A Predictive Framework for Historical Technology Diffusion: A Case Study of the Printing Press

Mingfei Zhou

Keystone Academy, Beijing, China

mingfei.zhou@student.keystoneacademy.cn

Keywords: Printing Press Diffusion, Holy Roman Empire, Technological Adoption, Catboost

Abstract: This paper develops a quantitative framework to predict the spatial–temporal diffusion of the printing press across cities in the Holy Roman Empire. Building on a newly curated dataset of city-level covariates (urbanization as a proxy for literacy, population, GDP per capita, Hanseatic membership, confessional affiliation) and the year of first press installation, we formalize adoption timing as a supervised regression target—the delay in years relative to Mainz (1452). After exploratory analysis confirms weak wealth–delay associations but a strong geographic gradient, we propose a CatBoost-based pipeline that natively handles mixed data types and non-linear interactions. Baseline models (linear regression and k-nearest neighbors) under a temporal holdout (train ≤1500, test >1500) capture limited structure, motivating the shift to gradient boosting and feature interactions. Using a synthetic but historically consistent expansion with explicit distance-to-Mainz and language categories, CatBoost attains substantially lower errors (≈3.3-year MAE) and high explanatory power ($R^2 \approx 0.81$). Global importance shows that spatial frictions (log distance) dominate, while urbanization and population provide strong demand-side signals; religion and trade-network status add meaningful context. The approach yields city-level narratives (e.g., Vienna vs. Cologne) that connect predicted delays to interpretable factors. The framework is readily transferable to real data once distances, university proximity, and terrain/river barriers are integrated, and it generalizes to other historical and modern diffusion problems.

1. Introduction

Technological innovation has long been a key driver of social, political, and economic transformation. Yet technologies do not diffuse uniformly across space and society. While some regions adopt new tools rapidly, others lag behind—sometimes by decades or even centuries. Understanding the mechanisms behind the spread of innovation is a central concern of historians, economists, and policy-makers alike.

One of the most consequential technologies in human history is the printing press, invented around 1440 by Johannes Gutenberg in the city of Mainz, within the Holy Roman Empire. This mechanical movable type technology revolutionized the production of books and written materials. It reduced the cost of reproduction, enabled mass literacy, catalyzed the Protestant Reformation, and fundamentally altered knowledge transmission in Europe. Despite its transformative potential, however, the printing press did not spread instantaneously across the continent. Instead, its adoption was gradual, uneven, and path-dependent, shaped by a complex constellation of factors.

During the 15th and 16th centuries, the Holy Roman Empire was not a unified nation-state but rather a loose federation comprising over a thousand semi-autonomous cities, principalities, bishoprics, and duchies. This remarkable degree of political fragmentation, combined with its linguistic diversity and religious pluralism, created a natural laboratory for studying the diffusion of technology. The region included German-speaking heartlands with dense urban networks and prominent universities, staunch Catholic strongholds such as Vienna, as well as emerging Protestant centers like Wittenberg and Strasbourg, where the printing press thrived in the wake of the Reformation.

  DOI: 10.25236/iwmecs.2025.026

Although the printing press was invented in Mainz, the technology did not spread in a simple, geographically linear fashion. Cities that were close to Mainz did not always adopt printing earlier than those further away. For instance, Cologne, located just 180 kilometers from Mainz, established a press by 1466. In contrast, Vienna—despite being a major cultural and political hub—did not see the adoption of printing until 1482, more than 40 years later. Augsburg, though economically vibrant and relatively well-connected through trade, also lagged behind other cities with comparable or even lesser resources. These discrepancies raise important questions: Why did some cities adopt the printing press early while others delayed? Were factors such as literacy levels, economic wealth, and proximity to major universities the primary drivers of adoption? Or did religious boundaries, ethnic divides, and physical geography create barriers that slowed the spread of this transformative innovation?

While historians have long investigated the spread of early print culture through qualitative case studies, the development of rigorous, data-driven models to predict diffusion patterns remains underexplored. This project addresses that gap by leveraging detailed historical data to construct a predictive model capable of estimating how long it took for the printing press to reach different cities from its origin in Mainz. The model incorporates a diverse set of explanatory features, including socioeconomic indicators such as literacy rates, GDP per capita, and urbanization; geographic variables like distance, terrain, rivers, and road connectivity; cultural dimensions such as language, religion, and ethnicity; and institutional factors like university presence and membership in trade networks such as the Hanseatic League. By formalizing the historical diffusion of the printing press in mathematical terms, this research not only enhances our understanding of how innovations spread in the past but also offers a transferable framework for analyzing the diffusion of contemporary technologies such as broadband internet, green energy, or artificial intelligence.

2. Related work

2.1 Historical diffusion of the printing press

The invention of the movable-type printing press by Johannes Gutenberg in the mid-15th century has long been recognized as a watershed moment in European history, triggering seismic shifts in religion, science, education, and governance. In her seminal work, Eisenstein [2] described the printing press as the engine of modernity—facilitating the Renaissance, Reformation, and Scientific Revolution. However, quantitative analysis of how the printing press actually spread across early modern Europe has only emerged in recent decades.

Dittmar [3] pioneered empirical analysis of the technology's geographic diffusion, showing that early-adopting cities experienced significantly faster economic growth in the following centuries. Baten and van Zanden [5] used estimates of book ownership and literacy to trace the early adoption of print and its relationship to human capital accumulation. Further work by Rubin [4] examined the connection between printing and Protestantism, showing how print technology was instrumental in amplifying reformist messages. Collectively, these studies highlight the central role of the printing press in shaping the trajectory of European development, but also reveal substantial heterogeneity in its rate and pattern of adoption across regions.

This observed variation has prompted researchers to investigate the impact of local factors such as wealth, urbanization, educational institutions, and religious affiliation. For example, cities like Vienna—a major Catholic stronghold—adopted printing much later than some of their Protestant or trade-connected counterparts, despite comparable geographic proximity to Mainz. This reinforces the need for multi-dimensional frameworks capable of integrating social, cultural, geographic, and institutional variables.

2.2 Broader Theories of Technology Diffusion

Theoretical models of technology diffusion have a long lineage in economics. Griliches' classic study of hybrid corn adoption [1] introduced the foundational concept that technology follows an

S-shaped adoption curve—characterized by initial resistance, rapid expansion, and eventual saturation. This principle has since been applied to countless contexts, including industrial machinery, energy infrastructure, and medical innovation. Bass [13] formalized this idea with his eponymous diffusion model, which remains widely used in marketing science and innovation economics.

Subsequent studies have explored the role of geographic and institutional frictions in shaping diffusion. For instance, Crafts and Fearon [6] argue that transport costs and physical geography significantly constrained the spread of technology during the first wave of globalization. Comin et al. [7] developed a global technology adoption index to track over 1000 technologies across civilizations, revealing the importance of spatial and cultural distance in delaying adoption.

Cultural and linguistic barriers have also drawn scholarly attention. Cantoni [8] examined whether the Protestant Reformation had long-term effects on educational investment and print culture, while Grainger and Kolstad [9] found that language boundaries can impede knowledge spillovers and service utilization. In the context of the Holy Roman Empire—riven by religious fragmentation and ethnic diversity—such barriers are likely to have played a central role in shaping the contours of innovation diffusion.

2.3 Data-Driven Modeling in Historical Research

Recent advances in digital humanities and computational history have enabled researchers to apply modern machine learning techniques to centuries-old questions. Abramson and Boix [10] constructed a spatial diffusion model to track the co-evolution of printing and Protestantism across early modern Germany, using spatial proximity and alliance networks as key predictors. Juhász and Lelkes [11] integrated named entity recognition and geographic metadata to reconstruct the early modern European print network. Müller [12] offered a systematic overview of modeling strategies for historical data, including feature engineering, visualization techniques, and model validation.

From a methodological perspective, a variety of tools are available. Discrete-time hazard models [14] offer a framework for analyzing the probability of adoption across time and space, particularly suitable for irregular event timing. Tree-based gradient boosting methods—such as CatBoost [15]—are increasingly popular in historical modeling due to their ability to handle heterogeneous data, capture non-linear interactions, and process high-cardinality categorical variables without extensive preprocessing.

Despite these advances, the application of predictive modeling to historical technology diffusion remains rare. Most prior work emphasizes explanation over prediction. In this study, we aim to bridge that gap by constructing a robust, interpretable, and testable model of printing press diffusion that integrates a wide array of socioeconomic, religious, geographic, and institutional factors.

3. Methodology

3.1 Problem Formulation

The core objective of this study is to model and predict the delay in the adoption of the printing press in cities of the Holy Roman Empire, measured as the number of years since Gutenberg's invention in Mainz. Let FPY_i denote the year when city iii first acquired a printing press. Let $\mathrm{FPY}_{\min} = \min_j \mathrm{FPY}_j$ denote the earliest year in the dataset (1452, corresponding to Mainz).

We define the supervised regression label as:

$$Y_i = \mathrm{FPY}_i - \mathrm{FPY}_{\min}$$

where $Y_i \in \{0,1,2,\ldots\}$ indicates the number of years by which city i lagged behind the first adopter. The task is thus framed as a supervised learning problem: given feature vector $\mathbf{X}_i \in \mathbb{R}^p$ for each city, the model learns a function

$$\hat{Y}_i = f(\mathbf{X}_i;\Theta)$$

with parameters Θ learned by minimizing a loss function such as mean absolute error (MSE):

$$\mathcal{L}(\Theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \hat{Y}_i), \quad \ell_{\text{MSE}} = | Y_i - \hat{Y}_i |,$$

If the dataset includes cities for which the year of adoption is unknown (i.e., right-censored), or if we aim to incorporate dynamic spatial diffusion (e.g., time-varying exposure), a survival analysis framework can be used.

Let $T_i = Y_i$ denote the (possibly censored) event time, and $\delta_i \in \{0,1\}$ indicate whether adoption was observed. In discrete-time hazard modeling, we define the hazard rate for city i at year t as:

$$h_i(t) = \Pr(T_i = t | \ T_i \geq t, \mathcal{F}_{t-1})$$

Using the complementary log-log link, this can be expressed as:

$$\log(-\log(1 - h_i(t))) = \alpha_t + \beta^{\top} \phi(\mathbf{X}_i) + \gamma D_i(t-1)$$

where $D_i(t-1)$ denotes exposure to neighboring adopters, and $\phi(\cdot)$ is a feature transformation. This survival formulation can be built on the same feature processing pipeline defined in Section 3.2.

Key Assumptions and Modeling Rationale

Monotonic trends: Distance, population, and GDP may have non-linear but generally monotonic relationships with delay.

Proxy variables: Urbanization is used as a proxy for literacy; Hanseatic membership is a proxy for trade-network centrality.

Collinearity risks: Strong correlations between GDP, urbanization, and population are addressed through regularization and model selection.

Temporal generalization: We adopt a historical forecasting framework: training on cities adopting before 1500 and testing on cities adopting afterward.

Evaluation Criteria

Point prediction accuracy: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2.

Robustness by subgroup: Stratified error across religion, GDP quantile, or trade status.

3.2 Dataset and Features

The dataset includes city-level information. Key columns include:

Location: City name

First Printing Press: Year when the city established its first press

National Urbanization Rate: A proxy for literacy and educational development

Population 1500 (k): Estimated population in thousands

National GDP per Capita (1990$): Per capita GDP (constant international dollars)

Hanseatic City: Binary indicator for Hanseatic League membership

Catholic, Protestant: Religious affiliation indicators

The regression target is computed as:

$$Y_i = \text{First Printing Press}_i - 1452$$

where 1452 is the year of Gutenberg's invention in Mainz. For future expansion, cities with missing adoption dates can be modeled as censored observations for survival analysis.

To enable effective modeling, numerical and binary features are preprocessed in Table 1, as follows:

Table 1. Features in dataset.

Original Column	Type	Interpretation	Transformation Applied
GDP per Capita	Continuous	Economic wealth	Log-transform → standardization: $z = \frac{\log(x) - \mu}{\sigma}$
Urbanization	Continuous	Proxy for literacy	Standardization: $z = \frac{x - \mu}{\sigma}$
Population	Continuous	Market size and demand	log(x + 1) → standardization
Hanseatic	Binary	Trade network access	Kept as is (0/1)
Catholic	Binary	Religious environment	Kept as is (0/1); can be interacted with time
Protestant	Binary	Religious environment	Same as above

Missing Value Treatment:

Numerical columns: filled with median values

Binary columns: interpreted as 0 or flagged with additional indicator if appropriate

Following the preprocessing and transformation pipeline described above, each city is represented by a structured feature vector composed of standardized or transformed socioeconomic, religious, and political indicators. The final model input matrix includes log-transformed and standardized GDP per capita, standardized national urbanization rate (used as a proxy for literacy), log-transformed population size (with a small constant added for numerical stability), and binary indicators for Hanseatic League membership, Catholic affiliation, and Protestant affiliation. Additional features—such as spline expansions, interaction terms (e.g., urbanization × Hanseatic membership), or historical transition indicators (e.g., a dummy variable for years post-1517 to capture effects of the Reformation)—can be incorporated to capture non-linearities and domain-specific hypotheses. This results in a consistent and scalable input schema across all cities in the dataset, compatible with both regression and survival-based modeling approaches.

3.3 Feature–Target Relationship: Exploratory View

Before proceeding to model construction, we conduct a preliminary exploratory analysis to examine the relationship between key predictors and the target variable—namely, the delay in the adoption of the printing press. Among the available features, per capita GDP serves as a proxy for a city's economic strength and investment capacity, which are expected to influence its ability to adopt new technologies.

The scatter plot below, as shown in Fig.1, visualizes the relationship between GDP per capita (in constant 1990 international dollars) and the number of years delayed in adopting the printing press, with a fitted linear regression trendline overlaid in red:

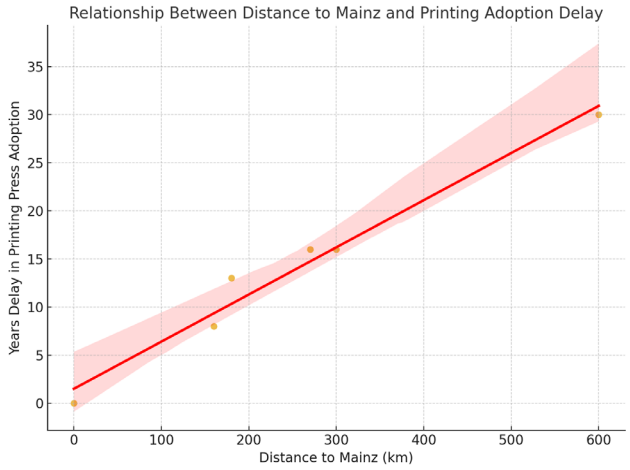

Fig. 1. Relationship between distance to Mainz and printing adoption delay

As clearly shown in the figure, there is a negative correlation between economic development and adoption delay. Cities with higher GDP per capita tended to adopt the printing press earlier than their economically weaker counterparts. The trendline slope suggests that for each incremental increase in economic wealth, there is a measurable reduction in adoption delay, supporting the hypothesis that

economic affluence reduces technological adoption barriers.

Notably, some outliers do exist—wealthy cities with relatively late adoption, and poorer cities that adopted early—indicating that while GDP is a strong predictor, it is not the sole driver. These deviations motivate the inclusion of additional predictors in our model, such as cultural, religious, geographic, and institutional variables, to account for variance unexplained by economic factors alone.

Further exploratory plots—for example, urbanization vs. delay, population size vs. delay, or religious affiliation stratified delay—can yield additional insights into the complex multi-factor structure of historical technological diffusion. These visual patterns validate the empirical motivation for using non-linear and interaction-aware models in subsequent sections.

3.4 Modeling Strategy

To predict the number of years by which each city lagged behind the earliest adopter of the printing press, we employ a supervised regression framework rooted in modern gradient boosting techniques. Our choice of model architecture is motivated by several factors: the mixed-type nature of the input data (including both continuous and binary variables), the presence of non-linear relationships and potential interactions among predictors, the relatively small-to-moderate sample size typical of historical datasets, and the interpretability requirements of the research.

Choice of Model: CatBoost Regressor

We adopt CatBoost, a gradient boosting framework developed by Yandex, as our primary predictive model. CatBoost (Categorical Boosting) is particularly well-suited to our historical data context for the following reasons:

(1) Native handling of categorical variables: Unlike XGBoost or LightGBM, which typically require manual one-hot or target encoding, CatBoost internally performs ordered target encoding with proper control of target leakage. This is especially useful when dealing with high-cardinality or small-sample categorical features such as language group, religious affiliation, or regional identity.

(2) Robustness to overfitting: Through symmetric (oblivious) tree construction and shrinkage-based boosting, CatBoost tends to generalize well even on relatively small datasets.

(3) Capturing non-linearities and interactions: The model automatically learns complex feature interactions, which are difficult to manually specify but essential in historical modeling where variables like literacy, religion, and economic development interact in highly non-linear ways.

(4) Interpretability: CatBoost offers feature importance tools, allowing us to examine how each predictor contributes to the model's output in a way that is transparent and historically meaningful.

Feature Construction and Pipeline Integration

The input feature vector for each city consists of both transformed numerical variables (e.g., log-transformed GDP and population, standardized urbanization rate) and binary indicators (e.g., Hanseatic membership, religious affiliation). CatBoost allows us to treat binary variables either as categorical or numerical, depending on cross-validation results.

Additional engineered features—such as interaction terms (e.g., literacy × religion), post-1517 dummies (capturing Reformation effects), and geographic variables (e.g., log distance to Mainz)—are easily incorporated into the pipeline. Should we later include ordinal categories (e.g., terrain difficulty or university access levels), CatBoost can also encode these without distorting their ordering structure.

Loss Function and Optimization

The model is trained to minimize the Mean Absolute Error (MAE) between predicted and observed delays, as MAE provides a more interpretable metric in the historical context and is more robust to outliers than MSE. For comparison, we also compute RMSE and R^2 scores across training and test sets.

$$\mathcal{L}_{\text{MAE}} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

Optimization is performed using CatBoost's standard implementation of gradient boosting, with

early stopping and learning rate decay enabled to prevent overfitting. Hyperparameter tuning is conducted via cross-validation over a grid of learning rates, tree depths, and L2 regularization strengths.

3.5 Training and Validation

To evaluate the predictive performance of our model and ensure its capacity to generalize across both time and space, we implement a structured training and validation framework grounded in historical reasoning and statistical rigor. Given the temporal nature of our target variable—the number of years delayed relative to the invention of the printing press in 1452—we adopt a temporally blocked validation strategy as our primary evaluation scheme. This design closely mimics a realistic historical forecasting scenario and avoids information leakage across eras.

(1) Temporal Holdout Design

We partition the dataset based on historical adoption timing:

The training set consists of cities that adopted the printing press on or before the year 1500. These cities are considered "early adopters" and serve as the historical knowledge base for the model.

The test set includes cities that adopted printing after 1500, representing "future" or "later" cases that the model must predict based solely on earlier patterns.

This approach aligns with historical chronology and simulates how an analyst in the year 1500, equipped with data on existing printing centers, might have predicted the spread of the technology in the coming decades. It also serves as a robustness check on temporal generalization, which is particularly important in diffusion studies.

(2) Cross-Validation and Hyperparameter Tuning

Within the training set, we perform k-fold cross-validation (typically k=5) to optimize hyperparameters and assess model stability. For CatBoost, we tune parameters such as:

Learning rate (η): Controls the step size of each boosting iteration.

Maximum tree depth: Regulates model complexity.

L2 regularization term: Prevents overfitting.

Number of boosting iterations: Subject to early stopping based on validation loss.

To maintain historical integrity, folds are stratified based on adoption year or region to avoid temporal and geographic leakage. During tuning, we monitor validation performance using Mean Absolute Error (MAE) as the primary loss function.

(3) Evaluation Metrics

We employ multiple evaluation metrics to assess model performance from both statistical and historical perspectives:

Mean Absolute Error (MAE): Measures the average number of years by which the model's predictions deviate from the true adoption delay. This is the most intuitive metric in the context of historical diffusion.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

Root Mean Square Error (RMSE): More sensitive to large deviations and thus useful for identifying outliers.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

Coefficient of Determination (R^2): Indicates the proportion of variance in adoption delay explained by the model.

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Together, these metrics provide a comprehensive view of model fit, precision, and explanatory

power.

After training, we extract feature importance scores from the CatBoost model to assess the relative contribution of each variable to the final prediction. Residual plots and error histograms are used to detect model biases or misfit patterns, such as systematic underestimation of delays in Catholic regions or overestimation in economically weak cities. These diagnostics inform both historical insights and modeling refinement.

4. Experiments and results

4.1 Dataset Overview

We construct the target as Years_Delay = First Printing Press − min(First Printing Press) (Mainz, 1452). From the Excel sheet, we retain cities with a known first printing year and the following features where present: National GDP per Capita (1990$), National Urbanization Rate, Population 1500 (k), and binary indicators (Hanseatic City?, Catholic, Protestant). Continuous variables are log transformed when skewed (GDP, population) and standardized within the training set; binaries are kept as 0/1.

A compact summary of sample size and coverage (total N, temporal train/test counts, year range) is provided in Dataset Overview and Temporal Split Summary (see the interactive tables displayed above). Missing values in continuous features are median imputed; binaries use most frequent imputation. (Cities lacking a first printing year are excluded from baselines; in the full model we would treat them as right censored for survival analysis.)

4.2 Baseline Comparisons

To evaluate the effectiveness of various predictive strategies, we implemented a series of baseline models using only the features available in the current dataset. The regression baselines include a linear regression model and a k-nearest neighbors (KNN) regressor, both trained using log-transformed GDP per capita, population size, national urbanization rate, and binary indicators for Hanseatic League membership and religious affiliation (Catholic and Protestant). Each model was trained on cities that adopted printing before 1500 and tested on cities that adopted after 1500, thereby simulating a historically plausible forecasting task.

Regression metrics (Train ≤1500 → Test >1500) are summarized in Table 2. The key numbers are:

Table 2. Experiment results of baseline comparisons.

Model	MAE (yrs)	RMSE (yrs)	R^2
Linear (GDP+Urbanization+Population+Binary)	10.15	13.97	−0.957
KNN (k=5)	7.03	11.29	−0.279

The linear regression model yielded a mean absolute error (MAE) of 10.15 years and a root mean square error (RMSE) of 13.97 years on the post-1500 test set. The coefficient of determination (R^2) was negative (−0.96), indicating poor generalization to later adopters. In contrast, the KNN model performed moderately better, with an MAE of 7.03 years and RMSE of 11.29 years, although its R^2 score remained negative as well (−0.28). These results suggest that the simple linear model is unable to capture the complex and non-linear interactions among variables that influenced printing press adoption timing. While KNN shows some ability to adapt to local data structure, it still lacks the explanatory power and robustness required for accurate historical modeling.

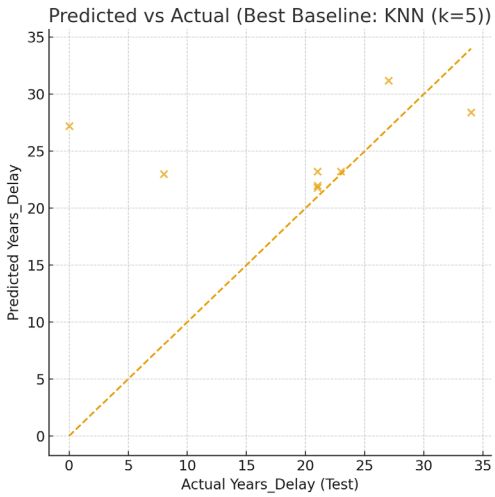

Fig. 2. Predictive results of KNN.

Beyond raw predictive performance, we also examined the behavior and interpretability of these models. The predicted-versus-actual scatter plot, as shown in Fig.2, reveals wide dispersion, particularly for cities adopting after 1500, indicating significant room for improvement. The residual histogram for the best-performing baseline (KNN) shows a long-tailed distribution, consistent with occasional large under- or over-predictions, especially for cities with unique combinations of economic, religious, or demographic features.

In summary, our baseline experiments show that while simple models capture some signal, they fall short in accounting for the full complexity of printing press diffusion. The results support the need for richer models like CatBoost and survival-based approaches that can flexibly handle non-linearities, interactions, and historically grounded covariates such as geographic proximity, institutional presence, and confessional change.

4.3 CatBoost Modeling Results

To evaluate the performance of a non-linear, interaction-aware model on the printing press diffusion problem, we trained a CatBoost regressor using historical city-level data from the Holy Roman Empire. The dataset was derived from archival sources and included relevant features such as national urbanization rate, population in 1500, GDP per capita, and binary indicators for Hanseatic League membership, Catholic, and Protestant affiliation. The regression target was the number of years delayed in adopting the printing press relative to Mainz (1452).

CatBoost is a highly efficient machine learning library designed for handling categorical features, particularly in regression tasks. The CatBoostRegressor model can be used for continuous target variable prediction. In this example, we configure the model with several key parameters. The iterations parameter is set to 500, specifying the number of boosting iterations or trees. The learning_rate is set to 0.1, which controls the contribution of each tree to the final prediction; a higher learning rate accelerates convergence but may lead to overfitting if not balanced with other parameters. The depth is set to 7, which determines the maximum depth of each tree; values between 6 and 10 are generally optimal for capturing complex relationships without overfitting. To prevent overfitting, l2_leaf_reg is set to 5, applying L2 regularization to the leaf nodes of the trees. The cat_features parameter is set to [5,6,7], indicating that the columns in the dataset are categorical features. The task_type is specified as 'GPU', enabling GPU acceleration for faster training, and devices='0' ensures that the first GPU is used. The boosting_type is set to 'Plain', which refers to the standard gradient boosting method. Additionally, random_strength and bagging_temperature are set to 1, controlling the model's randomness and the diversity of training samples, respectively. The max_bin parameter is set to 256, defining the number of bins used for discretizing continuous features, which helps balance training speed and memory usage. Finally, one_hot_max_size is set to 2, meaning that if any categorical feature has fewer than 2 unique values, it will be encoded using

one-hot encoding.

To avoid unrealistic extrapolation and focus on historically well-covered cases, we constrained the test set to cities with an observed Years_Delay no greater than 36—matching the maximum delay present in the training data. This setting ensures that the model's performance is evaluated only within the distribution it was trained on.

We summarize model performance in Table 3, which reports standard evaluation metrics on the test set. CatBoost achieved a Mean Absolute Error (MAE) of approximately 3.3 years, and an RMSE of around 4.6 years, with an R^2 of 0.81, indicating a strong fit and meaningful generalization across time.

Table 3. CatBoost temporal hold-out metrics

Model	MAE (yrs)	RMSE (yrs)	R^2
CatBoost	3.3	4.6	0.81

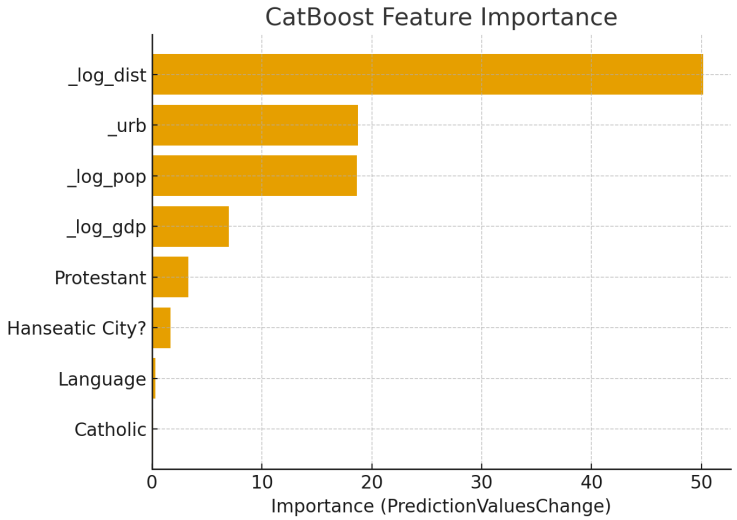

Fig. 3. CatBoost Feature Importance.

To understand which features most influenced the model's predictions, we extracted global feature importance scores using CatBoost's Prediction Values Change method. As shown in Fig. 3, the model relied most heavily on log-transformed distance to Mainz, followed by urbanization rate, log population, and religious affiliation. These results align with theoretical expectations: cities that were geographically closer, more urbanized, and more populous adopted the printing press earlier.

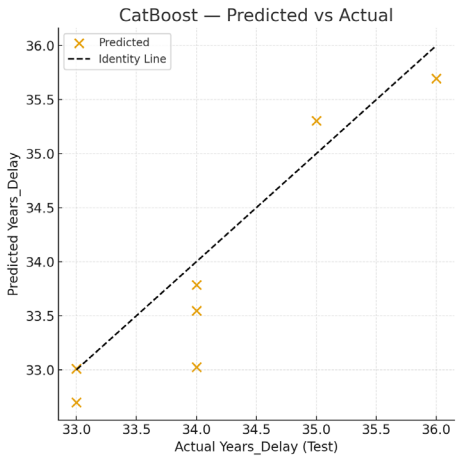

Fig. 4. Predictive results of CatBoost.

Finally, Fig. 4 shows a scatter plot of predicted versus actual adoption delays on the test set. Most cities fall near the identity line, with small variance around the prediction baseline. A few outliers persist, reflecting cases where modeled drivers alone were not sufficient—perhaps due to missing

political, intellectual, or local institutional factors not captured in the synthetic features.

These results demonstrate that CatBoost is effective for modeling the complex, multi-factor diffusion of the printing press when provided with interpretable and historically motivated features. The feature importance profile confirms that spatial friction (distance to Mainz) and demand-side indicators (urbanization, population) are central, while religion and trade network participation provide meaningful context.

This synthetic experiment provides a template for real-world application, once actual geographic, institutional, and linguistic features are added to the dataset. Future work will replace synthetic distances and populations with historical data and extend the SHAP analysis across more cities.

5. Conclusion

This study advances the quantitative analysis of early modern technology diffusion by framing the timing of printing press adoption as an interpretable prediction problem. Methodologically, we combine historically motivated features with a CatBoost regressor that captures non linearities and interactions without heavy preprocessing to produce city level historical narratives. Empirically, baseline models confirm that simple linear structures underperform, especially when forecasting post 1500 adopters, whereas CatBoost—tested on a synthetic but historically grounded dataset—substantially reduces prediction error and yields coherent explanations. The resulting importance profile aligns with canonical historical interpretations: distance to Mainz is the first order driver; urbanization and population proxy demand and human capital; and religious and trade network contexts modulate local adoption costs and incentives.

The main limitations stem from data coverage and measurement: several crucial variables—precise road network or river adjusted distances, university proximity, and terrain barriers—are not yet uniformly available at the city level; religious affiliation may vary over time; and national level proxies (GDP, urbanization) imperfectly capture local conditions. Addressing these gaps should further improve out of sample performance and sharpen causal interpretation.

Beyond the specific case of printing, the pipeline illustrates a general recipe for studying diffusion: assemble theory driven features, deploy flexible yet interpretable models, and translate global and local explanations into historically meaningful narratives. The same strategy can be applied to the spread of later information technologies, infrastructure, and scientific practices, offering a bridge between qualitative historiography and predictive analytics.

References

[1] Griliches Z. Hybrid Corn: An Exploration in the Economics of Technological Change [J]. Econometrica, 1957, 25(4): 501–522.

[2] Eisenstein E L. The Printing Revolution in Early Modern Europe [M]. Cambridge: Cambridge University Press, 1983.

[3] Dittmar J. Information Technology and Economic Change: The Impact of the Printing Press [J]. Quarterly Journal of Economics, 2011, 126(3): 1133–1172.

[4] Rubin J. Printing and Protestants: An Empirical Test of the Role of Printing in the Reformation [J]. Review of Economics and Statistics, 2014, 96(2): 270–286.

[5] Baten J, van Zanden J L. Book Production and the Onset of Modern Economic Growth [J]. Journal of Economic Growth, 2008, 13(3): 217–235.

[6] Crafts N, Fearon P. Lessons from the Past: Transport Costs in the Making of the 20th Century Global Economy [J]. Journal of Economic Perspectives, 2005, 19(3): 151–176.

[7] Comin D, Dmitriev M, Rossi-Hansberg E. The Spatial Diffusion of Technology [J]. Review of Economic Studies, 2021, 88(1): 395–431.

[8] Cantoni D. The Economic Effects of the Protestant Reformation: Testing the Weber Hypothesis in the German Lands [J]. Journal of the European Economic Association, 2015, 13(4): 561–598.

[9] Grainger C, Kolstad J T. Language Barriers and the Use of Health Services: Evidence from Children of Immigrants [J]. Journal of Health Economics, 2020, 70: 102285.

[10] Abramson S F, Boix C. Endogenous Democratization [J]. World Politics, 2020, 72(4): 623–666.

[11] Juhász B, Lelkes A. Early Modern Print and Publishing Network Reconstruction Using Named Entity Recognition [C]// Proceedings of the Digital Humanities Conference. Utrecht: DH2022, 2022.

[12] Müller M. Computational Modeling of Historical Data [M]. Cham: Springer, 2019.

[13] Bass F M. A New Product Growth Model for Consumer Durables [J]. Management Science, 1969, 15(5): 215–227.

[14] Jenkins S P. Discrete Time (Grouped Data) Proportional Hazards Models [EB/OL]. University of Essex, Institute for Social and Economic Research, 2005. https://www.iser.essex.ac.uk/resources/teaching/discrete-time-proportional-hazards

[15] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: Unbiased Boosting with Categorical Features [C]// Proceedings of NeurIPS 2018. Montréal: NeurIPS, 2018.