PREDICTING HOTEL BOOKING CANCELLATIONS WITH MACHINE LEARNING

PREDICTING HOTEL CANCELLATIONS: A DATA-DRIVEN STRATEGY TO REDUCE NO-SHOWS AND OVERBOOKINGS

By Ayoola Ososanya

PROJECT **SUMMARY**

Objective:

To build a predictive model for hotel booking cancellations using machine learning. This helps hotel managers optimise bookings, reduce no-shows, and avoid costly overbooking.

Key Business Question:

Can we accurately predict whether a guest will cancel their booking using features like lead time, room type, and booking history?

METHODS **USED**

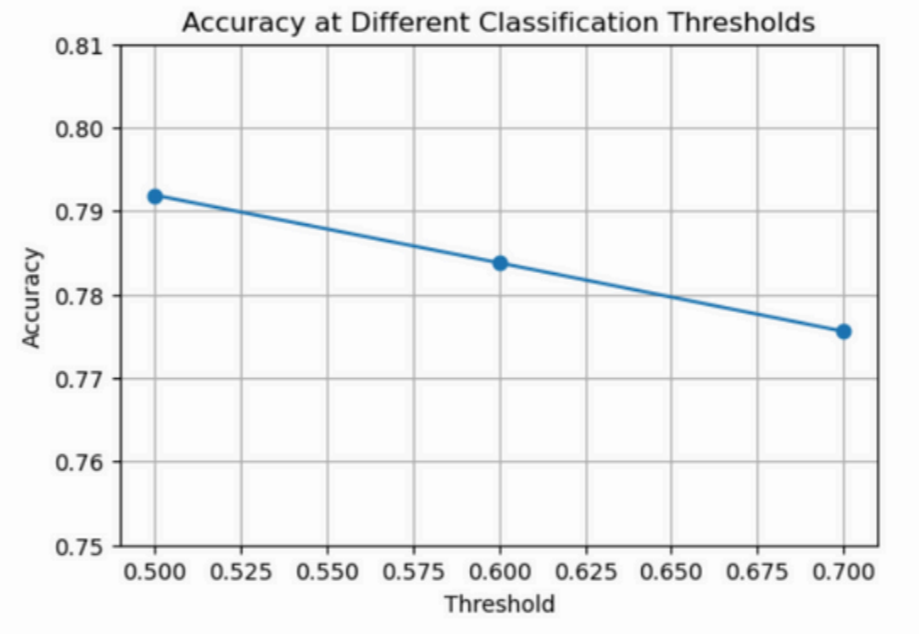
- Logistic Regression (with threshold tuning)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Ridge & Lasso Regression with Cross-Validation
- PCA for dimensionality reduction
- Feature reduction and interpretability analysis

KEY **RESULTS**

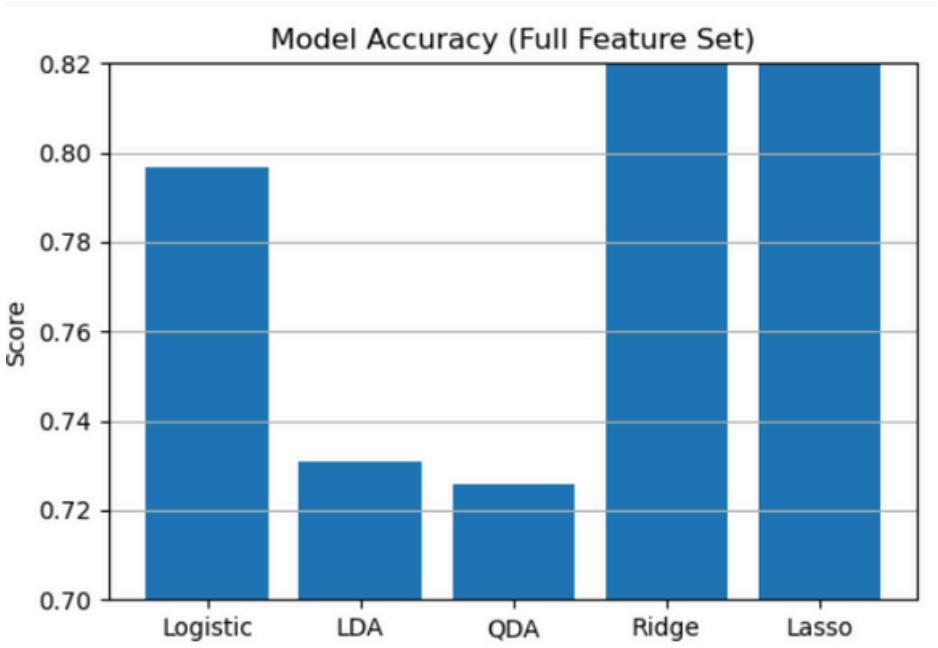
Model	**Accuracy(Full)**	**Accuracy (Reduced)**
Logistic Regression	79.7%	68.5%
LDA	73.1%	66.7%
QDA	72.6%	68.4%
Ridge	MSE = 0.1475	MSE = 0.2105
Lasso	MSE = 0.1496	MSE = 0.2105

VISUAL **INSIGHTS**

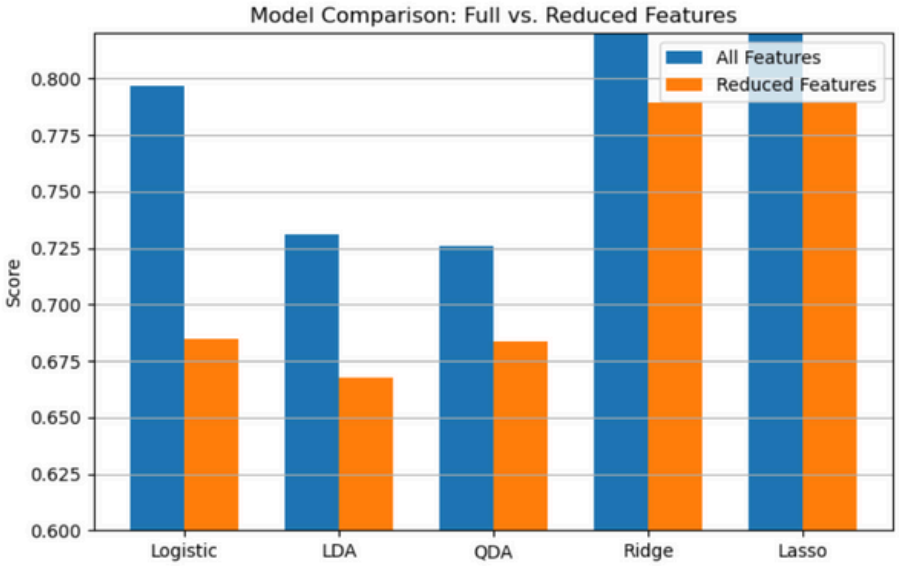
1. Threshold vs. Accuracy



2. Model Accuracy (Full Features)



3. Full vs Reduced Comparison



BUSINESS **TAKEAWAYS**

- Logistic Regression outperformed all other models for binary classification
- Threshold tuning (0.6) helps reduce false positives → fewer overbookings
- Feature reduction consistently led to weaker performance, especially in Ridge and Lasso
- Lasso identified 27 important predictors, aiding feature selection and operational focus

TOOLS **USED**

- PYTHON (PANDAS, SCIKIT-LEARN, MATPLOTLIB, SEABORN)
- JUPYTER NOTEBOOK

REFLECTION

As someone deeply interested in understanding patterns behind human behaviour — especially through my tutoring work — I was curious about what makes people cancel bookings. This project sharpened my ability to balance predictive accuracy with business implications like cost, risk, and operational capacity.