Секция «16.10 Историческая информатика»

Применение методов машинного обучения для прогнозирования цен книг и рукописей на аукционе «Sotheby's»

Научный руководитель – Афиани Виталий Юрьевич

Павлов Александр Вячеславович

Студент (магистр)

Российский государственный гуманитарный университет, Историко-архивный институт, Москва, Россия

E-mail: alexandrpavlov2001@mail.ru

Авторы работ, посвященных проблематике применения методов машинного обучения для прогнозирования результатов аукционов, фокусируются преимущественно на анализе рынка изобразительного искусства [1, 2, 6, 8, 9]. Несмотря на успешный опыт применения этой технологии в данной области, использованные ими подходы не могут в полной мере быть применимы для анализа рынка книг и рукописей.

Целью исследования является создание модели, которая была бы способна успешно конкурировать с экспертами в прогнозировании цен на книги и рукописи. Для достижения этой цели модель должна эффективно выполнять две задачи: прогнозировать цену лотов и прогнозировать, будет ли лот продан или нет. Гипотеза исследования заключается в том, что эту модель возможно построить, используя информацию о лотах с сайта «Sotheby's».

С помощью веб-скрейпинга и парсинга данных был сформирован датасет, содержащий информацию о соответствующих лотах. Несмотря на наличие структурированных данных, значительная часть информации содержалась в неструктурированных описаниях. Для работы с неструктурированными описаниями был применен TF-подход (Textual Factors) [3]. Данный подход позволяет создавать структурные представления, сохраняя синтаксическую и семантическую информацию. TF-подход балансирует три измерения: вычислительная масштабируемость, лингвистическая сложность и экономическая интерпретируемость.

Для извлечения имен из неструктурированных данных была использована модель *«dslim/bert-large-NER»* [5]. На основе полученной информации были созданы признаки, отражающие частоту упоминания имени в описаниях лотов.

Расчет реальных цен лотов осуществлялся в два этапа: приведение к единой валюте (USD) по среднегодовым обменным курсам года аукциона и корректировка на инфляцию с помощью индекса потребительских цен США на основе данных Федерального бюро статистики труда для приведения к покупательной способности базового года.

Мультиязычная модель *«BAAI/bge-m3»* [4] применялась для создания эмбеддингов названий аукционов, чтобы сохранить важные сведения об особенностях лотов. Информация о месте проведения аукциона была закодирована с помощью метода One-Hot Encoding.

Был проведен сравнительный анализ следующих алгоритмов, использованных для решения указанных задач:

1. Регрессия: линейные модели (Ridge, Lasso) и ансамблевые методы (RandomForest, XGBoost, CatBoost, LightGBM);

2. Классификация: логистическая регрессия, RandomForest, XGBoost.

Оптимизация гиперпараметров осуществлялась с помощью библиотеки *«Optuna»* [10]. Для оценки качества прогнозов использовались следующие метрики:

1. Регрессия: RMSE, MAE, R2, Adjusted R2, sMAPE;

2. Классификация: Accuracy, Precision, Recall, F1, ROC-AUC.

Для интерпретации прогнозов модели использовался метод SHAP (SHapley Additive exPlanations) [7]. Результаты исследования продемонстрировали сложности, связанные с применением TF-подхода для анализа неструктурированных данных в сочетании с табличными признаками.

Источники и литература

1) Aubry M. et. al. Machines and Masterpieces: Predicting Prices in the Art Auction Market // HEC Research Papers Series. 2019. URL: https://ebslgwp.hhs.se/heccah/abs/heccah1332.htm (дата обращения: 02.03.2026).
2) Borisov M. et. al. The influence of color on prices of abstract paintings // arXiv preprint arXiv:2206.04013. 2022. URL: https://arxiv.org/abs/2206.04013 (дата обращения: 02.03.2026).
3) Cong L. W. et. al. Textual Factors: A Scalable, Interpretable, and Data-Driven Approach to Analyzing Unstructured Information // NBER Working Paper. 2024. №. W33168. URL: https://www.nber.org/papers/w33168 (дата обращения: 02.03.2026).
4) HuggingFace. BAAI/bge-m3: https://huggingface.co/BAAI/bge-m3.
5) HuggingFace. dslim/bert-large-NER: https://huggingface.co/dslim/bert-large-NER.
6) Lucińska A. The regression model of the art market in Poland // Argumenta Oeconomica. 2025. Vol. 55. № 2. URL: https://journals.ue.wroc.pl/aoe/article/view/1119 (дата обращения: 02.03.2026).
7) Lundberg S., Lee S. A Unified Approach to Interpreting Model Predictions // arXiv preprint arXiv:1705.07874. 2017. URL: https://arxiv.org/abs/1705.07874 (дата обращения: 02.03.2026).
8) Mauer P., Paszkiel S. Tabular Data Models for Predicting Art Auction Results // Applied Science. 2024. Vol. 14. Issue 23. URL: https://www.mdpi.com/2076-3417/14/23/11006 (дата обращения: 02.03.2026).
9) Mei J. et. al. Deep Learning for Art Market Valuation // arXiv preprint arXiv:2512.23078. 2025. URL: https://arxiv.org/abs/2512.23078 (дата обращения: 02.03.2026).
10) Optuna: https://optuna.org/.