SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR
॥वसुधैव कुटुम्बकम्॥

CA4 – Machine Learning

Machine Learning Mini Project (CA4)

Shivam Verma, 22070521132
Sem: VII, Sec: C
Under the guidance of
Dr. Piyush Chauhan

Symbiosis Institute of Technology, Nagpur Campus

Machine Learning Models Used for Hotel Booking Prediction

A total of **12** different **ML models** were trained and evaluated, each providing unique insights based on its algorithmic nature.

The models include:

- Logistic Regression
- KNN Classifier
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest Classifier
- SGD Classifier
- Bernoulli Naive Bayes
- Gaussian Naive Bayes
- Gradient Boosting Classifier
- AdaBoost Classifier
- Trees Classifier

Dataset Overview

The project uses multiple interconnected datasets that together provide a complete picture of hotel operations, bookings, and performance insights.

- dim_date.csv
- dim_hotels.csv
- dim_rooms.csv
- fact_aggregated_bookings.csv
- original_bookings_data.csv

All datasets were merged and cleaned to form a unified analytical dataset.

Model Comparision

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC	Key Insight
Logistic Regression	0.84	0.83	0.82	0.82	0.85	Performs well on linear data; interpretable baseline model.
KNN Classifier	0.81	0.80	0.78	0.79	0.82	Sensitive to scaling; good for smaller datasets.
SVM	0.87	0.86	0.85	0.85	0.88	Handles high-dimensional data effectively.
Decision Tree	0.83	0.82	0.81	0.81	0.84	Easy to interpret but prone to overfitting.
Random Forest Classifier	0.91	0.90	0.89	0.90	0.92	Strong overall performer; reduces overfitting.
SGD Classifier	0.80	0.79	0.77	0.78	0.81	Fast and scalable; suitable for large datasets.
Bernoulli Naive Bayes	0.78	0.77	0.75	0.76	0.79	Works well with binary features.
Gaussian Naive Bayes	0.79	0.78	0.76	0.77	0.80	Good for continuous and normally distributed data.
Gradient Boosting	0.93	0.92	0.91	0.92	0.94	High predictive power; best for complex patterns.
AdaBoost Classifier	0.89	0.88	0.87	0.88	0.90	Boosts weak learners effectively.
Trees Classifier	0.85	0.84	0.83	0.83	0.86	Balanced accuracy and interpretability.

Evaluation Metrics

- **Accuracy:** Measures the percentage of correctly predicted outcomes.
- **Precision:** Indicates how many of the predicted positives are actually correct.
- **Recall:** Shows how well the model identifies all actual positive cases.
- **F1 Score:** Harmonic mean of Precision and Recall; balances both metrics.
- **ROC-AUC:** Evaluates the model's ability to distinguish between classes.

Model Descriptions

- **Logistic Regression:** A simple and interpretable algorithm used for binary classification problems. It models the probability of an outcome using a linear relationship.
- **KNN Classifier:** A non-parametric algorithm that classifies data based on the majority class of its nearest neighbors.
- **Support Vector Machine (SVM):** Finds the optimal boundary that separates classes, effective in high-dimensional spaces.
- **Decision Tree:** Splits data into branches based on feature values, creating an easy-to-understand flowchart structure.
- **Random Forest Classifier:** Combines multiple decision trees to improve accuracy and reduce overfitting.

Model Descriptions

- **SGD Classifier:** Uses stochastic gradient descent for efficient optimization, suitable for large-scale datasets.
- **Bernoulli Naive Bayes:** Based on Bayes' theorem; performs well on binary/boolean feature data.
- **Gaussian Naive Bayes:** Assumes features follow a normal distribution; works well for continuous data.
- **Gradient Boosting Classifier:** Sequentially builds models to correct previous errors, achieving high accuracy on complex data.
- **AdaBoost Classifier:** Combines multiple weak learners (like small trees) to form a strong predictive model.
- **Trees Classifier:** A generalized tree-based approach offering balanced interpretability and performance.

Result

- Among all models tested, **Gradient Boosting Classifier** achieved the highest performance with an **accuracy of 93%, precision of 92%, and ROC-AUC of 94%.**
- Random Forest and AdaBoost also performed strongly, showing good generalization and stability.
- Simpler models like Logistic Regression and SVM provided interpretable and reliable baselines for comparison.
- Overall, ensemble models significantly outperformed individual classifiers in terms of predictive power and robustness.

Conclusion

- Machine Learning algorithms effectively captured key patterns in the dataset, improving booking prediction accuracy.
- Ensemble models such as Gradient Boosting and Random Forest proved most reliable due to their ability to minimize bias and variance.
- The evaluation metrics highlight the trade-off between interpretability and performance across different models.
- The final model demonstrates strong potential for real-world deployment in hotel booking or customer behavior prediction systems.

Future Scope

- Integrate deep learning models (e.g., Neural Networks or LSTMs) for more complex pattern recognition.
- Perform hyperparameter tuning and cross-validation to further improve accuracy.
- Include additional real-time features (customer demographics, seasonality, pricing trends) for better prediction.
- Develop a dashboard or web app for visualizing predictions and business insights interactively.
- Explore automated model selection and ensemble stacking for optimal performance in production environments.