Jeffery Dirdem
ITAI 1371
December 8, 2024

Module 5: Final Project - Predicting Sales Revenue (Turnover) - EXTRA CREDIT Report

The goal of this study was to use a synthetic dataset to forecast sales income (turnover) for a fictional smart electronics store. The dataset was created using several characteristics that are known to influence sales, such as the number of goods ordered, the price of each item, the discount applied, marketing spend, and customer reviews. These features were used to train and assess many machine learning models, including Linear Regression, Lasso Regression, Random Forest, Gradient Boosting, and XGBoost.

First, I built a synthetic dataset, ensuring that each feature was appropriate for a retail setting. Sales income was determined using a method that takes into account the quantity ordered, the price of each item, and the applicable discount. Marketing spending and customer evaluations were incorporated to replicate their impact on sales. After preprocessing the data and scaling the features, I divided it into training and test sets. Feature scaling is especially relevant for models such as Lasso and Linear Regression, where the size of the features can have a major impact on model performance. After preparing the data, I trained the models and evaluated their performance with measures like Mean Squared Error (MSE) and R-squared (R^2). The Linear Regression model, which assumes a linear connection between the characteristics and the objective variable, provided a baseline for performance. Lasso Regression, which uses regularization to prevent overfitting, performed similarly to the Linear Regression model.

Random Forest, an ensemble model that incorporates many decision trees, outperformed linear models because it can capture more complicated correlations in data. Gradient Boosting and XGBoost, both tree-based models that use boosting techniques, outperformed all other models, demonstrating their ability to handle nonlinear relationships and feature interactions.

The Random Forest, Gradient Boosting, and XGBoost models produced more accurate sales revenue projections than simpler models such as Linear Regression. These models had lower MSE values and higher R^2 values, indicating they explained more of the variance in sales revenue. This experiment demonstrated that machine learning algorithms, particularly tree-based models such as Random Forest and XGBoost, may be extremely effective at predicting sales revenue. They can capture the intricate connections between different factors, such as pricing, quantity, and marketing spend, that linear models may struggle to do accurately. For a real-world application, I would propose focusing on more complicated models for sales revenue prediction, as they perform best in this synthetic case.

Finally, machine learning models can be an effective tool for estimating sales income in retail environments. Businesses can acquire a better understanding of their sales dynamics and make more informed pricing, discount, and marketing decisions by utilizing models such as Random Forest, Gradient Boosting, and XGBoost. This predictive skill has the potential to help organizations streamline their operations and increase profitability.

Cited Sources

Scikit-learn. (2024). *User Guide*. Retrieved from https://scikit-learn.org/stable/user_guide.html

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Retrieved from https://xgboost.readthedocs.io/en/latest/

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.