Importance Analysis

6.1 eBay Model Top 10 Features

Rank	Feature	Importance	Category
1	ebay_sold_median	24.8%	Market Signal
2	ebay_sold_min	11.5%	Market Signal
3	age_years	10.2%	Book Attribute
4	demand_score	8.9%	Demand
5	ebay_sold_max	7.3%	Market Signal
6	ebay_sold_price_spread	6.1%	Market Signal
7	is_signed	1.4%	Special Attribute
8	ebay_sold_count	5.8%	Market Signal
9	ebay_active_median	4.7%	Market Signal
10	ebay_active_vs_sold_ratio	3.2%	Market Signal

6.2 Key Insights

• **Market data dominates:** eBay-specific features account for ~70% of total importance

• **Age matters:** Book age is the 3rd most important feature (10.2%)

• **Signed premium works:** is_signed feature successfully captures collectible value (1.4%)

• **Demand signals:** Composite demand_score ranks 4th overall

• **Price spread information:** Range between min/max sold prices is highly predictive

7. Current Performance Metrics

7.1 eBay Specialist Model

- **Test MAE:** $1.63 (mean absolute error)
- **Test RMSE:** $6.74 (root mean squared error)
- **Test R^2:** 0.788 (explains 78.8% of variance)
- **Test MAPE:** 11.9% (mean absolute percentage error)
- **Training ISBNs:** 11,134 books with eBay sold comps
- **Training Date:** November 2025

7.2 AbeBooks Specialist Model

- **Test MAPE:** 17.7%
- **Training ISBNs:** ~8,000 books with AbeBooks data
- **Focus:** Collectible and rare book market segment

7.3 Performance by Book Category

Model performance varies by book type. The system performs best on books with strong market data (many comps) and struggles with rare or unusual items.

- **Mass market paperbacks:** Excellent (5-10% MAPE) - abundant comps
- **Popular fiction/non-fiction:** Very good (10-15% MAPE) - good comp coverage
- **Collectible first editions:** Good (15-25% MAPE) - requires signed book data
- **Textbooks:** Variable (15-30% MAPE) - edition sensitivity
- **Rare/antiquarian books:** Challenging (>30% MAPE) - limited comps

8. API Integration & Deployment

8.1 Price Estimation Endpoint

The primary API endpoint provides real-time price estimates with confidence scores:

```
Endpoint: POST /api/books/{isbn}/estimate_price
```

Request Body:

```
{"condition": "Good", "is_signed": false, "is_first_edition": true, "is_hardcover": true,
"is_paperback": false}
```

Response:

```
{"price": 45.23, "confidence": 0.85, "reason": "ML prediction based on 89% of features",
"deltas": [{"attribute": "is_signed", "delta": 28.50}], "from_metadata_only": false}
```

8.2 Fallback Logic

When a book is not in the database, the API automatically fetches metadata from Google Books API and generates a prediction using available features. This enables predictions for any book with an ISBN, even without historical market data.

1. Check if book exists in local database

2. If not found, fetch metadata from Google Books API

3. Convert metadata to BookMetadata object

4. Apply user-selected attributes (signed, first_edition, condition)

5. Call ML estimator with metadata only

6. Return prediction with reduced confidence (70% multiplier)

8.3 Attribute Deltas

The API calculates incremental price deltas for toggleable attributes, enabling interactive UI updates. Users can see how signing, first edition status, or condition changes affect estimated value.

9. Recent Improvements (2025)

9.1 Signed Book Premium Detection

Problem: Original models severely undervalued signed first editions ("Clear and Present Danger problem").

Solution: Collected 138 new signed first edition comps from 10 collectible authors, increasing signed book training data by 60% (226 → 362 listings).

Result: eBay model now correctly predicts signed premium of $51.41 (+214.9%) for Tom Clancy signed first editions. The is_signed feature now has 1.4% importance.

9.2 Amazon FBM Integration

Integrated Amazon Fulfilled-by-Merchant (FBM) pricing data as a new feature category. This provides insights into third-party seller pricing strategies distinct from Amazon's own pricing.

9.3 API Fallback for New Books

Implemented intelligent fallback logic to predict prices for books not in the database. The system now automatically fetches metadata and generates estimates using the unified model, enabling predictions for any ISBN with sufficient metadata.

10. Recommendations for Future Work

10.1 Data Collection

- **Expand signed book coverage:** Continue collecting signed/first edition comps, target 1,000+ listings
- **Temporal data enrichment:** Track price changes over time to capture market trends
- **International markets:** Expand Zvab coverage for European pricing insights
- **Auction results:** Integrate auction house data for high-value collectibles

10.2 Model Improvements

- **Deep learning exploration:** Test neural network architectures for complex patterns
- **Series-aware modeling:** Leverage series data more effectively (book 1 vs book 7)
- **Multi-task learning:** Jointly predict price, sell-through rate, and time-to-sale
- **Uncertainty quantification:** Implement prediction intervals, not just point estimates

10.3 System Architecture

- **Real-time updates:** Implement streaming data pipeline for live market data
- **A/B testing framework:** Test model variants in production with controlled experiments
- **Feature store:** Centralize feature engineering for consistency across models
- **Model monitoring:** Track prediction accuracy and data drift in production

11. Conclusion

The ISBN Lot Optimizer represents a sophisticated application of machine learning to book price prediction. The stacking ensemble architecture effectively combines insights from multiple marketplaces, while the feature engineering pipeline captures both market dynamics and book attributes.

With an eBay model MAPE of 11.9% and R^2 of 0.788, the system provides reliable price estimates for the majority of books. Recent improvements to signed book detection and API fallback logic have significantly expanded the system's capabilities.

The system is production-ready and actively used for book pricing decisions. Continued data collection and model refinement will further improve accuracy, particularly for collectible and rare books where market signals are sparse.

Document Information

Generated: November 09, 2025 at 07:31 PM

System Version: 1.0

Contact: ISBN Lot Optimizer Development Team