Library Late Book Returns Use Case

Milan Stankovic
30/10/2024

Contents

1. Business objective

The library is facing a problem with late book returns (books are considered late if not returned within 28 days of checkout), which affects book availability for other patrons and disrupts book management

1. Conduct root cause analysis to identify factors associated with late returns
2. Build a model to predict the likelihood of a late return of any book at checkout

2. Data analysis

TRAINING DATA CONSISTS OF 4 CSV FILES: LIBRARIES, CHECKOUTS, BOOKS AND CUSTOMERS

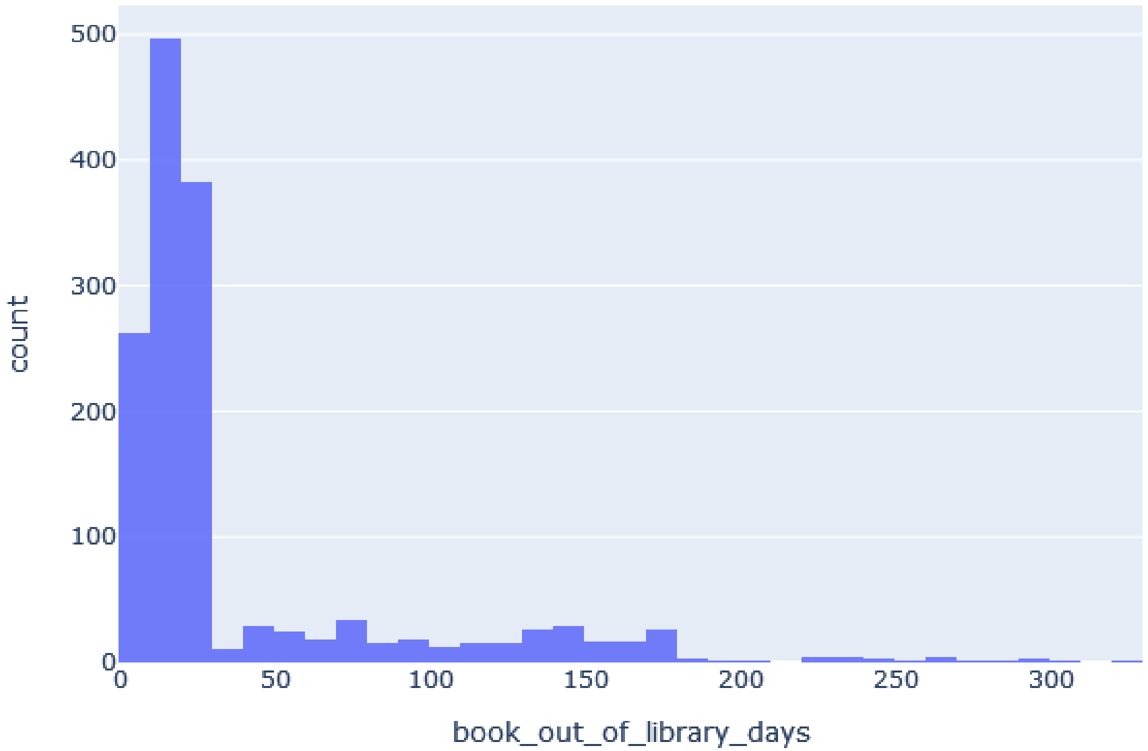
DATA QUALITY ISSUE (MISSING VALUES, DATA INCONSISTENCIES)

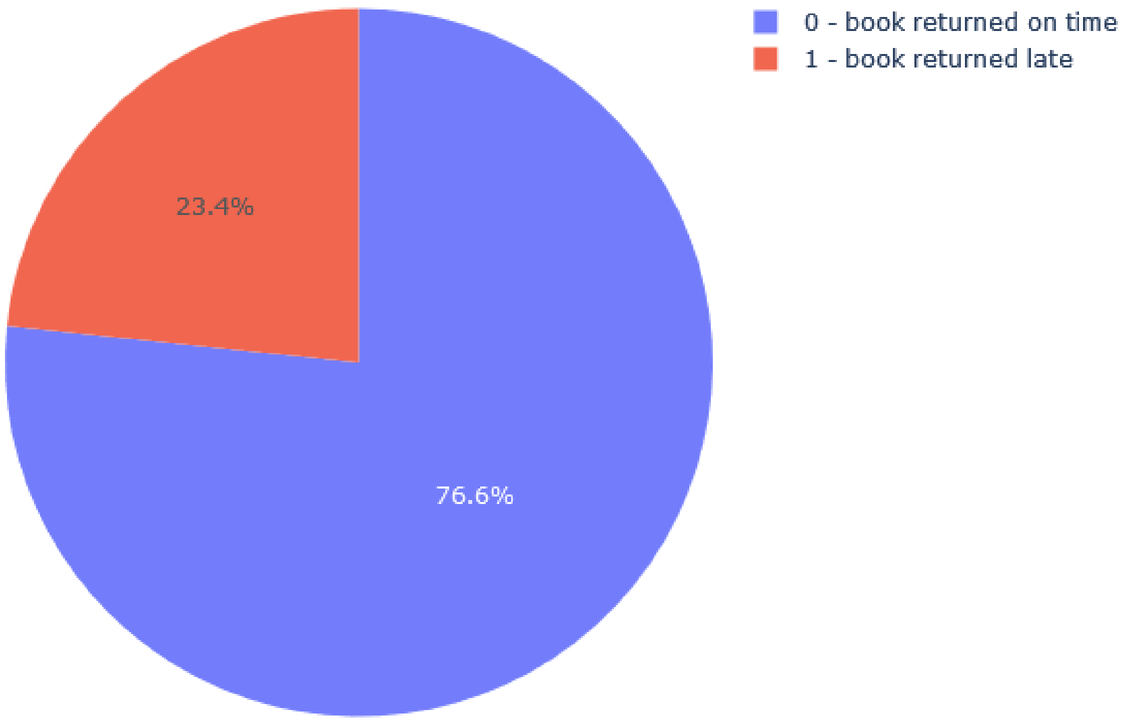
DATA CLEANING AND FEATURE ENGINEERING ARE PERFORMED BEFORE EXPLORATORY DATA ANALYSIS

3 NEW FEATURES ARE ADDED: CUSTOMER AGE, BOOK AGE AND CUSTOMER LIBRARY DISTANCE

APPROXIMATELY 1500 INSTANCES ARE AVAILABLE FOR TRAINING THE MODEL

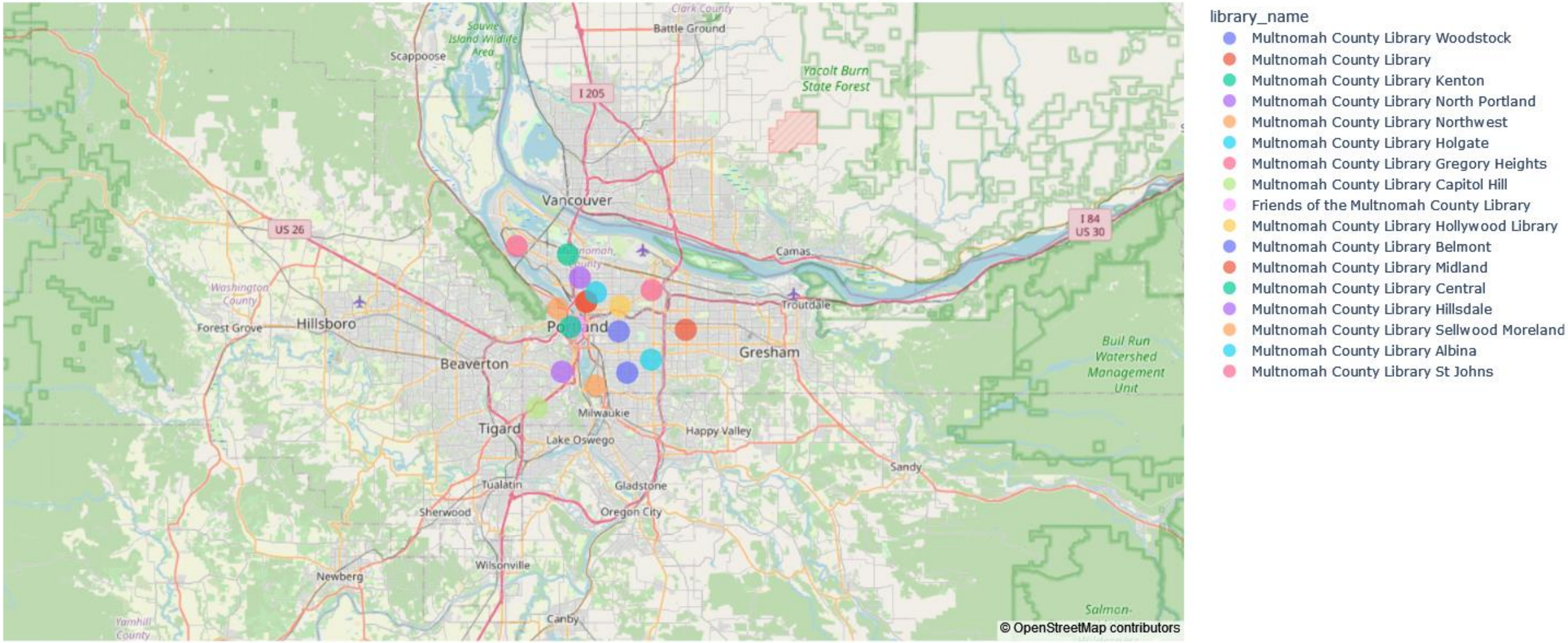
2.1 Target distribution



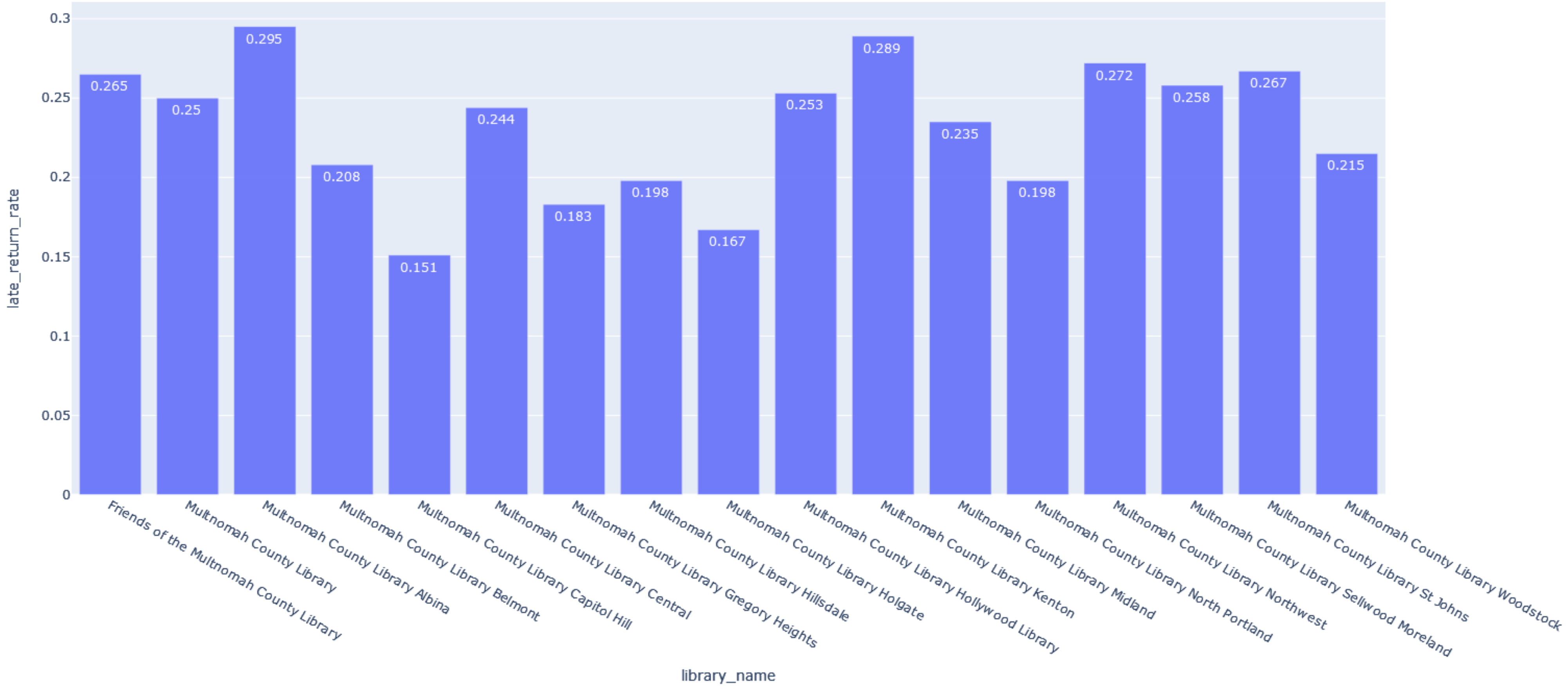


The mean borrowing period is 39.3 days. Half of the books are returned within 19 days of checkout and 75% of the books are returned within 27 days of checkout.

2.2 Libraries



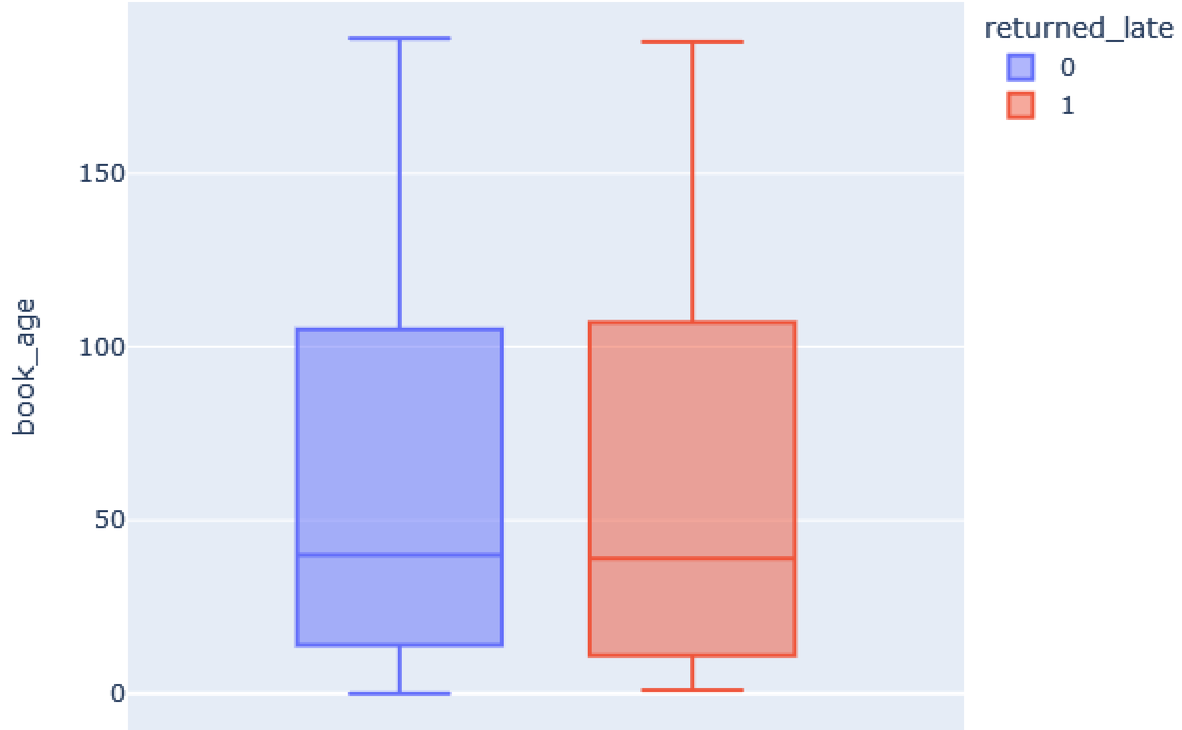
2.2 Libraries

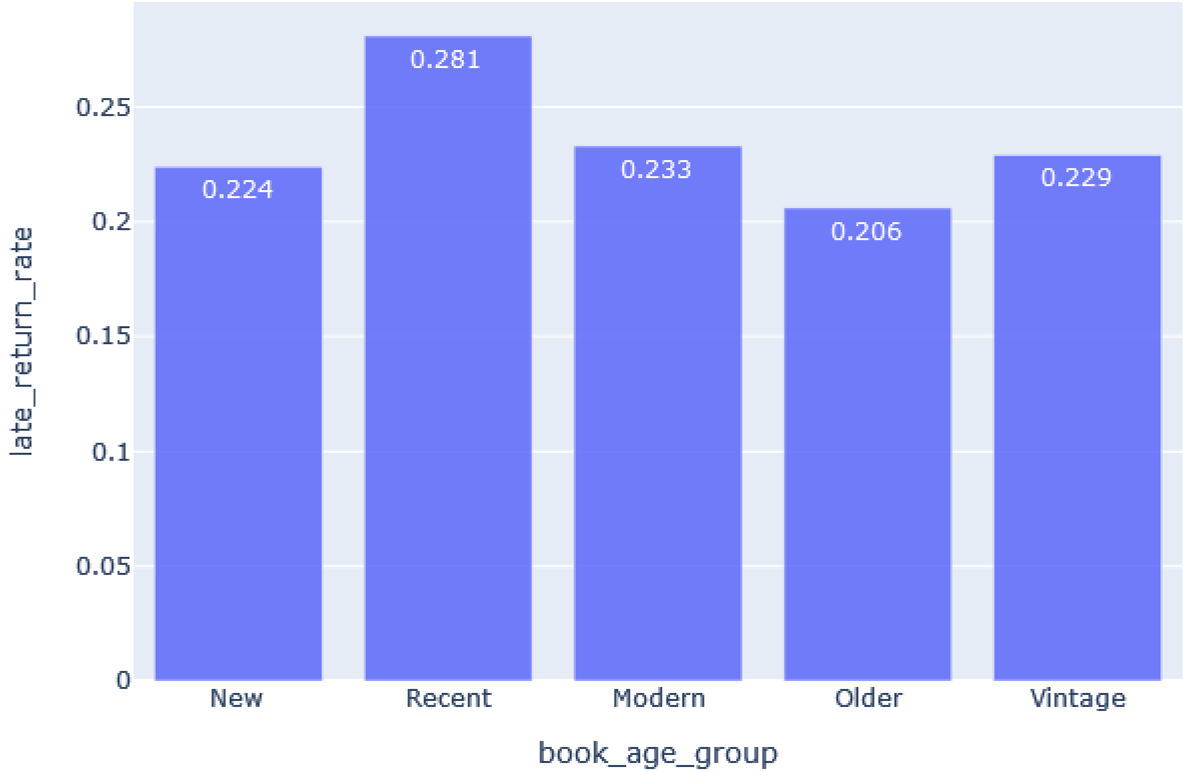


Multnomah County Library Albina has the highest percentage of late returns (29.5%).
Multnomah County Library Capitol Hill has the lowest percentage of the late returns (15.1%).

2.3 Book age

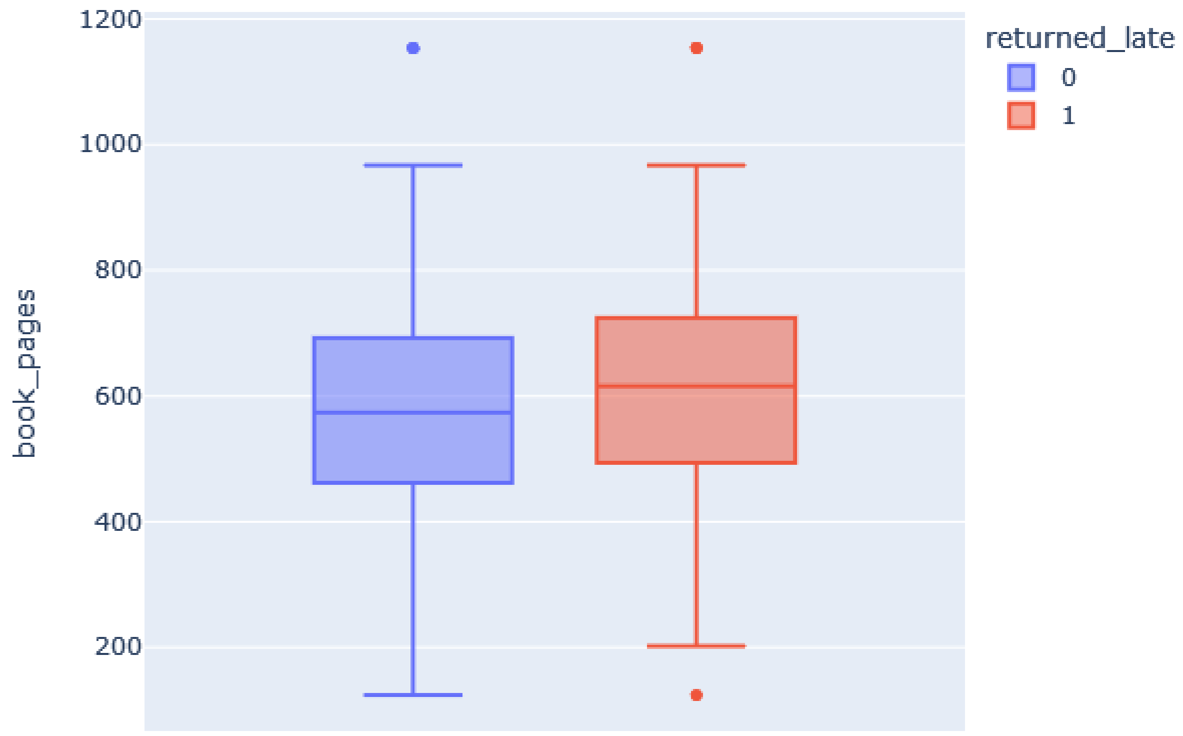
New	0-2 years
Recent	3-10 years
Modern	11-30 years
Older	31-50 years
Vintage	51+ years

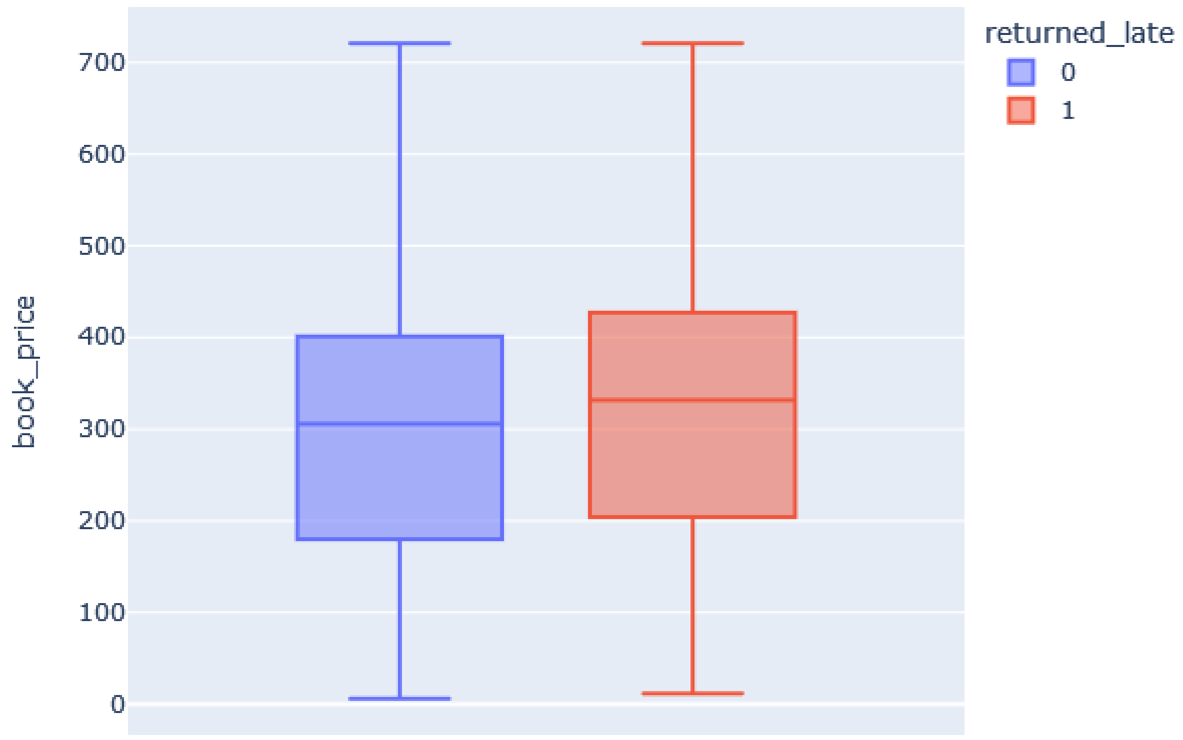




The mean book age is 59.5 years.
Books published within the last 3-10 years have the highest percentage of late returns (28.1%).
Books published within the last 31-50 years have the lowest percentage of late returns (20.6%).

2.4 Book pages and price

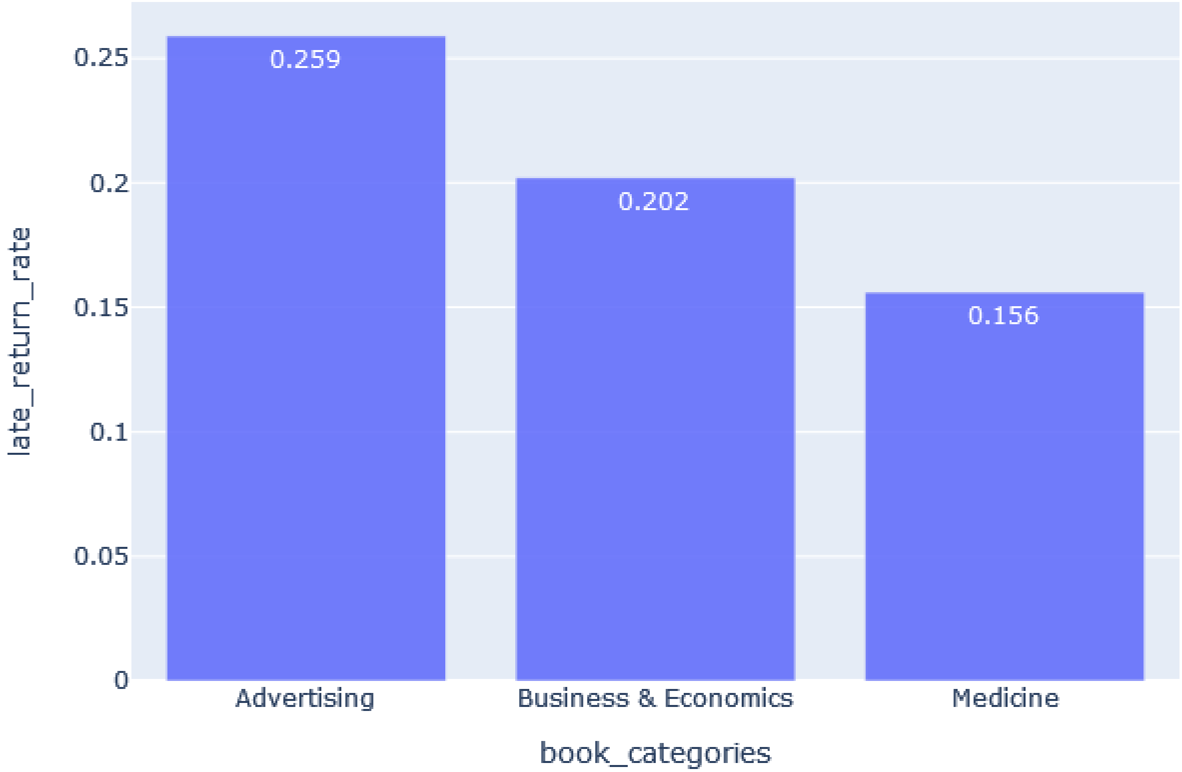




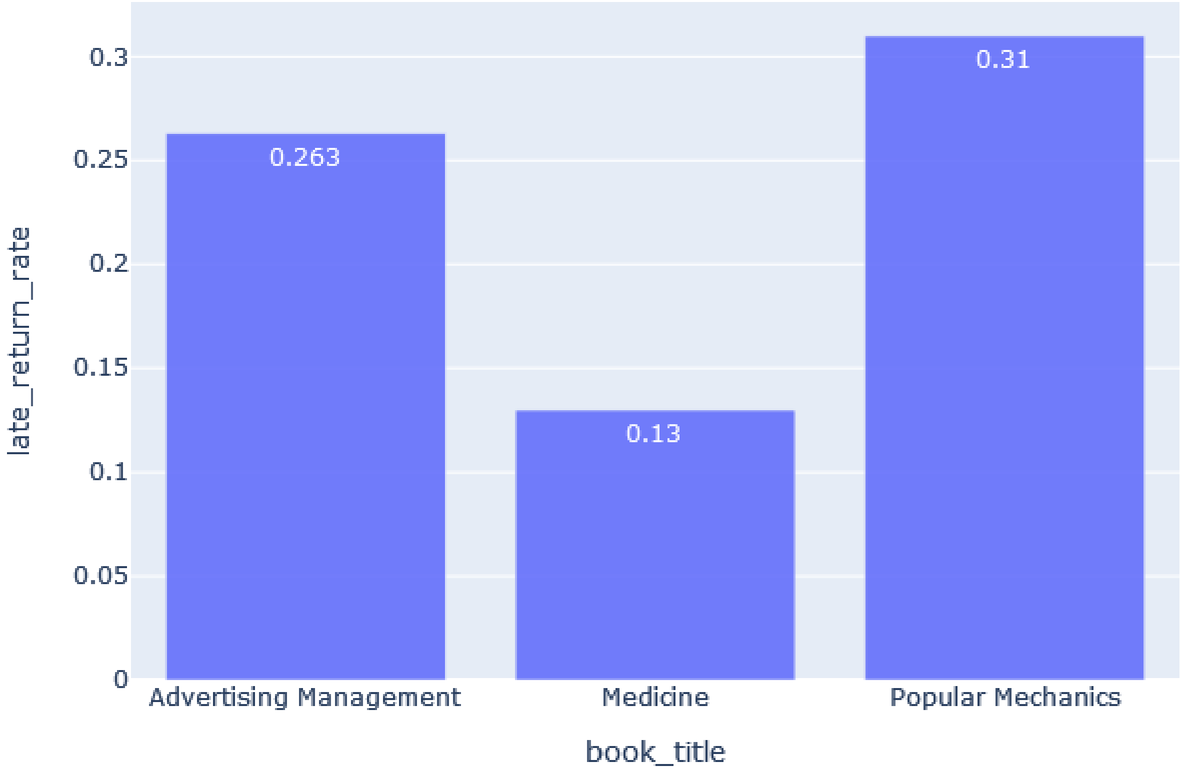
Books with more pages are more likely to be returned late. On average, books returned late have 27 more pages than those returned on time.

Books with higher prices are more likely to be returned late. On average, books returned late are 21$ more expensive than those returned on time.

2.5 Book categories and titles

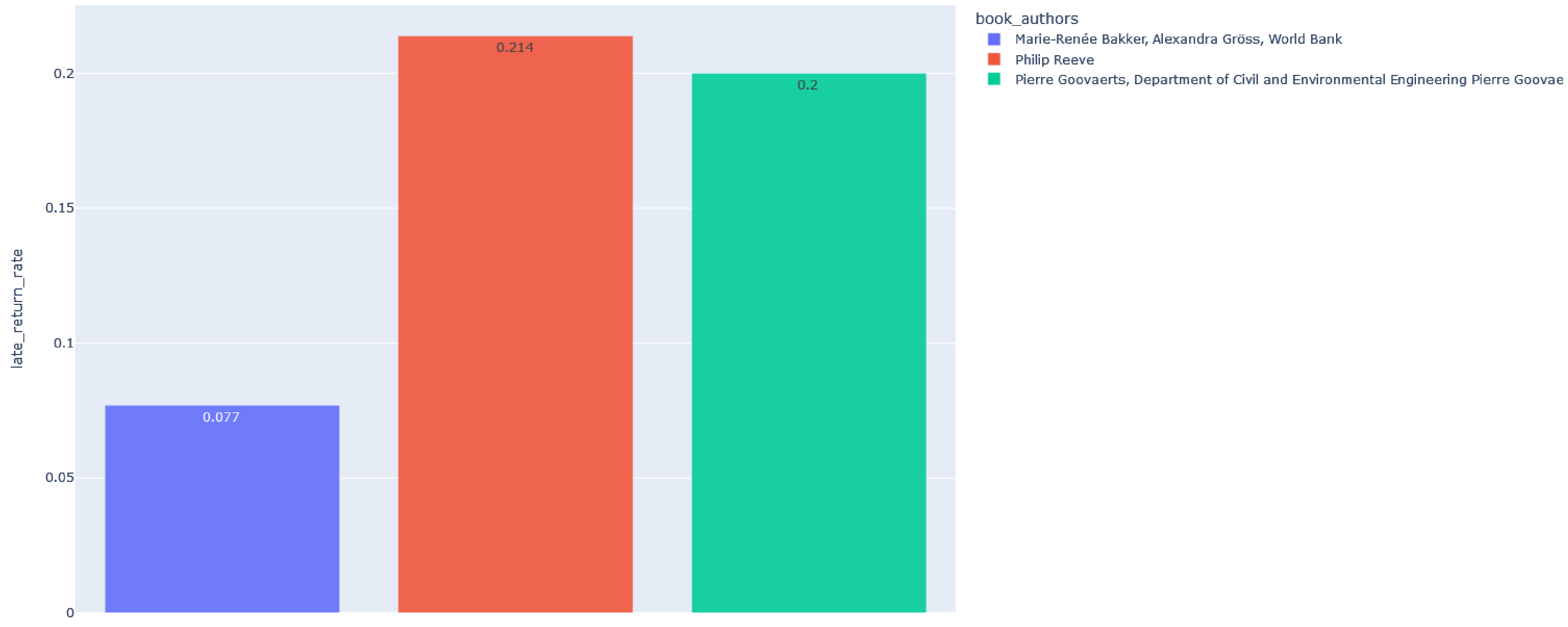


Top 3 most frequent book categories. Medicine has the lowest late return rate (15.6%). Advertising has the highest late return rate (25.9%).



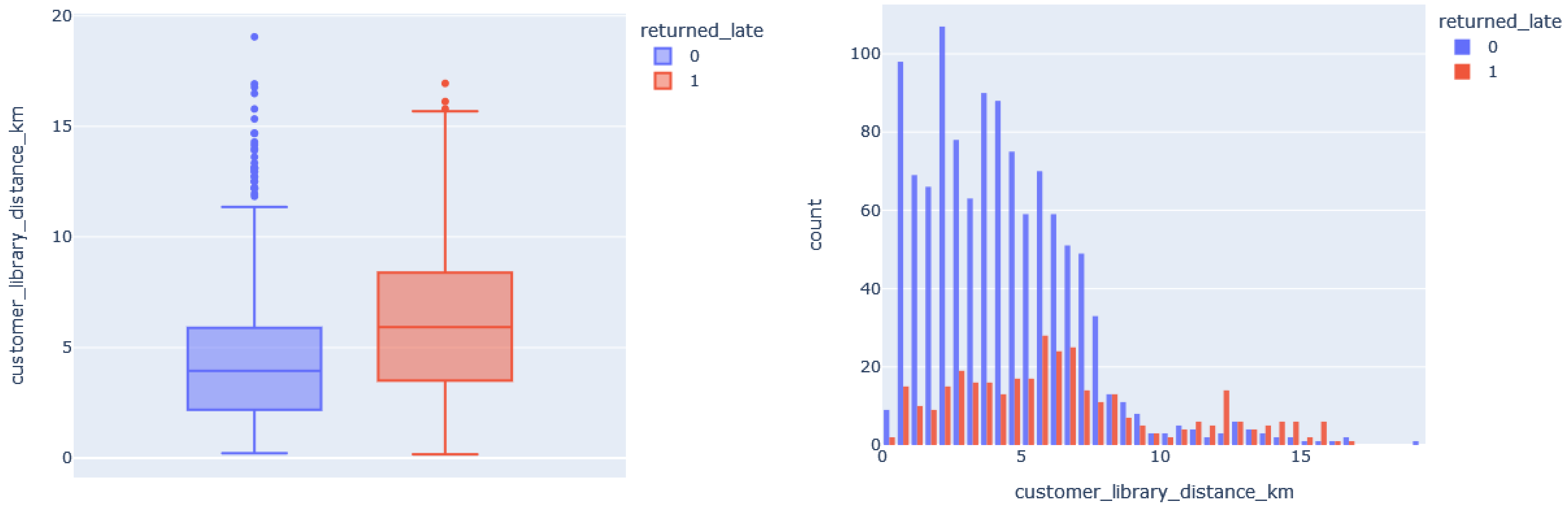
Top 3 most frequent book titles. Medicine has the lowest late return rate (13%). Popular Mechanics has the highest late return rate (31%).

2.6 Book authors



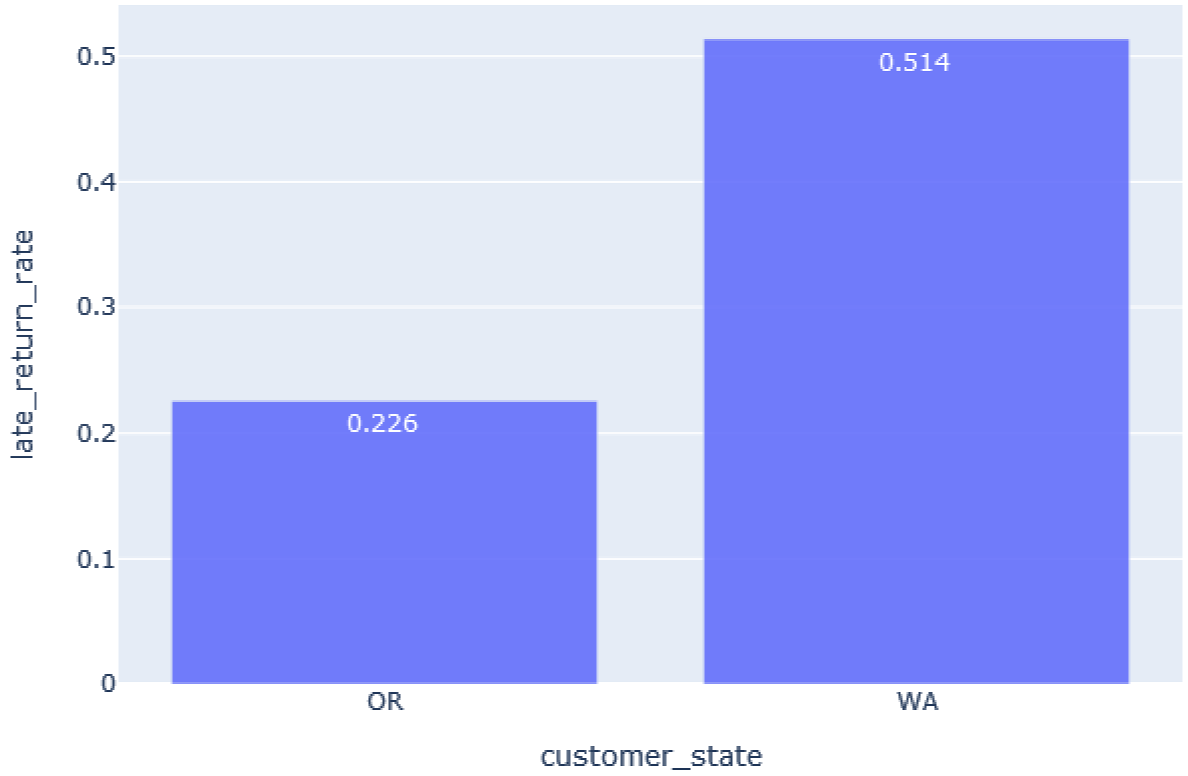
Top 3 most frequent book authors. Books written by Marie-Renée Bakker, Alexandra Gröss, World Bank have the lowest late return rate (7.7%). Books written by Philip Reeve have the highest late return rate (21.4%).

2.7 Distance between customer and library



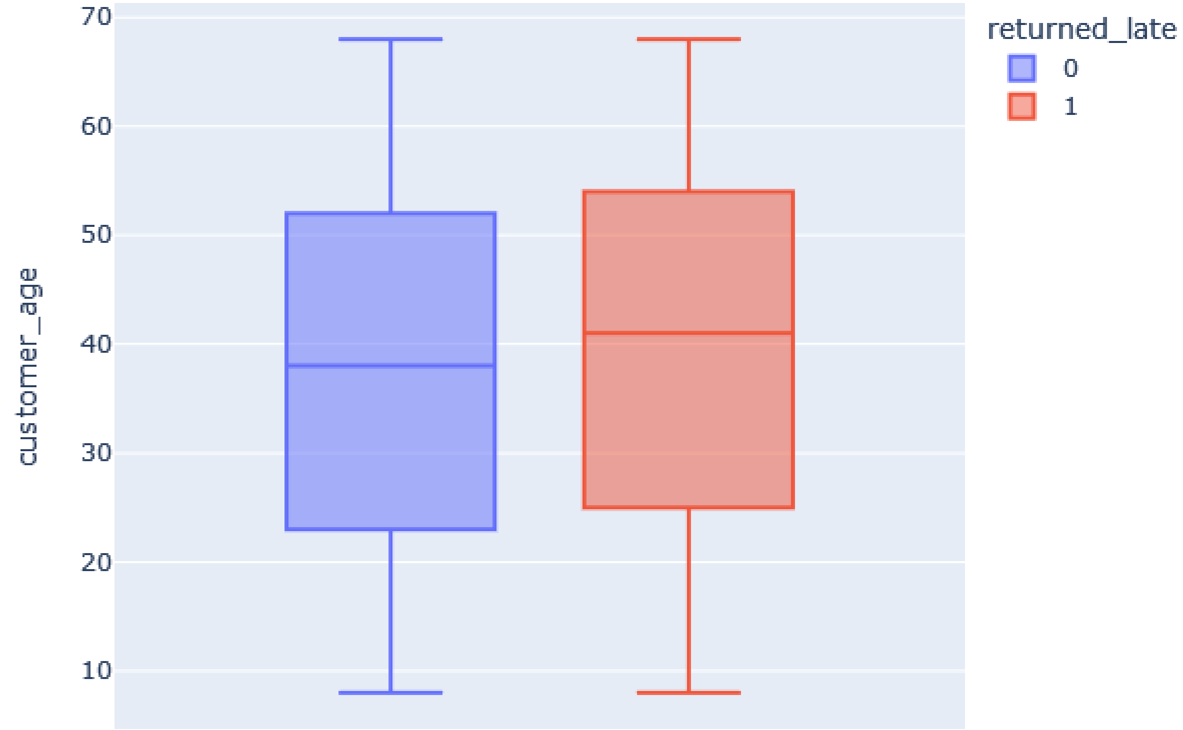
Customers who live farther from the library are more likely to return books late. On average, customers who return books late live 2.24 km farther from the library than those who return books on time.

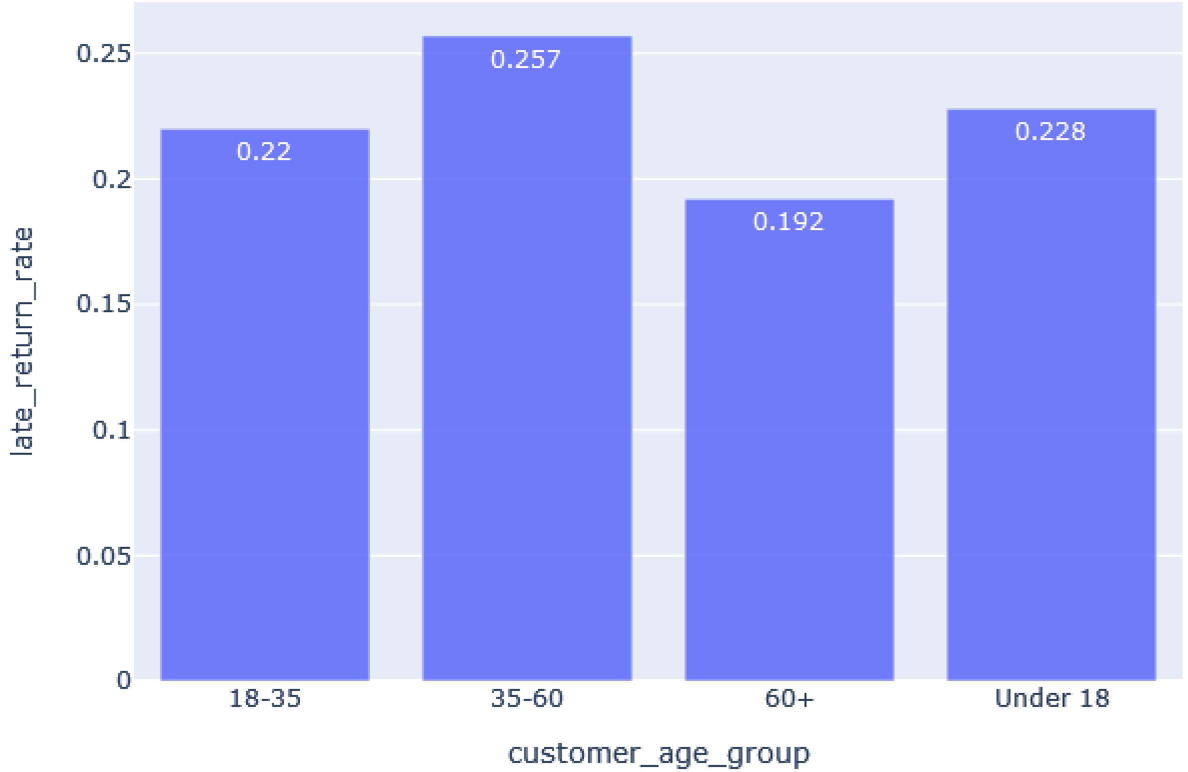
2.8 Customer state



Customers from Washington (2.5%) have a much higher late return rate than those from Oregon, 51.4% compared to 22.6% (which is expected, as all libraries are located in Oregon).

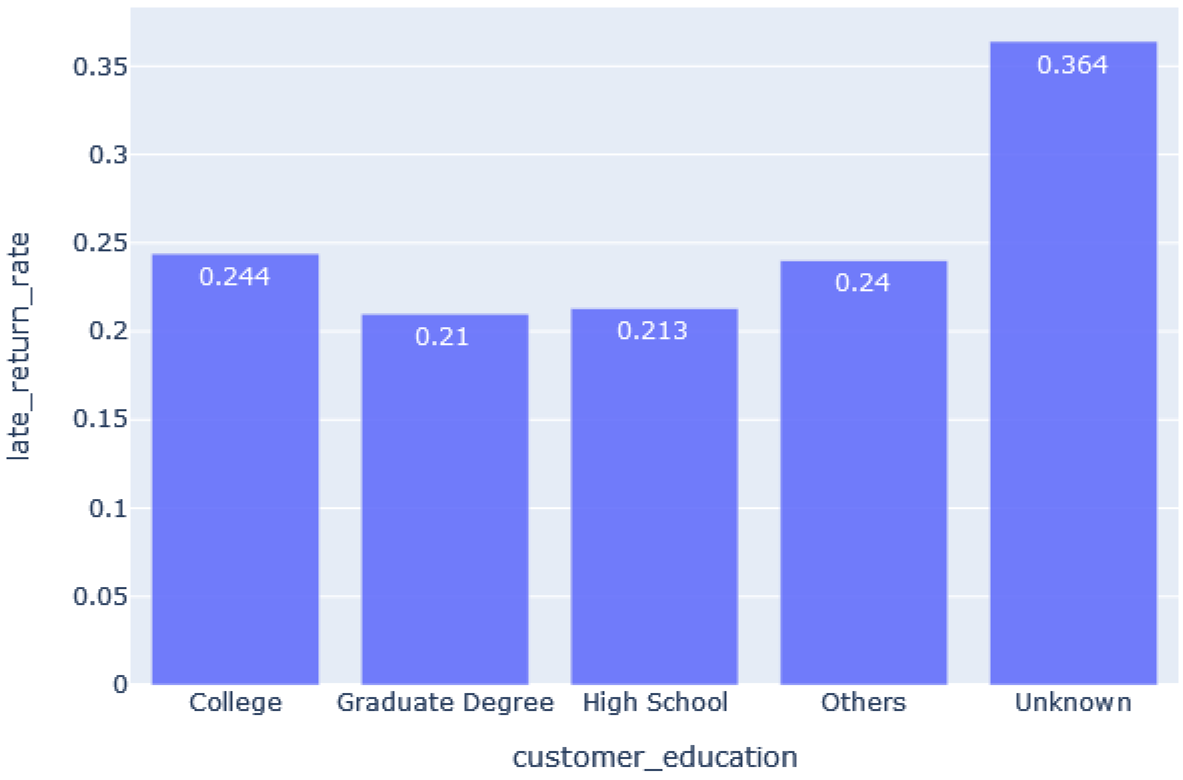
2.9 Customer age





The lowest percentage of late returns is found among older customers (19.2%) - likely retirees, with more time for reading. The highest percentage of late returns is found among 35 to 60-year-olds, who are typically working and have less time for daily reading (25.7%).

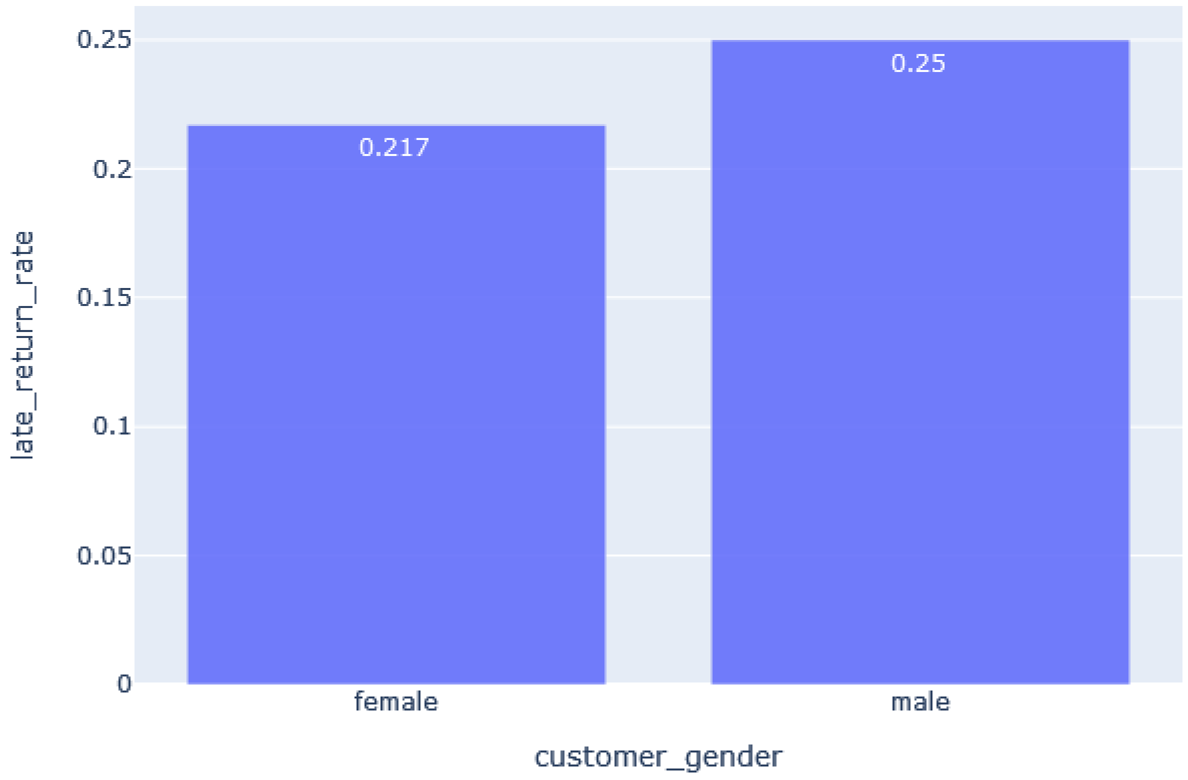
2.10 Customer occupation and education



The highest percentage of late returns is found in the group of customers with tech occupations (26.6%). The lowest percentage of late returns is found in the group of customers with education and health occupations (20.8%).

Customer education is missing in 5.2% of cases. Graduates have the lowest late return rate (21%).

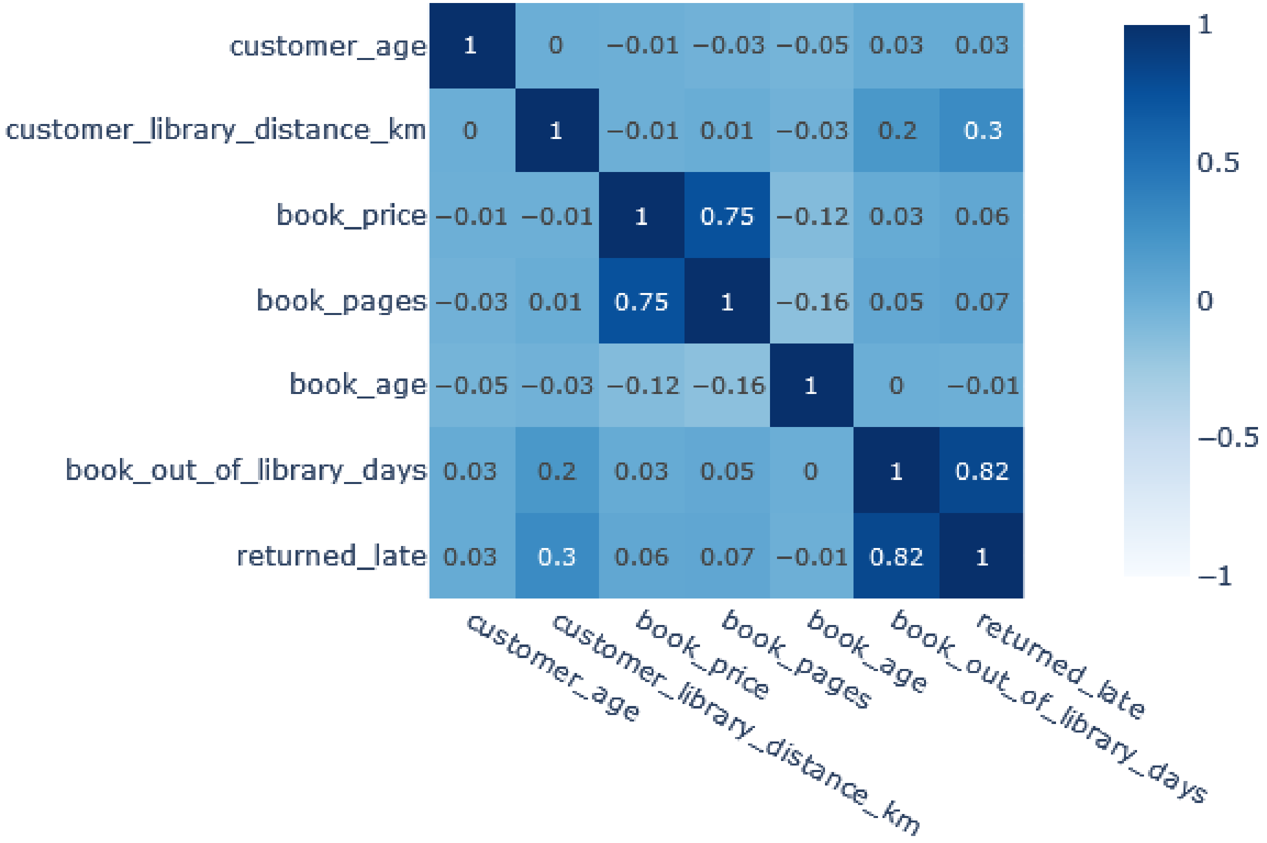
2.11 Customer gender

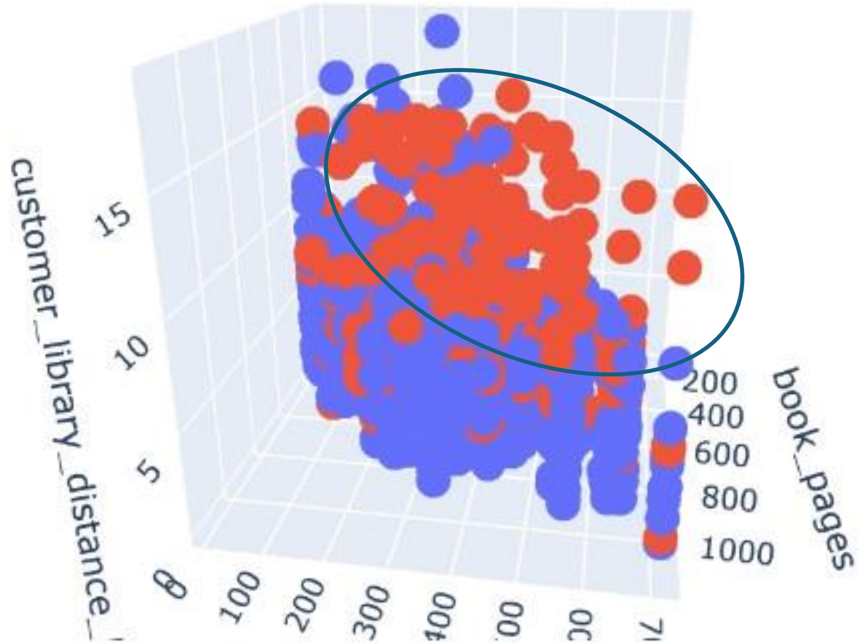


Men return books late more frequently (25% late return rate compared to 21.7% for women).

2.12 Correlation

Features customer_library_distance, book pages and book price are the most correlated with the target variable (returned_late).





2.13 Key observations

75% of the books are returned within 27 days of checkout.

Customers who live farther from the library are more likely to return books late.

Books with more pages are more likely to be returned late as they require a longer time to read.

Books with higher prices are more likely to be returned late.

Customers aged 60 and over tend to return books on time, likely because they are retirees and have more time for reading.

Men return books late more frequently than women.

Medicine books have the lowest late return rates.

3. Model selection

Dataset is split into train and test sets, taking class imbalance into account.

Data preprocessing includes scaling numerical features and encoding categorical features.

Several ML models are tested, considering class imbalance: Logistic Regression, Random Forest, XGBoost and CatBoost. Building a global model – one model for all libraries.

Hyperparameters tuning is performed using grid search and cross validation.

Used model evaluation metrics are ROC AUC, balanced accuracy and weighted f1 score (imbalanced dataset).

3.1. Models overview

	Logistic Regression		Random Forest		XGBoost		CatBoost	
	train	test	train	test	train	test	train	test
ROC AUC	0.69	0.67	0.72	0.7	**0.72**	**0.67**	0.71	0.69
BAL ACC	0.63	0.67	0.65	0.63	**0.65**	**0.64**	0.64	0.65
WEIG F1	0.68	0.7	0.69	0.7	**0.72**	**0.71**	0.7	0.72

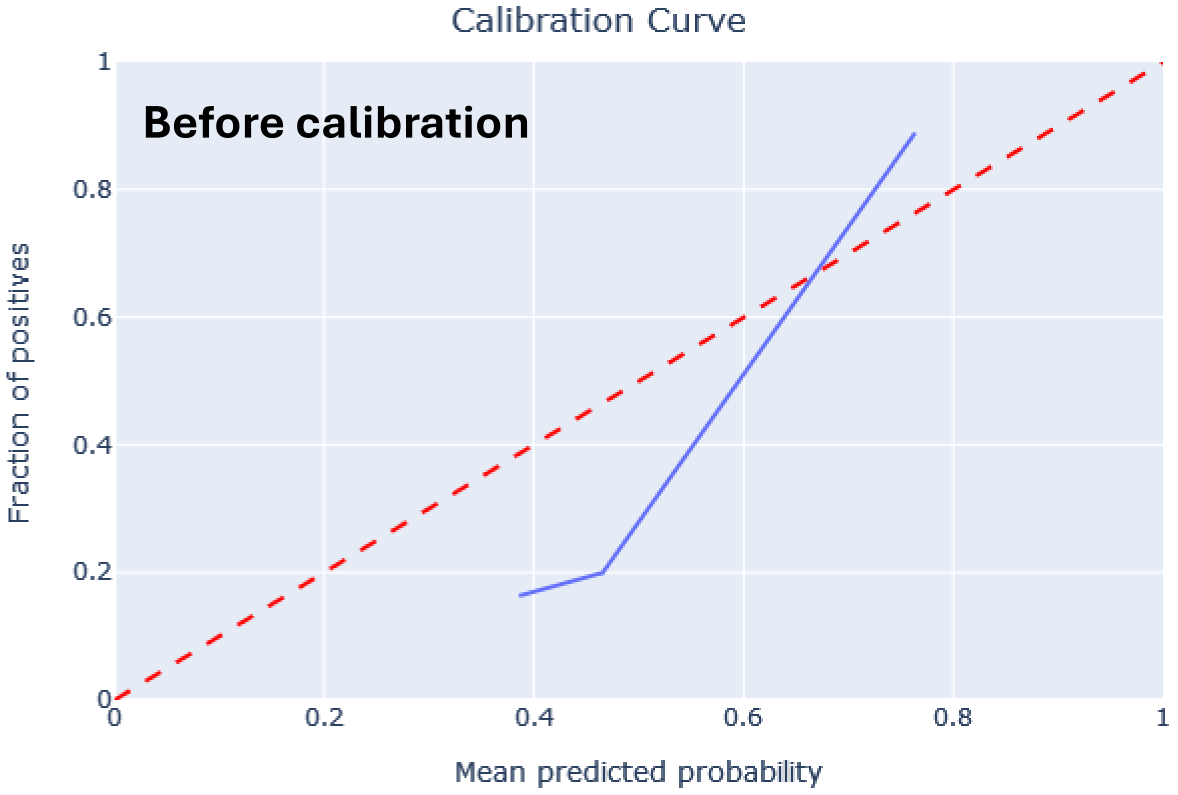
All models show similar performance. XGBoost is selected as go-to model because it operates with the smallest number of features (next slide). Model explainability.

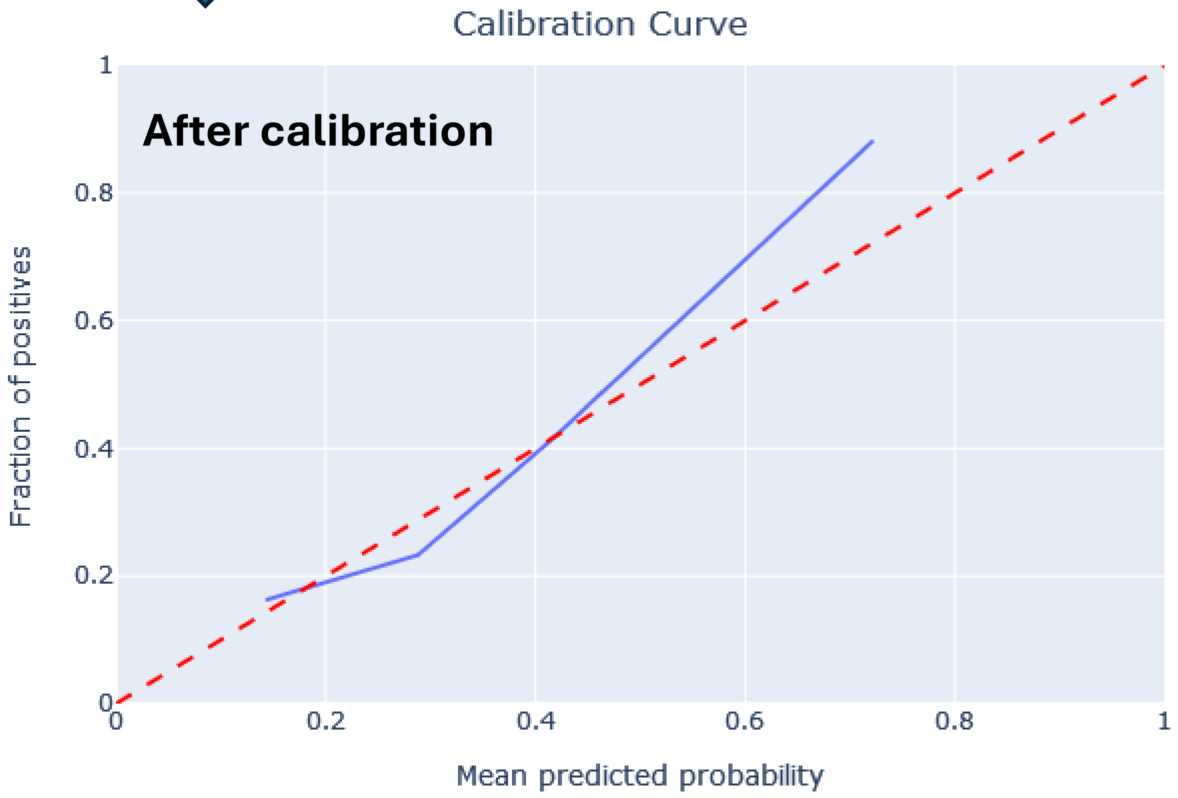
3.2 Final model - XGBoost

	XGBoost trained with top features	
	train	test
ROC AUC	0.72	0.66
BAL ACC	0.65	0.63
WEIG F1	0.71	0.7

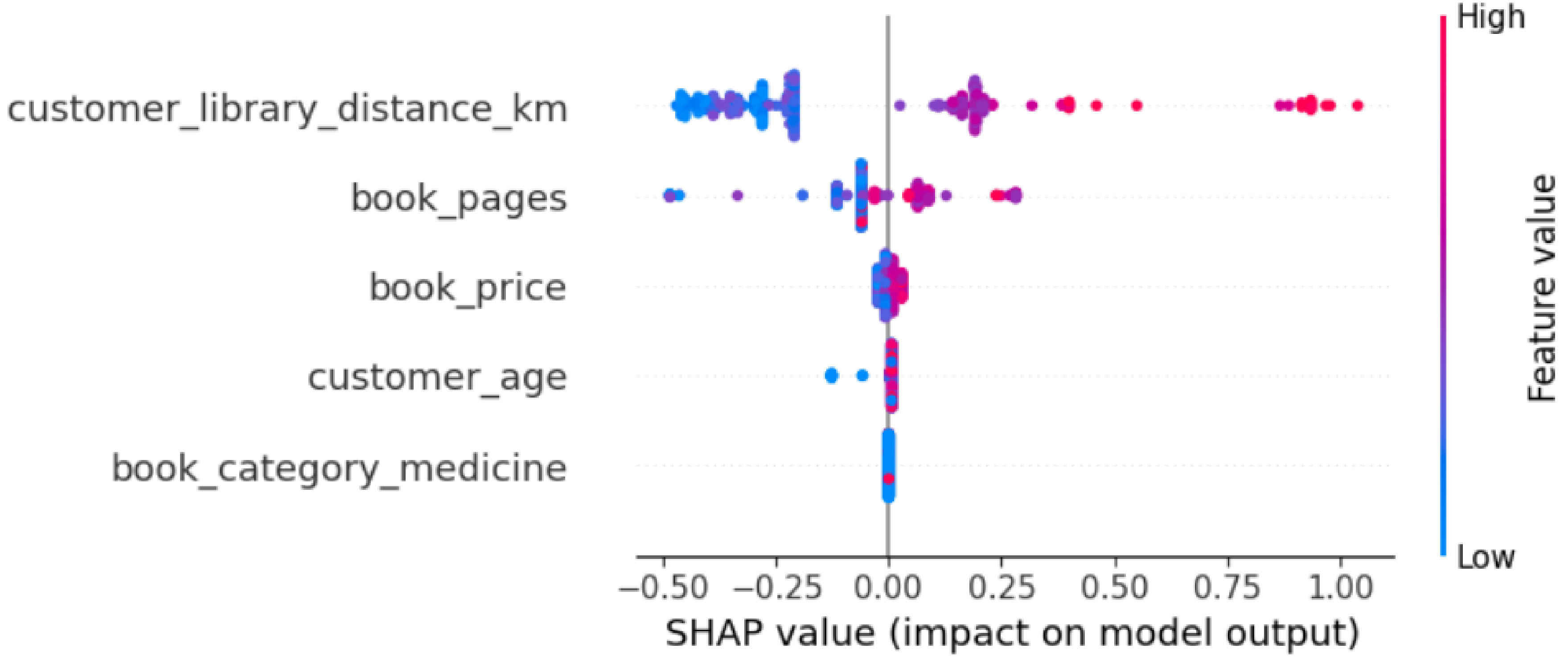
3.3 Model calibration

Calibrated model achieves weighted f1 score of 0.77, which is 7% higher than the uncalibrated model (0.7). Calibration is essential in this case because we are interested in the likelihood of book late returns.





3.4 Global model interpretability (SHAP)



3.5 Local model interpretability (SHAP)

Late return instance

prob = 0.61

On time return instance

prob = 0.26

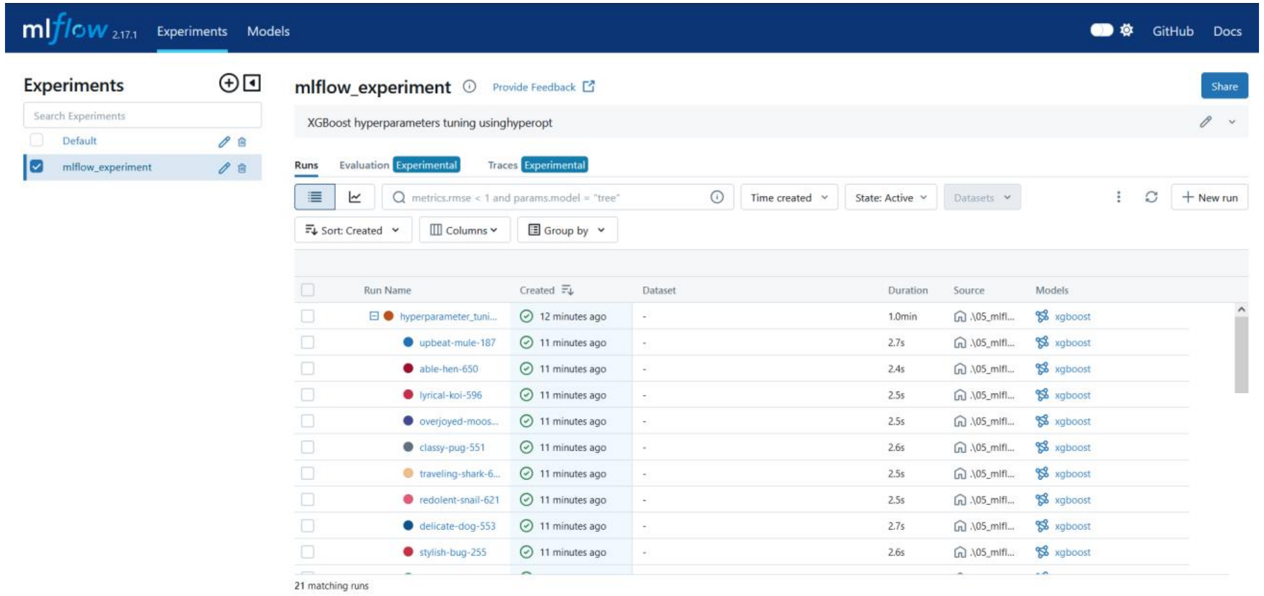
4. Solution

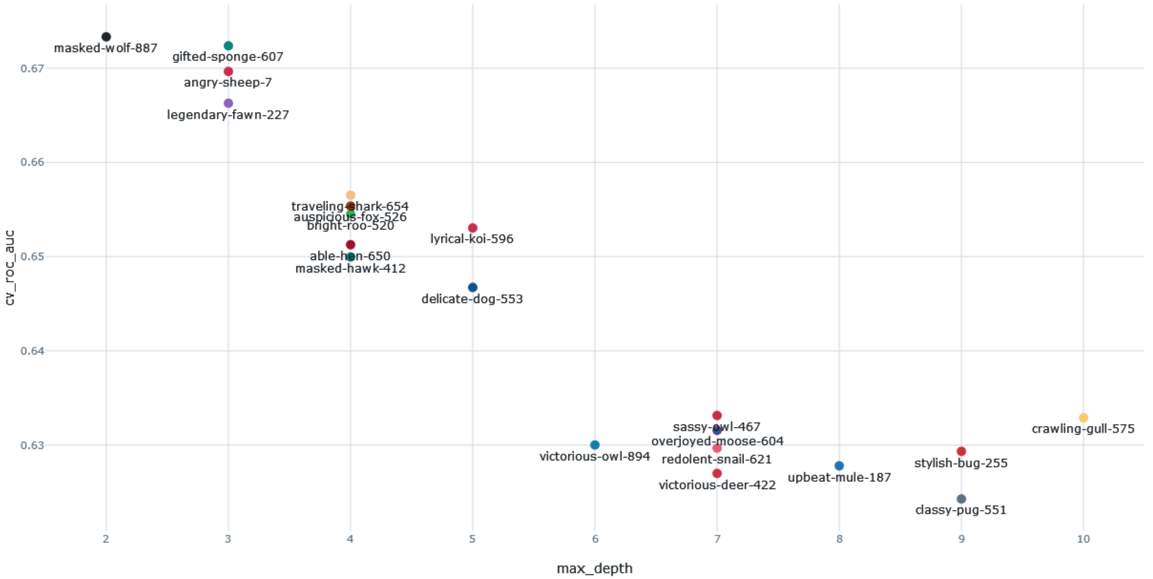
- We identified key factors related to late book returns: distance between the customer and the library, the number of book pages, book price, customer age and books belonging to medicine category.
- To mitigate the risks of late returns, we propose using developed library ML model in day-to-day operations to assess the likelihood of a late return at checkout for any customer. If the model classifies a customer as a high-risk, the library can offer them some rewards for on time book returns, such as discount on the library subscription for the next month or another relevant benefit. Implementing a penalty system is not recommended, as it may negatively affect customer retention.
- Library model is available via API endpoint (appendix) and it can be integrated into the library app. This internal web-based application will provide valuable insights to library staff. It may include a simple dashboard highlighting high-risk customers, real-time prediction functionality, book tracking and a built-in notification system to alert employees for return deadlines.
- Business value: optimization of library operations, reduced late return rate, increased book availability and enhanced customer satisfaction.

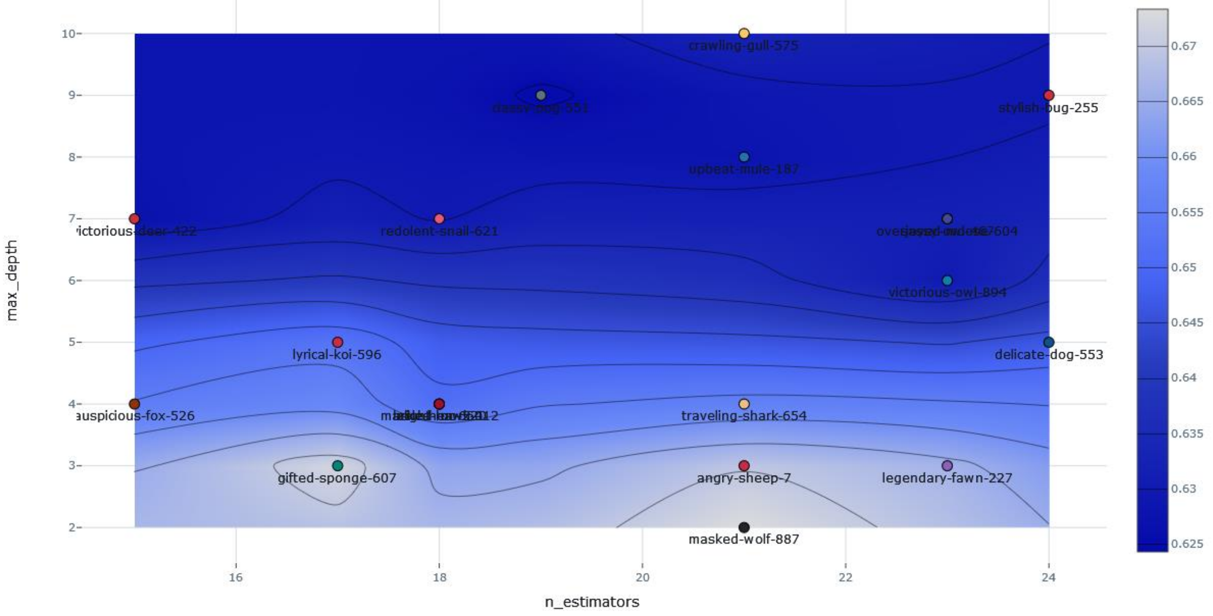
5. Next steps

1. Go/No-go decision. If go, proceed with the following.
2. ML: Implement Airflow for ML workflow orchestration (Train/Predict/Refit DAGs). Set-up a database for storing model predictions (batch prediction).
3. ML: Build a model monitoring system to track production model performance on real-world data.
4. SW: Backend/Frontend development of the library application.
5. QA and app deployment.
6. Cost/effort estimation for the above tasks.
7. Generate the project plan with milestones/timelines and kick-off the work.

Appendix 1 - MLflow

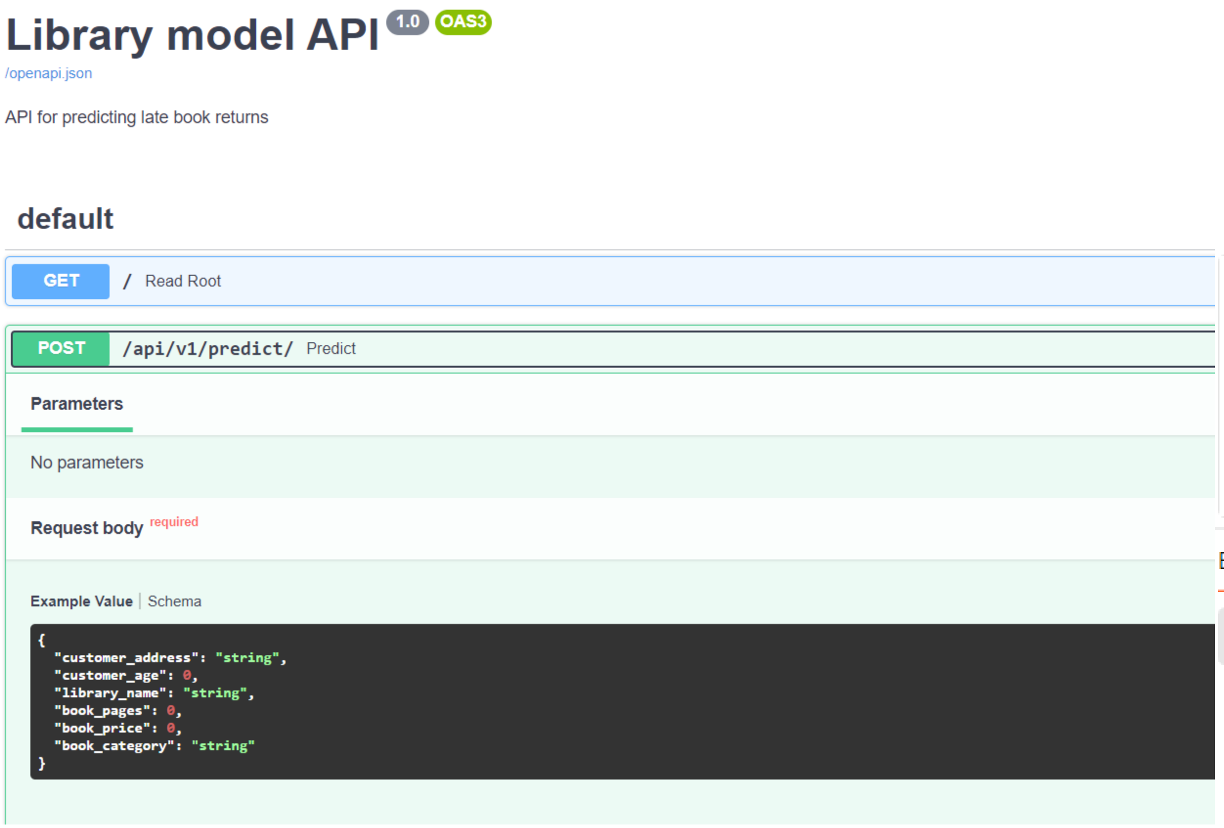




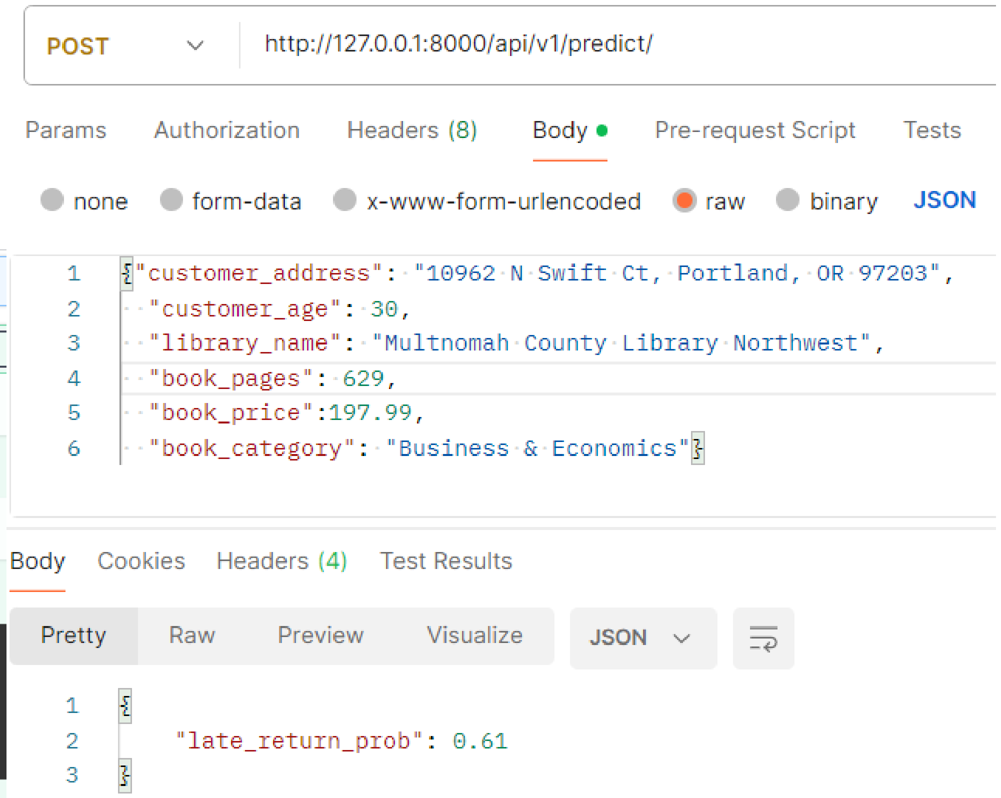


MLflow was used to track XGBoost hyperparameters tuning runs with Hyperopt (Bayesian optimization).

Appendix 2 - API



Late return instance



FastAPI endpoint for real-time predictions.

Thank you!