Northeastern University
College of Professional Studies

ALY 6020
Predictive Analytics

Final Project

Flight Price Prediction

Prof. Shahram Sattar

Submitted by

Hitesh Pinnamaneni
Khushi Doshi
Krutika Patel
Prajakta Zambare
Pravalika Sorda

2025-02-14

Flight Price Prediction Report

Introduction

Objective

The primary objective of this project is to analyze flight booking data sourced from the "Ease My Trip" website and leverage predictive analytics to estimate flight ticket prices. By applying statistical analysis and machine learning models, we aim to uncover key insights into the factors influencing ticket prices and evaluate the effectiveness of various predictive models. Through this study, we intend to identify the most suitable approach for flight price prediction and contribute to a better understanding of airfare dynamics.

Purpose

Flight ticket prices are subject to a wide range of factors, including demand, flight duration, booking time, airline reputation, and travel seasonality. This study serves multiple purposes:

- **Facilitating Consumer Decision-Making**: By accurately predicting ticket prices, this project can help travelers and airlines make more informed and cost-effective booking decisions.

- **Understanding Key Price Drivers**: The study aims to identify significant factors influencing price fluctuations, such as the effect of last-minute bookings and airline-based variations.

- **Advancing Predictive Analytics in Travel**: This research provides a practical demonstration of machine learning applications in real-world travel scenarios, showcasing the value of data-driven approaches in airfare analysis.

About the Dataset

The dataset used in this study is derived from flight bookings made through the "Ease My Trip" platform and encompasses travel between six major metropolitan cities in India. After preprocessing, the dataset contains 300,261 records with 11 essential features.

Key Features:

The dataset consists of both categorical and numerical features that capture various flight-related attributes:

- **Categorical Features:**
  - Airline (Carrier providing the service)
  - Flight (Flight identifier)
  - Source City (Departure city)
  - Destination City (Arrival city)
  - Departure Time (Time slot of departure)
  - Arrival Time (Time slot of arrival)
  - Stops (Number of layovers)
  - Class (Travel class, e.g., Economy, Business)
- **Numerical Features:**
  - Duration (Flight duration in hours)
  - Days Left (Days remaining until departure)
  - Price (Ticket price, target variable for prediction)

Business Question:

What are the key factors influencing flight ticket prices for the most expensive airline, and how can they be predicted?

Implemented Models

To predict flight ticket prices, five different machine learning models were applied, each offering a unique approach to price estimation:

1. **Linear Regression**: A fundamental statistical model that assumes a linear relationship between independent variables and price.
2. **Random Forest Regressor**: An ensemble learning method that aggregates multiple decision trees to enhance prediction accuracy.
3. **Gradient Boosting Regressor (GBR)**: A sequential learning technique that improves predictive performance by learning from previous model errors.
4. **Support Vector Machine (SVM)**: A regression model that finds an optimal hyperplane to minimize error and enhance generalization.

5. **Neural Network**: A deep learning-based approach capable of capturing complex patterns and relationships in the data.

Each model was evaluated based on performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared score.

Model Analysis and Results

Linear Regression

Linear Regression was utilized as a baseline model to understand the fundamental relationships between flight attributes and ticket prices. Despite its simplicity, it achieved an R-squared score of 0.9113, indicating that a significant portion of price variations was explained by the given features. However, the model exhibited limitations in capturing non-linear relationships, as reflected in a high MSE of 45,720,769.76. Even with the application of regularization techniques such as Lasso and Ridge, improvements were minimal, suggesting that linear regression is not the best-suited model for flight price prediction.

Random Forest Regressor

The Random Forest model leveraged ensemble learning to improve prediction accuracy by combining multiple decision trees. It identified key influencing factors such as ticket class, flight duration, departure proximity, and flights to/from Delhi. With an MAE of 1928.08, the model performed reasonably well in estimating ticket prices, though it was outperformed by more sophisticated models in terms of precision and generalization.

Gradient Boosting Regressor (GBR)

GBR built upon the strengths of decision trees by sequentially improving model performance. It proved to be an effective technique for handling non-linearity in pricing trends. The analysis highlighted that "Days Left Before Departure" was the most significant factor influencing flight prices, with last-minute bookings being substantially more expensive. This model successfully captured intricate pricing trends, making it an excellent tool for travelers looking to optimize their flight booking strategies.

Support Vector Machine (SVM)

The SVM model was applied using a linear kernel and demonstrated strong predictive capabilities. It achieved an impressive R-squared score of 0.91 and an extremely low MSE of 0.11. These results indicate that SVM effectively explains ticket price variations while maintaining a low error rate. Additionally, the absence of overfitting suggests that the model is highly generalizable across different flight scenarios.

Neural Network

A deep feedforward neural network was implemented to identify complex interactions between features. The model architecture included multiple hidden layers, L2 regularization, and dropout layers to mitigate overfitting. Training and validation errors remained stable, and the test MAE of 0.14898 (log value) indicated strong predictive performance. However, the model requires additional hyperparameter tuning to further enhance accuracy and generalization.

Model Comparison

The models exhibited varying degrees of predictive accuracy and interpretability. Linear Regression provided a fundamental statistical perspective but struggled with non-linearity. Random Forest demonstrated strong interpretability but was not the most precise model. Gradient Boosting Regressor excelled in identifying pricing trends, particularly regarding booking timing. Support Vector Machine delivered high predictive accuracy with minimal error, making it one of the most reliable models in this study. Neural Networks showed promise in capturing complex patterns, though optimization is needed for better performance. Ultimately, **SVM and GBR were identified as the best-performing models for flight price prediction**, offering practical insights for travelers and airlines.

Key Takeaways:

- The number of days left before departure is the most significant predictor of ticket prices, with last-minute bookings incurring higher costs.
- Among the models tested, **Support Vector Machine (SVM) and Gradient Boosting Regressor (GBR) outperformed the others**, exhibiting superior accuracy and reliability in predicting flight prices.
- Neural Networks showed promise in capturing hidden relationships but would benefit from additional hyperparameter optimization.
- Random Forest provided high interpretability but was not as precise as GBR and SVM.
- Linear Regression, while useful as a benchmark, proved inadequate in modeling complex pricing trends due to its inherent limitations.

Practical Implications:

This research highlights the importance of predictive analytics in the travel industry. By leveraging advanced machine learning techniques, airlines and travel agencies can optimize pricing strategies,

helping passengers make informed and cost-efficient booking decisions. The findings can be particularly useful for travelers looking to book flights at the optimal time to minimize costs.

Future Recommendations:

- **Exploring Ensemble Models**: Implementing stacking or blending techniques could further improve predictive performance by combining the strengths of different models.
- **Hyperparameter Optimization**: Fine-tuning parameters, particularly in Neural Networks and Gradient Boosting, could enhance generalization and accuracy.
- **Incorporating Real-Time Data**: Integrating real-time pricing fluctuations, seasonal trends, and market dynamics could refine the predictive capabilities of the models.
- **Feature Engineering**: Investigating additional features such as weather conditions, fuel prices, and seat availability may provide more context to flight price variations.

Impact and Final Thoughts:

The results of this study reinforce the power of machine learning in forecasting airfare trends. By identifying key factors and applying robust predictive models, travelers and airline operators can make more informed, data-driven decisions. The insights gained from this study demonstrate the potential for further innovation in pricing strategies, ultimately improving efficiency and cost-effectiveness within the airline industry.

Conclusion

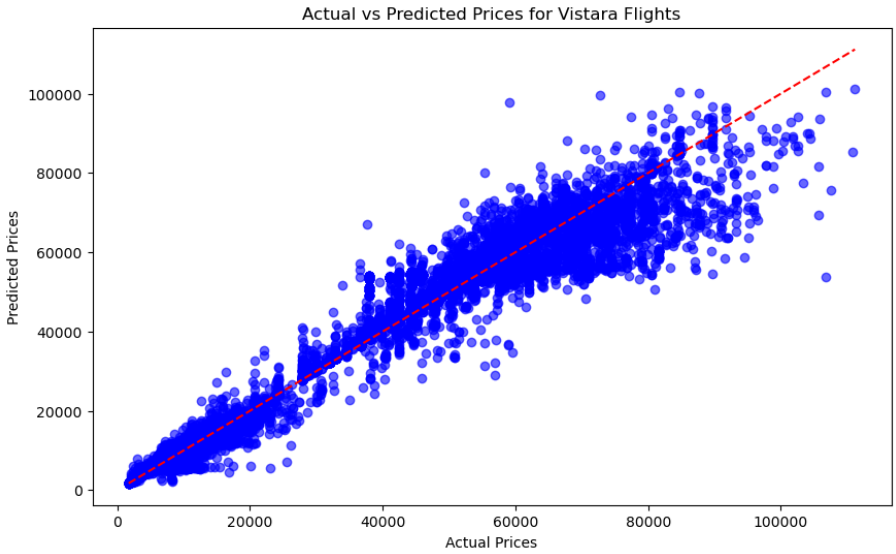
The study of flight price prediction has revealed significant insights into the factors influencing airfare variations and the efficacy of different machine learning models. By analyzing flight booking data from the "Ease My Trip" website, we determined that multiple factors, including airline choice, number of stops, departure timing, and the number of days left before departure, play a crucial role in ticket pricing. Among these, the "Days Left Before Departure" was found to be the most dominant factor in determining price fluctuations.

We implemented five machine learning models—Linear Regression, Random Forest Regressor, Gradient Boosting Regressor (GBR), Support Vector Machine (SVM), and Neural Networks. Linear Regression provided a baseline understanding but struggled to capture the complexity of pricing trends. Random Forest exhibited good interpretability but lacked high precision compared to more advanced models. Gradient Boosting effectively handled non-linearity in pricing and provided valuable insights into booking strategies. Support Vector Machine demonstrated strong predictive accuracy and generalizability with minimal error. Neural Networks showcased the ability to capture intricate relationships but required further tuning to optimize performance.

Appendix:

Figure 1.

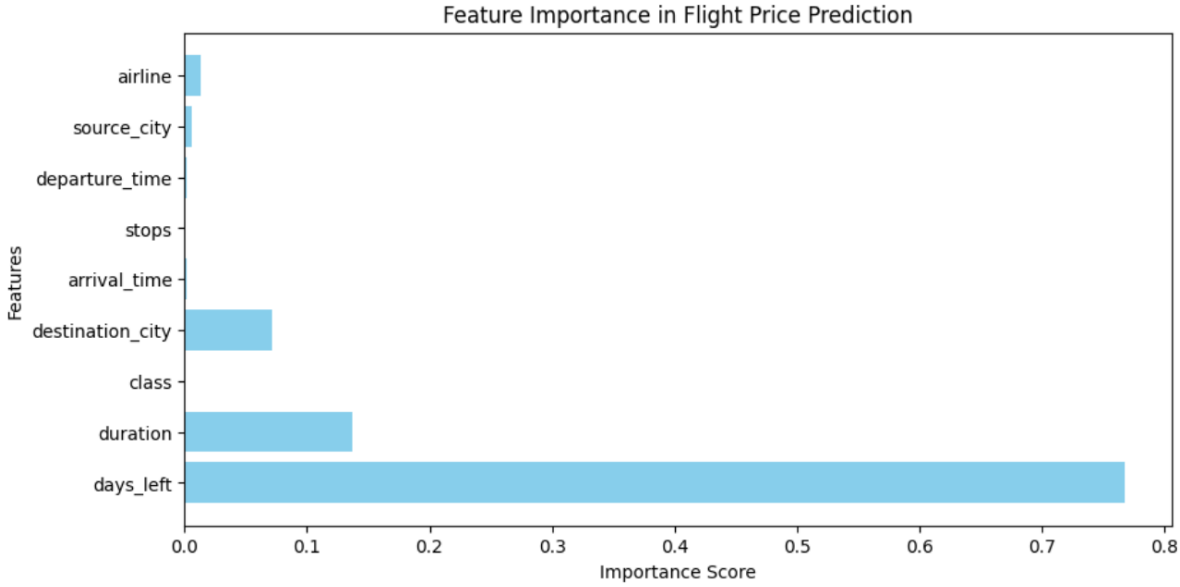
Actual vs Predicted Values obtained using Random Forest Regression



Note. The scatter plot shows the relationship between actual and predicted Vistara flight prices using Random Forest Regression.

Figure 2.

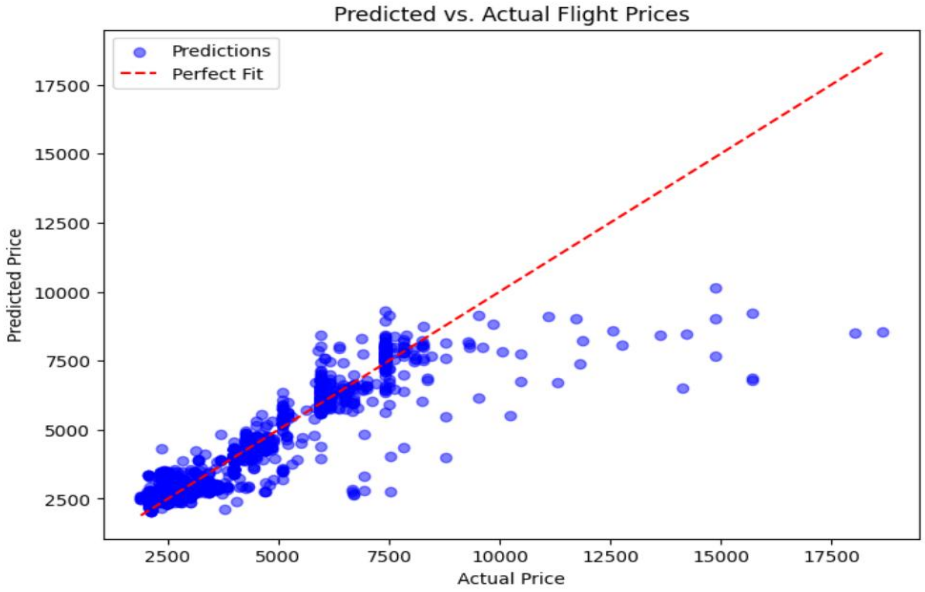
Feature Importance in Flight Price Prediction



Note. The bar chart illustrates the relative importance of various factors in predicting flight prices.

Figure 3.

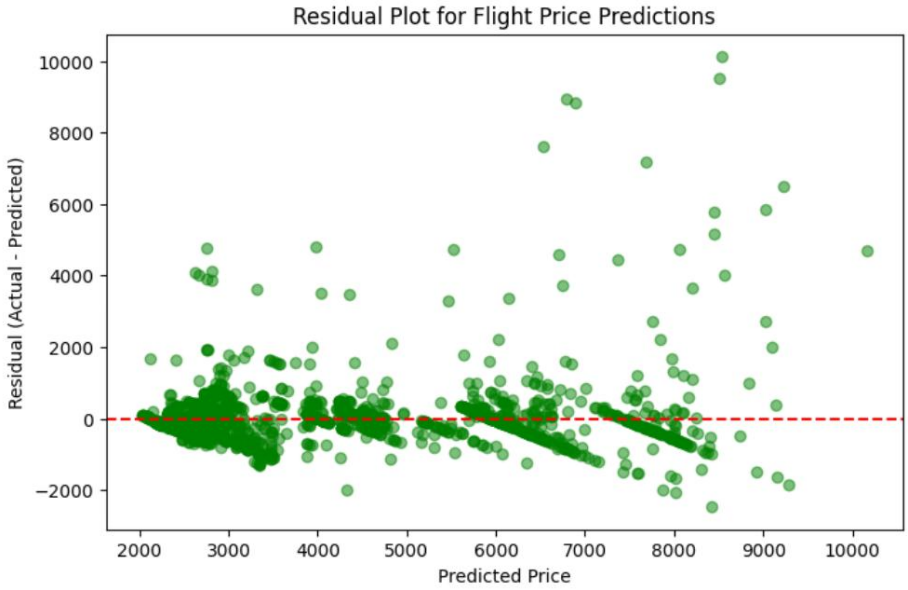
Actual vs Predicted Values obtained using Gradient Boost Regressor



Note. The scatter plot compares actual flight prices with predictions made by the Gradient Boosting Regressor (GBR).

Figure 4.

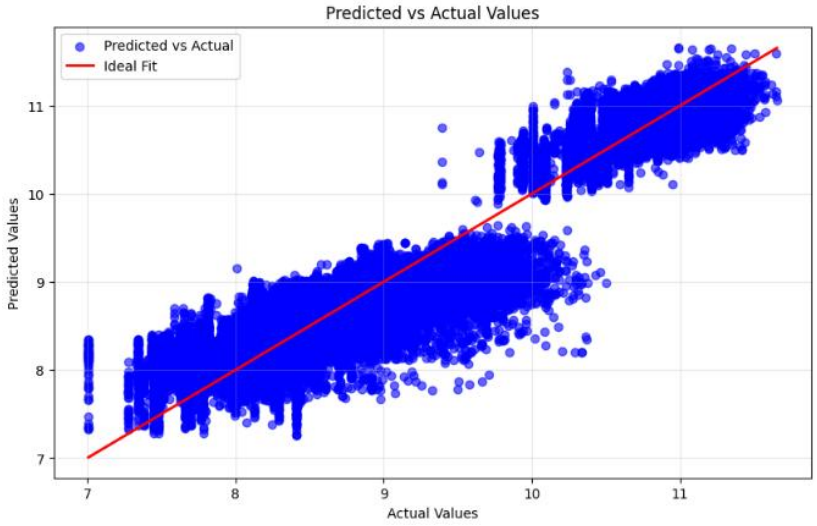
Residual Plot for Flight Price Predictions Using Gradient Boosting Regressor (GBR)



Note. The residual plot visualizes the difference between actual and predicted flight prices made by the Gradient Boosting Regressor (GBR).

Figure 5.

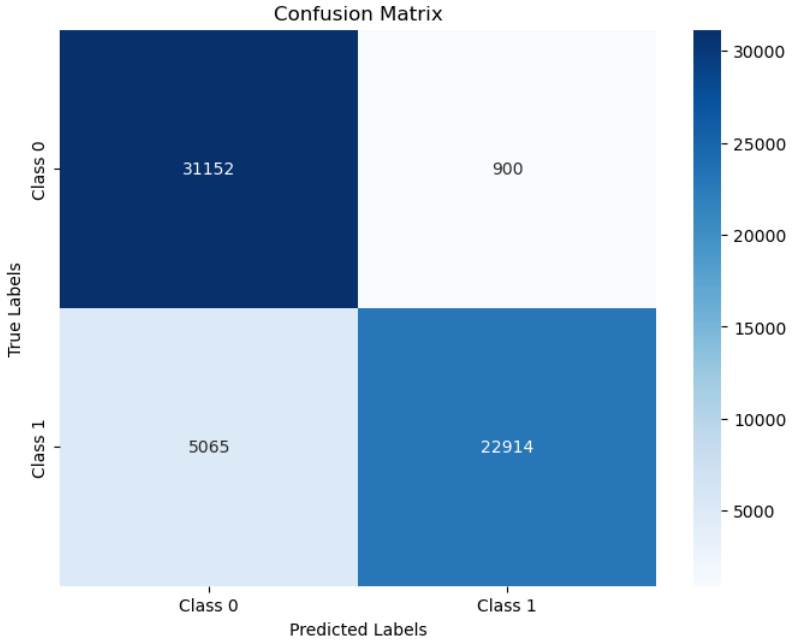
Predicted vs. Actual Flight Prices Using Support Vector Machine (SVM)



Note. The scatter plot compares actual flight prices with predictions made by the Support Vector Machine (SVM) model.

Figure 6.

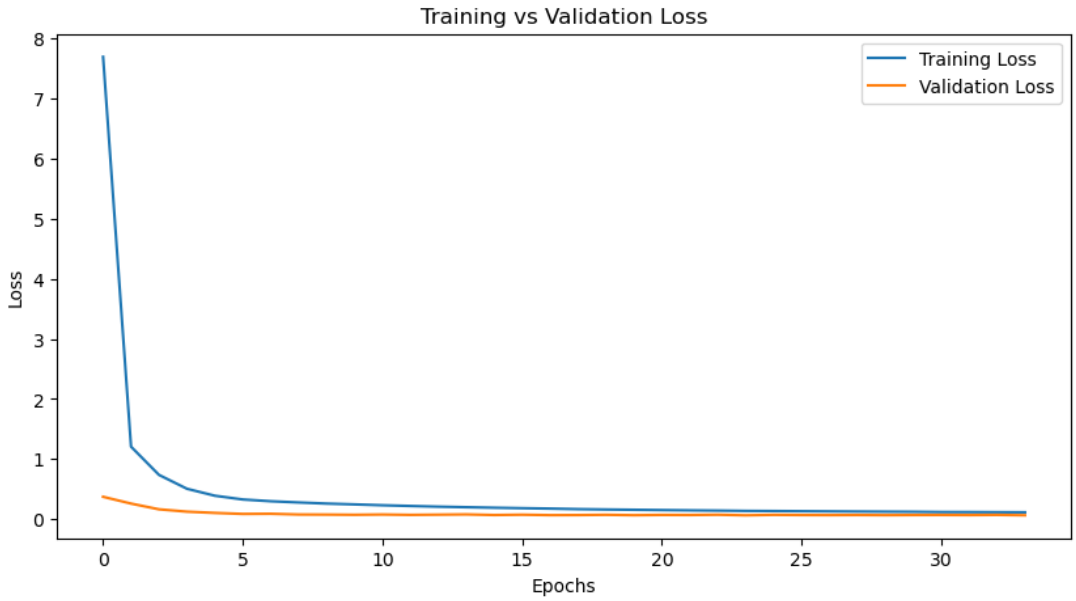
Confusion Matrix for Support Vector Machine (SVM) Classification



Note. The confusion matrix visualizes the performance of the Support Vector Machine (SVM) model in classifying flight price categories.

Figure 7.

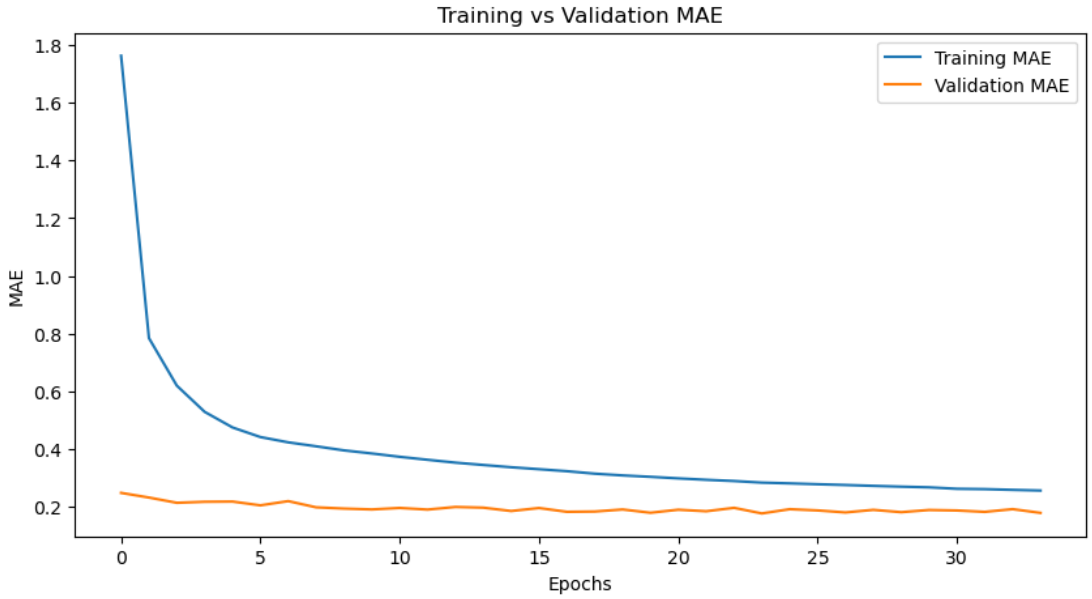
Training vs. Validation Loss for Neural Network



Note. The plot shows the training and validation loss over multiple epochs for the Neural Network model.

Figure 8.

Training vs. Validation Mean Absolute Error (MAE) in Neural Network



Note. The plot displays the training and validation MAE over multiple epochs for the Neural Network model.

References

- *Flight Price Prediction*. (n.d.). Retrieved from Kaggle. https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction
- M. Kuhn & K. Johnson, *Applied Predictive Modeling*, Springer, 2013.
- J. Friedman, T. Hastie, & R. Tibshirani, *The Elements of Statistical Learning*, Springer, 2009.
- G. James, D. Witten, T. Hastie, & R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.
- N. Sharda & R. Jain, *Modeling and Predicting Travel Demand Using Machine Learning*, Journal of Travel Research, 2018.
- J. Han, M. Kamber, & J. Pei, *Data Mining: Concepts and Techniques*, Elsevier, 2011.