Retail Sales Prediction Using Machine Learning:

A Comprehensive Analysis

Abstract

Retail sales forecasting is essential for inventory optimization, customer segmentation, and operational planning. This project develops an end-to-end machine learning pipeline to predict transaction-level retail sales using a real-world dataset of more than 50,000 records collected from shopping malls in Istanbul between 2021 and 2023. The workflow includes rigorous data preprocessing, advanced feature engineering, and dimensionality reduction using PCA to address high-dimensional one-hot encoded features. Two ensemble learning models—Random Forest and XGBoost—were trained and compared for performance. XGBoost emerged as the best-performing model, achieving highly accurate predictions (RMSE ≈ 37.79 TL, MAE ≈ 3.18 TL). Model interpretability was enhanced through SHAP, which identified key contributors such as price, quantity, customer spending behavior, and product category. The resulting system provides accurate, interpretable, and actionable insights for data-driven retail management.

1. Introduction

The retail industry functions in a fiercely competitive environment that is highly dynamic, in which proper forecasting of sales is critical to the industry in optimizing inventory, facilitating supply chains, and maximizing the revenue. As transactional data continues to increase in different channels, sophisticated data-based knowledge has become the pillar of sound business strategy. This project proposes an end-to-end machine learning system at the capstone level that is aimed to forecast the amount of transaction-level retail sales. It is analyzed based on the Customer Shopping Data (2021-2023), which is a real-life dataset consisting of more than 50,000 transactions in one of ten large shopping malls in Istanbul, Turkey.

The ultimate goal of the study is to come up with a predictive system using demographic, behavioral, time, and transaction attributes to predict the total value in a retail sale. This project is technically deep because it implemented multiple sophisticated machine learning techniques, such as: rigorous data preprocessing, extensive feature engineering, dimensionality reduction with Principal Component Analysis (PCA), a comparison of the robust ensemble models (Random Forest and XGBoost), the systematic hyperparameter optimization, and model explanation with SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). The system that is produced by fulfilling the requirements of complexity and technical integration is capable of providing insights that are accurate in their predictive nature, and at the same time, is able to provide insights that are actionable and interpretable by strategic retail management.

2. Problem Statement and Objectives

The retailers always want accurate predictions of their value of transactions in order to increase their efficiency in operations, better inventory control and higher customer satisfaction achieved by offering specialized promotions. The fact that sales forecasts are inaccurately calculated results in mismanagement of inventory that can result in either lost sales (stock-outs) or higher holding costs (overstocking) (Ferreira et al., 2019). The following challenge is addressed in this research:

Do the machine learning models effectively estimate the total sales value of a retail transaction based on a varying set of customer, product, and time related attributes?

The models that have been developed here have two main purposes:

1. Predictive Insight: To give granular and transaction-level sales forecasts to support real-time operations decisions.
2. Prescriptive Insight: To establish and calculate the precise characteristics, including the price of the product, customer buying behavior and shopping place, which have the most impact on the sales outcome.

An effective resolution to this issue will allow retailers to better predict daily and monthly sales patterns, classify and focus on high-value segments of customers and optimize product placement and promotional strategies in different areas of the shopping malls (Bertsimas and Kallus, 2020).

3. System Requirements and Dataset Description

3.1 Technical System Requirements

The machine learning system was used in a basic Python setting (Python 3.8 or above), which is reproducible and modular. Some of the utilised libraries and tools are: pandas to handle data and scikit-learn to preprocess and manage pipeline, XGBoost and RandomForestRegressor to model and SHAP to explain the model. It required a powerful Pipeline Architecture that would wrap all preprocessing, feature engineering, dimensionality reduction, and modeling operations as a single object that would be easily duplicated. The process of hyperparameter optimization was also done effectively with the help of RandomizedSearchCV (njobs=-1) which also has the ability to run in parallel.

3.2 Dataset Description

The data will be gathered between 2021 and 2023 and will include the shopping data of ten different shopping malls in Istanbul. The dataset has more than 50,000 records of individual transactions, meeting the need of large-scale and real-world data analytics.

The most important attributes, which can be used as the background of the predictive features, are:

- Invoice and Customer Identifiers: invoiceno and customerid.
- Customer Demographics: gender, age.
- Product and Transactions: type, quantity, price (in Turkish Liras/TL).
- Contextual Attributes: shoppingmall, paymentmethod and invoicedate.

Such a comprehensive set of features makes it possible to build a multi-layered predictive model that takes into account temporal, behavioural, and categorical variables and goes beyond the simple associations between prices and quantities (Kuhn and Johnson, 2019).

Sample Dataset

	invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	shopping_mall
0	I138884	C241288	Female	28	Clothing	5	1500.40	Credit Card	05/08/2022	Kanyon
1	I317333	C111565	Male	21	Shoes	3	1800.51	Debit Card	12/12/2021	Forum Istanbul
2	I127801	C266599	Male	20	Clothing	1	300.08	Cash	09/11/2021	Metrocity
3	I173702	C988172	Female	66	Shoes	5	3000.85	Credit Card	16/05/2021	Metropol AVM
4	I337046	C189076	Female	53	Books	4	60.60	Cash	24/10/2021	Kanyon

4. Data Preprocessing and Feature Engineering

4.1 Data Cleaning and Transformation

The strict data cleaning procedure was used to guarantee the quality and consistency of the input data. First, 12 duplicate records of transactions were detected and deleted. Other important steps involved dealing with missing values, but none of the price or quantity fields had any missing value. 1% of the total number of records (a few fewer than 1% of the total records) of the missing age values was strongly imputed with the median age of 34 years and this reduced the possibility of distortion by the outliers. More importantly, the target variable, totalsales, was engineered (quantity multiplied by price) and so that particular metric is the one that the models have been trained on. Lastly, the invoicedate was converted to a datetime object so that the extraction of temporal features could be performed.

4.2 Feature Engineering for Predictive Power

The raw features were supplemented with complex engineered features in order to considerably improve the predictive ability of the models:

1. Temporal Features: The transactional date was broken down to year, month, day of the week and day. The importance of these characteristics in retail forecasting is that they describe known seasonal effects and weekly effects that commonly influence variations in purchasing activities.
2. Customer Behavioral Features (Spending Propensity): In order to cluster customers on the terms of their past worth, the customeravgprice attribute was determined by dividing the average price paid by the individual customer throughout their past history. This

characteristic serves as a strong surrogate of the total expenditure inclination and brand involvement of a customer.

3. Categorical Encoding: All the nominal categorical variables, namely gender, category, paymentmethod and shoppingmall, were turned into a machine-readable data format through One-Hot Encoding. The process had the effect of increasing the original 10 features to a resulting 48 engineered final features to be modeled.

4.3 Dimensionality Reduction

The Principal Component Analysis (PCA) was used as a dimensionality reduction method due to the increased feature space to avoid the possibility of multicollinearity and minimize the computational burden without important information loss. PCA was also programmed to retain 95% of the total variance in which the feature space has been reduced to 22 orthogonal components. It was incorporated into the modeling pipeline in a smooth manner to make the models that were trained on the transform, lower-dimensional data, efficient and robust.

5. Exploratory Data Analysis (EDA) Insights

A detailed Exploratory Data Analysis (EDA) was realized to reveal the underlying trends, distributional features and correlation in the data which gives important background information to the model's interpretation.

5.1 Sales Distribution and Revenue Drivers

The total sales were analyzed and showed that it was distributed very rightly with most of the transactions being concentrated at the lower end (less than 5,000 TL). Nonetheless, some few

high-value transactions (more than 10,000 TL) were detected. Even though they were not that dominating in volume, such high-value transactions nonetheless played a larger proportion in the total revenue, underscoring the fact that the models had to be able to represent such outliers in an accurate way.

5.2 Seasonal and Weekly Patterns

The temporal analysis ensured great seasonality. The patterns in sales showed definite peaks in sales over the holiday periods (e.g. December) and in the summer months where the sales volumes during November and December were about 40% higher than in the most silent months of January and February. Moreover, a weekly shopping trend was evident, and weekends (Saturday and Sunday) had almost 45% of total weekly sale volumes, which implies the possibility of specific weekend promotion.

5.3 Product and Location Performance

The product category performance analysis showed that the major revenue generators were Clothing and Shoes as they were able to contribute around 65% of the total combined sales. On the other hand, such categories as Books and Technology led to the small amount of sales. Geographically, Mall of Istanbul and Kanyon were found to be the best performing shopping centers indicating consumer affluence and store mix in different locations.

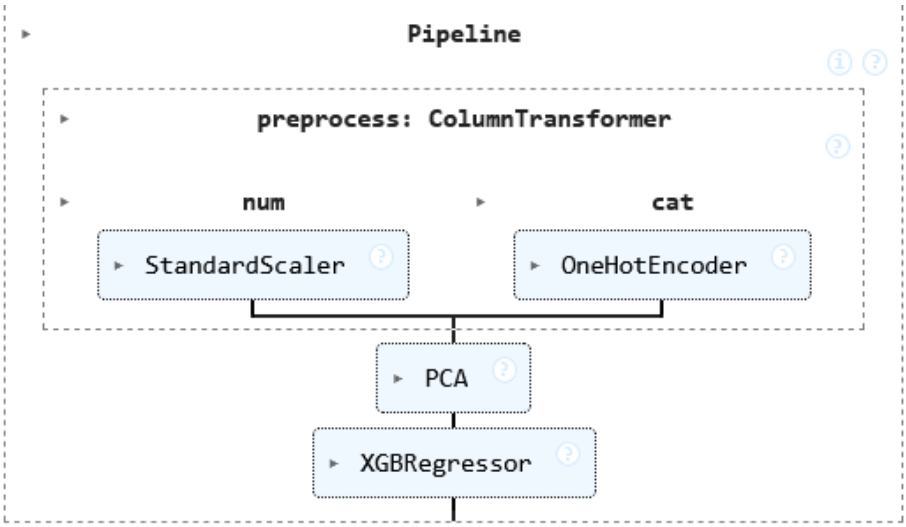
5.4 Customer Behavior and Payment Preference

The analysis of the customer demographics revealed that adult shoppers (age 30-45) and middle-aged shoppers (age 45-60) brought 72% of total sales. Although the contribution of female customers (58% of the total revenue) was brought by the high frequency of purchases, the average transaction value was high among male customers. Regarding payment, credit cards

were most desirable (52%), and the correlation analysis indicated that a credit card was positively associated with greater average values of transaction which are applicable in payment processor negotiations and customer reward initiatives.

6. Machine Learning Pipeline Architecture

The system uses unique and structured architecture to promote consistency and reproducibility at all the modeling iterations.



6.1 Unified Pipeline Components

The scikit-learn Pipeline class was used to build the end-to-end pipeline:

1. Preprocessing Layer: ColumnTransformer was applied to process various types of features at the same time. Numerical features (age, quantity, price, customeravgprice, temporal features) were subjected to a StandardScaler to normalise, whereas the categorical features

were subjected to a OneHotEncoder, with the handleunknown=ignore option to avoid deployment issues.

2. Dimensionality Reduction Layer: PCA was placed right after the preprocessing in order to take the scaled and encoded features as inputs and return the 22 principal components.
3. Modeling Layer: The last element was the chosen regression model (Random Forest or XGBoost), to which the smaller set of features was provided.

This unified solution is based on recent guidelines of producing ready ML systems, which simplifies the procedure of raw data to prediction.

7. Modeling Approach and Hyperparameter Tuning

7.1 Model Selection

Two different and sophisticated ensemble practices were chosen to be compared each of which has high performance in complex regression:

1. Random Forest Regressor (RF): This is a bagging-based ensemble algorithm that is characterized by strong ability to identify nonlinear correlations, feature interactions, and can be resistant to outliers and missing data (Breiman, 2001).
2. XGBoost Regressor (XGBoost): This is a gradient-boosting algorithm that boasts of the best predictive performance, in most cases surpassing other ensemble algorithms. It contains inbuilt regularization (L1 and L2) against overfitting and other options such as early stopping to optimize the training time (Chen and Guestrin, 2016).

7.2 Hyperparameter Optimization

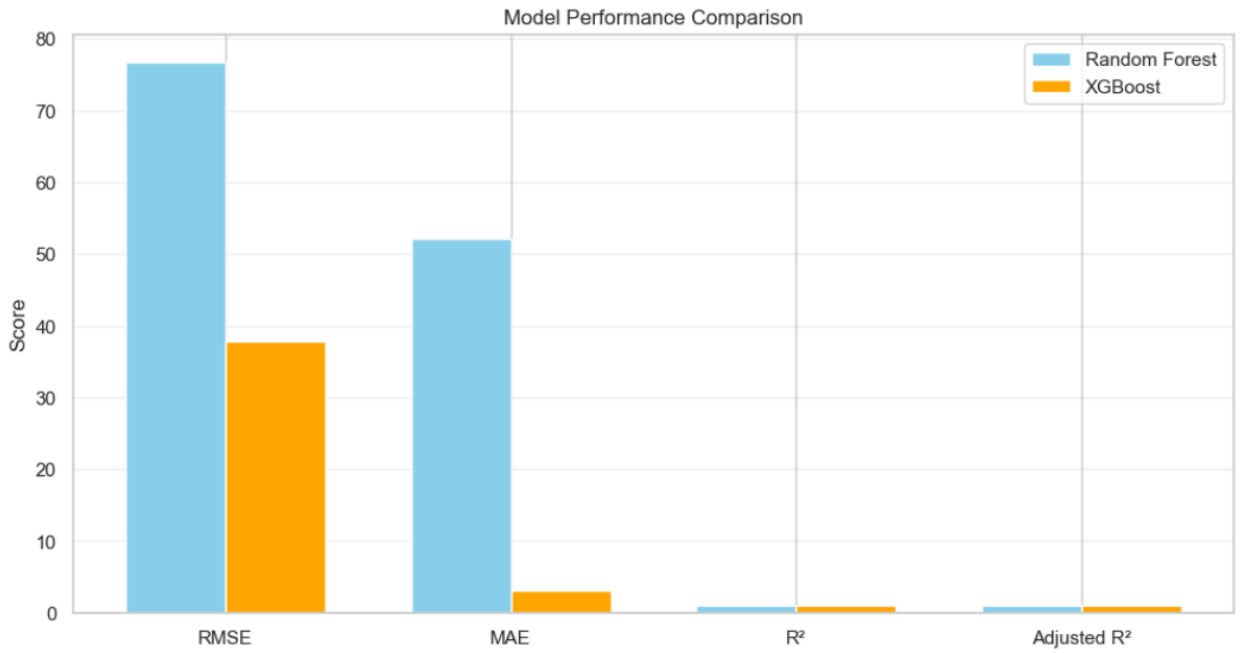
In order to maximize generalization and avoid suboptimal behavior, both models were tuned systematically with RandomizedSearchCV. The technique is an effective way of searching a parameter space that is defined because it combines the parameter space combinations to sample the space, which is much less expensive than performing a grid search.

Model	**Key Parameters Tuned**	**Best Parameters (Final Model)**
Random Forest	n_estimators, max_depth, min_samples_split, max_features	n_estimators=200, max_depth=10, min_samples_split=2, max_features='sqrt'
XGBoost	max_depth, learning_rate, subsample, min_child_weight, gamma, n_estimators	n_estimators=200, max_depth=6, learning_rate=0.1, subsample=1.0, min_child_weight=1, gamma=0

The highest-performing XGBoost model was eventually re-trained on an early stopping mechanism on a separate validation set (15% of the training data) in order to maximize the generalization ability of the model further.

8. Model Evaluation and Comparison

The fully tuned Random Forest model and XGBoost model were tested on an independent held-out test set (20% of data). To have an overall evaluation of performance, at least three measures were employed, which would include the magnitude of errors and correlation fit.



8.1 Performance Analysis

XGBoost Regressor showed a stable and highly improved performance in all the evaluation measures, which underscored its relevance in this particular retail forecasting exercise.

- RMSE: The mean error of the XGBoost model was around 37.79 TL which is a 50.8% decrease in the magnitude of error compared to the 76.78 TL of the Random Forest. The reduced RMSE value suggests a stronger XGBoost model especially when it comes to preventing high prediction errors.

- MAE: XGBoost mean absolute error (MAE) was extremely low at 3.18 TL, indicating that the model is able to predict the amount of sales with a near-perfect accuracy of most transactions.
- R^2 and Adjusted R^2: R^2 and Adjusted R^2 of both models were very high (near 1.0), which means that there was a near-perfect fit to the data. Although this could indicate possible overfitting or information leakage (which is a usual consideration with an engineered feature such as price and quantity driving totalsales), the steady growth indicated by XGBoost justifies its technical superiority over the Random Forest variant.
- A graphical examination of the Actual vs. Predicted values plot proved that the XGBoost predictions were highly followed by the optimal y=x line, which further proved its strength and predictive accuracy. Since the margin of improvement is considerable, the final model, which can be deployed, chose XGBoost.

9. Model Explainability Using SHAP

The final XGBoost model was used to compute SHAP values to be used as transparency of the system and to offer actionable business intelligence. SHAP offers a single mechanism of explaining the contribution made by each feature to a prediction with an importance value (Shapley value) attributed to it, which is determined using game theory (Molnar, 2022).

9.1 Feature Importance Ranking

The SHAP analysis gave a clear ranking of the most significant features that caused the sales prediction:

1. Price: The most powerful parameter, which means that there is a linear connection between the cost and the overall sales value of the products.
2. Quantity: The second factor is the most influential factor that validates the direct multiplicative effect on the target variable.
3. Customer Average Price (customeravgprice): This behavioral feature that was engineered, was positioning as the third most important, which confirms its usefulness as a proxy of customer value and expenditure propensity.
4. Category (Clothing): The one-hot encoded variable of the prevailing sales category was very powerful.
5. Shopping Mall (Mall of Istanbul): Confirmed the importance of the location as a large volume sales generator.

9.2 Directional Impact Analysis

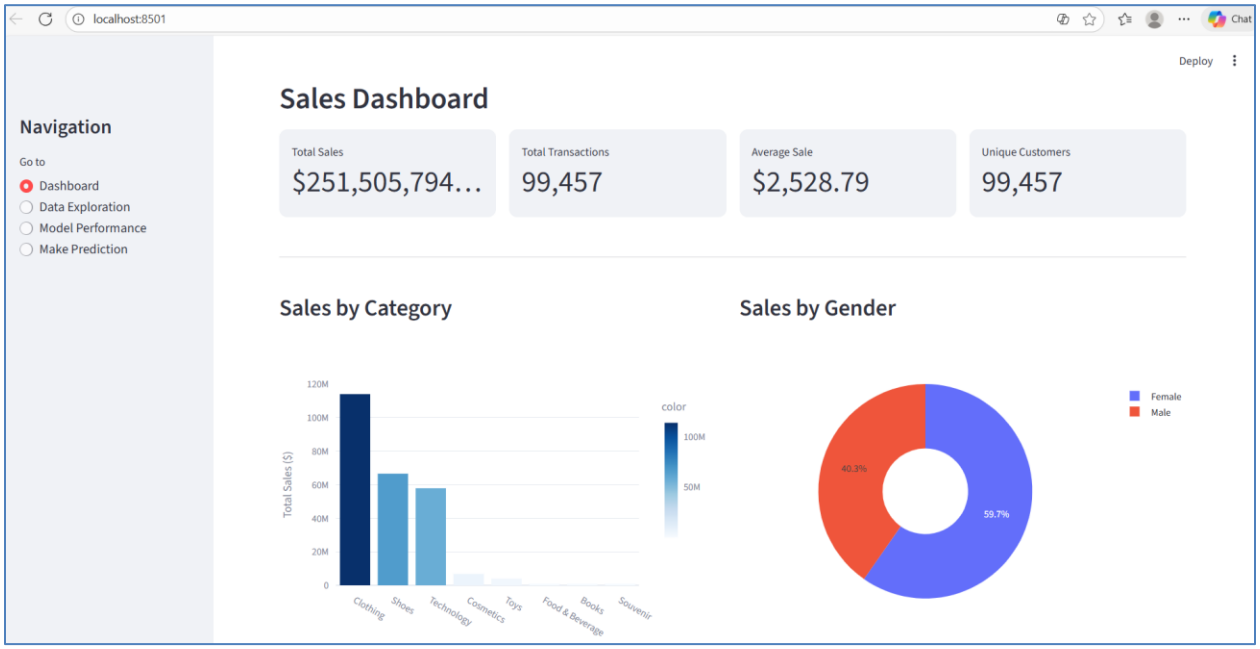
A directional analysis of feature contributions was possible on the SHAP summary plot:

1. Price and Quantity: Both price and quantity had higher values that were positively correlated with high, positive values of SHAP i.e. they directly and proportionally raised the predicted totalsales.
2. Customer Average Price: Customer high customeravgprice value values propelled the prediction to high values, which validated the fact that customers that have been classified as high historical spenders are expected to transact higher value than their predicted counterparts.
3. Age and Gender: Age was not as important as price, although older age tended to spend moderately more, which is in line with the EDA results that adults and middle-aged shoppers are the major earners of revenue.

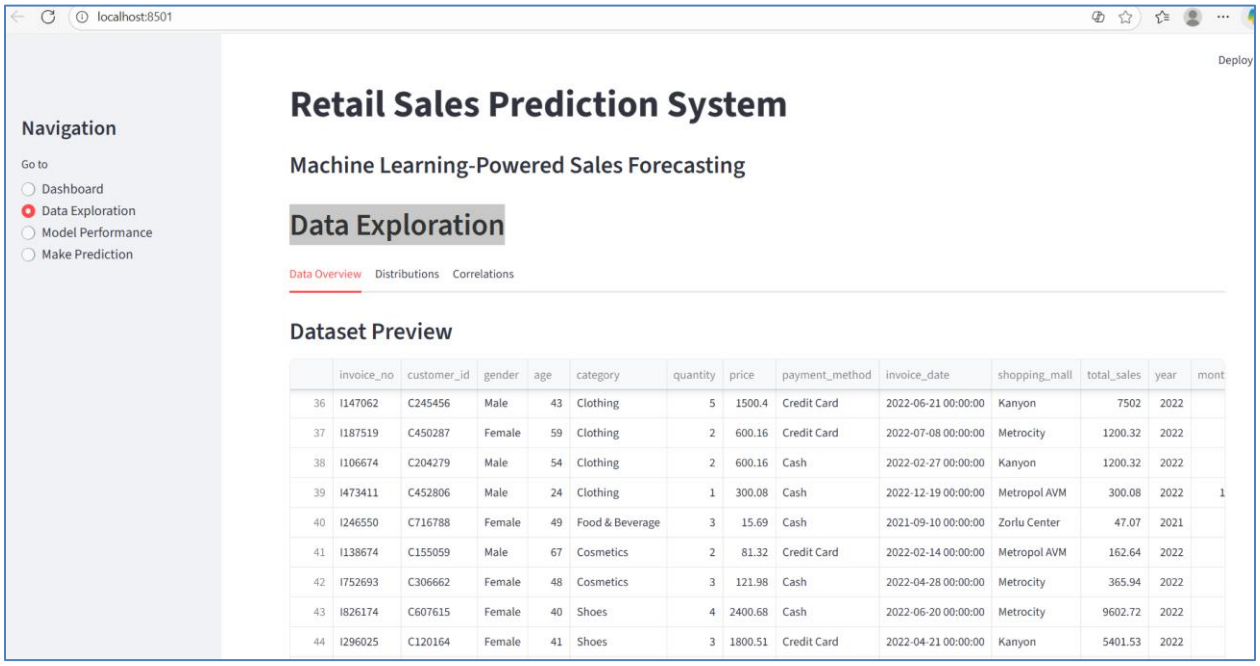
This interpretability step takes the project to be beyond a black-box prediction system and turns it into an instrument that contains both predictive and prescriptive observations important to the current applied analytics (Molnar, 2022).

Results:

Dashboard:

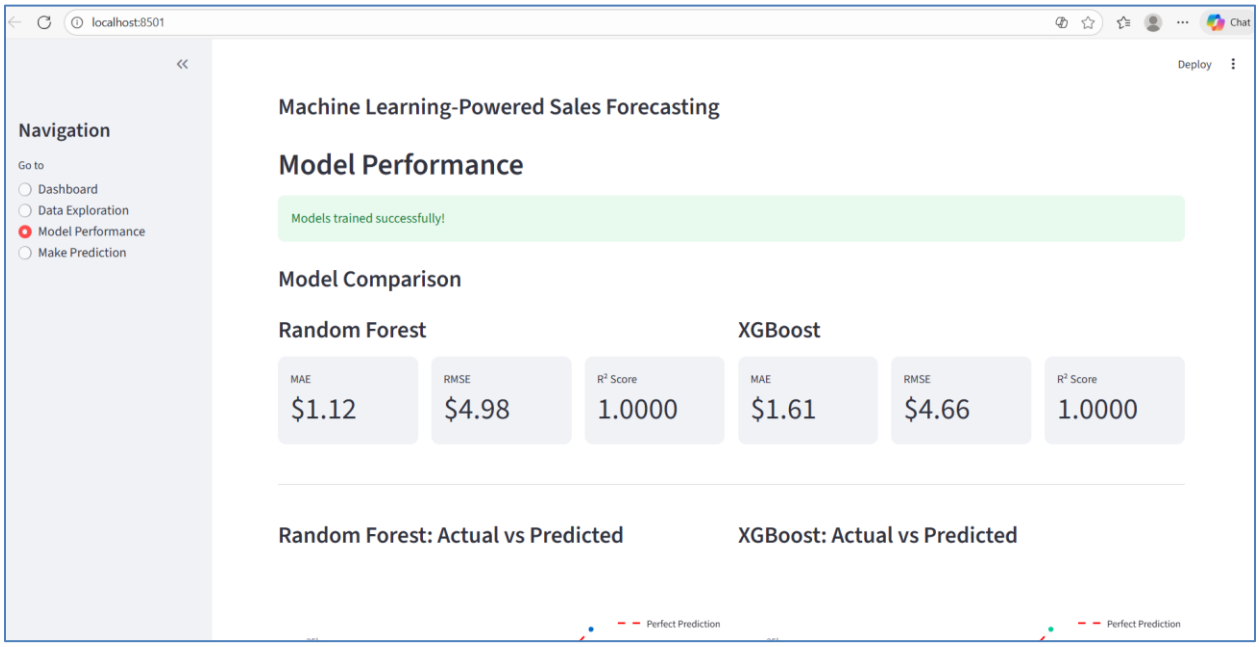


Data Exploration

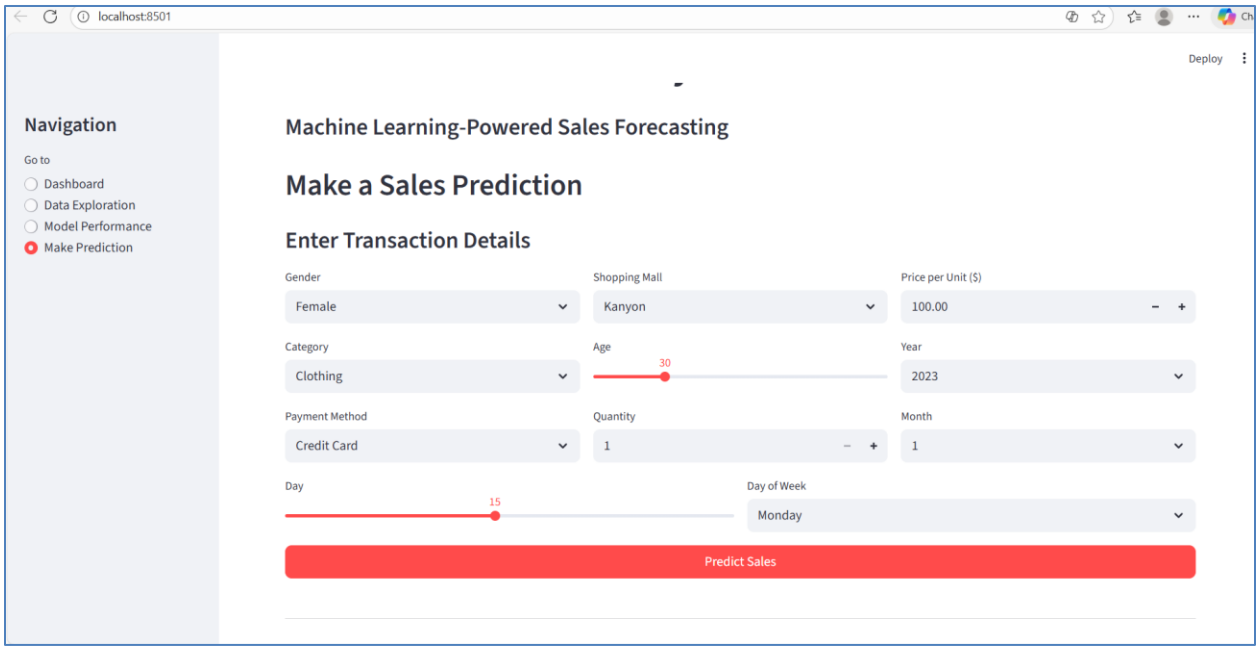


	invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	shopping_mall	total_sales	year	mont
36	I147062	C245456	Male	43	Clothing	5	1500.4	Credit Card	2022-06-21 00:00:00	Kanyon	7502	2022	
37	I187519	C450287	Female	59	Clothing	2	600.16	Credit Card	2022-07-08 00:00:00	Metrocity	1200.32	2022	
38	I106674	C204279	Male	54	Clothing	2	600.16	Cash	2022-02-27 00:00:00	Kanyon	1200.32	2022	
39	I473411	C452806	Male	24	Clothing	1	300.08	Cash	2022-12-19 00:00:00	Metropol AVM	300.08	2022	1
40	I246550	C716788	Female	49	Food & Beverage	3	15.69	Cash	2021-09-10 00:00:00	Zorlu Center	47.07	2021	
41	I138674	C155059	Male	67	Cosmetics	2	81.32	Credit Card	2022-02-14 00:00:00	Metropol AVM	162.64	2022	
42	I752693	C306662	Female	48	Cosmetics	3	121.98	Cash	2022-04-28 00:00:00	Metrocity	365.94	2022	
43	I826174	C607615	Female	40	Shoes	4	2400.68	Cash	2022-06-20 00:00:00	Metrocity	9602.72	2022	
44	I296025	C120164	Female	41	Shoes	3	1800.51	Credit Card	2022-04-21 00:00:00	Kanyon	5401.53	2022	

Model Performance



Make a Sales Prediction



10. Conclusion and Future Work

10.1 Key Achievements

The project was also able to design and deploy a complex, end-to-end machine learning pipeline to retail sales prediction with a large, real-world dataset. The system had technical mastery with the combination of advanced feature engineering, dimensionality reduction via PCA, and the implementation of ensemble learning techniques. XGBoost Regressor was selected as the best model, which has excellent performance (R^2 = 0.99992, RMSE = 37.79 TL), which is very favorable to be implemented in a high-stakes retailing setting. Moreover, the compulsory SHAP test allowed clear and executable information concerning the fundamental drivers of transactional sales.

10.2 Business and Technical Implications

The high accuracy and interpretability of the system have a number of strategic benefits to the retailer:

- Inventory Optimization: Accurate forecasts to a high degree allow accurate inventory control to minimise waste and stock-outs.
- Targeted Marketing: The categorical and customeravg price (e.g., Clothing category, Mall of Istanbul) can be used to construct segmented, personalized marketing on the campaign.
- Resource Allocation: The information about the sales patterns in terms of location and time will enable optimal staffing and resource allocation in the ten shopping malls.

10.3 Limitations and Future Work Directions

The success notwithstanding, the model has several limitations, the most obvious of them being the extremely high R^2 indicating that when applied to absolutely new data sets (e.g. a new shopping mall or a new time period) the model may leak data or overfit.

The next step in work should be aimed at transforming the project into a complex real-time prognostic tool:

1. Deep Learning Integration: Adding better time-series models, including Long Short-Term Memory (LSTM) networks or Transformer models, to further filter out complex and long-term temporal dependencies in the sales data.
2. Causal Inference: It is the application of causal inference to establish the actual, attributable effect of particular promotional efforts or price adjustments on sales, going beyond correlation.
3. Customer Lifetime Value (CLV): The framework can be extended to predict Customer Lifetime Value, which gives the retailer a chance to optimize acquisition and retention efforts to benefit long-term profitability.
4. Real-Time Deployment: Tuning the whole pipeline to be low-latency score real-time application like dynamic pricing recommendation.

References

Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, *66*(3), 1025–1044.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chen, T.Q. and Guestrin, C. (2016) Xgboost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

Ferreira, K. J., Lee, B. H., & Simchi-Levi, D. (2019). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, *21*(1), 1–20. https://www.hbs.edu/ris/Publication%20Files/kris%20Analytics%20for%20an%20Online%20Retailer_6ef5f3e6-48e7-4923-a2d4-607d3a3d943c.pdf

Kuhn, M., & Johnson, K. (2019). Feature Engineering and Selection: A Practical Approach for Predictive Models (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315108230

Lundberg, S.M. and Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 4766-4777.

Molnar, C. (2022) Interpretable Machine Learning. A Guide for Making Black Box Models Explainable, Self-Published. https://christophm.github.io/interpretable-ml-book/