# Prediction Results

## Logistic Regression Model

The Logistic Regression model achieved an **accuracy of 82.4%**, correctly predicting the booking status in most cases.

**Confusion Matrix Analysis**

|  |  |  |
| --- | --- | --- |
|  | **Predicted: Not Canceled** | **Predicted: Canceled** |
| **Actual: Not Canceled** | 13,599 (True Negative) | 1,434 (False Positive) |
| **Actual: Canceled** | 2,768 (False Negative) | 6,077 (True Positive) |

* The model is highly reliable at detecting non-cancellations (TN = 13,599).
* It correctly identifies **6,077 out of 8,845** actual cancellations.
* However, **2,768 cancellations are missed**, indicating room for improvement in capturing potential cancellations.

**Classification Metrics**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Not Canceled (0)** | **Canceled (1)** |
| Precision | 0.83 | 0.81 |
| Recall | 0.90 | 0.69 |
| F1-Score | 0.87 | 0.74 |
| Support | 15,033 | 8,845 |

* **Precision (0.81)**: When the model predicts a booking will be canceled, it is correct 81% of the time.
* **Recall (0.69)**: The model captures 69% of all actual cancellations—**missing 31%**, which is notable for operational planning.
* **F1-score (0.74)**: A balanced score, indicating decent performance in predicting cancellations, but not yet optimal.

**Key Insights & Business Implications**

* **Strength**: Very strong at detecting non-cancellations (high recall = 90%).
* **Weakness**: Moderate performance in detecting true cancellations (recall = 69%).
* **Implication**: For a hotel, **missing cancellations can result in empty rooms and lost revenue**. Improving recall for cancellations should be a priority, even at the cost of slightly more false positives.

## K-Nearest Neighbors (KNN)

The KNN model achieved a **high overall accuracy of 90.9%**, correctly predicting 40 out of 44 cases in the test set.

**Confusion Matrix Analysis**

|  |  |  |
| --- | --- | --- |
|  | **Predicted: Not Canceled** | **Predicted: Canceled** |
| **Actual: Not Canceled** | 39 (True Negative) | 2 (False Positive) |
| **Actual: Canceled** | 2 (False Negative) | 1 (True Positive) |

* Strong performance in identifying **non-cancellations**.
* Poor performance in identifying **cancellations**—only 1 out of 3 actual cancellations was detected.

**Classification Metrics**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Not Canceled (0)** | **Canceled (1)** |
| Precision | 0.95 | 0.33 |
| Recall | 0.95 | 0.33 |
| F1-Score | 0.95 | 0.33 |
| Support | 41 | 3 |

* **Precision & Recall for Canceled Bookings** = 0.33:
  + Only 1 in 3 predicted cancellations is correct.
  + Only 1 in 3 actual cancellations is identified.
* **F1-score for Canceled Bookings** = 0.33:
  + Indicates poor balance between precision and recall for this critical class.

**Macro vs Weighted Averages**

* **Macro Average F1-score = 0.64**: Treats both classes equally; reveals imbalance in performance.
* **Weighted Average F1-score = 0.89**: Skewed by the dominant non-cancelled class.

**Key Insights & Business Implications**

* **Strength**: Accurately predicts bookings that will not be canceled.
* **Weakness**: Very poor at identifying cancellations (only 33% recall and precision).
* **Problem**: Likely a **class imbalance**—KNN is biased toward the majority class.
* **Business Risk**: Failing to predict cancellations can lead to overbooking issues, revenue loss, and poor operational planning.

## ML Model Comparison

**Model Performance Comparison: Logistic Regression vs KNN**

|  |  |  |
| --- | --- | --- |
|  | **Logistic Regression** | **K-Nearest Neighbors (KNN)** |
| **Accuracy** | 82.4% | **90.9%** |
| **Mean Squared Error** | 0.176 | **0.0909** |
| **True Negatives (TN)** | 13,599 | 39 |
| **False Positives (FP)** | 1,434 | 2 |
| **False Negatives (FN)** | 2,768 | 2 |
| **True Positives (TP)** | 6,077 | 1 |
| **Precision (Canceled)** | 0.81 – 81% of predicted cancellations were correct | 0.33 – only 33% were correct |
| **Recall (Canceled)** | 0.69 – identified 69% of true cancellations | 0.33 – missed 67% of actual cancellations |
| **F1-score (Canceled)** | 0.74 – solid balance of precision and recall | 0.33 – very weak performance |

**Conclusion & Recommendation**

|  |  |  |
| --- | --- | --- |
|  | **Winner** | **Why** |
| **Overall Accuracy** | KNN | Higher raw accuracy (but misleading due to imbalance) |
| **Cancellation Detection** | Logistic Regression | Much higher precision and recall for predicting class 1 |
| **Interpretability** | Logistic Regression | Better for business reporting and actionable insight |
| **Imbalance Robustness** | Logistic Regression | Performs better with class imbalance |
| **Scalability** | Logistic Regression | More efficient on larger datasets |

**Result Summary**

In this project, both Logistic Regression and K-Nearest Neighbors (KNN) were evaluated to predict hotel booking cancellations using historical booking data.

KNN achieved a higher overall accuracy (90.9%) compared to Logistic Regression (82.4%). However, this accuracy is misleading due to class imbalance—most bookings are not canceled, which skews the performance of KNN.

Logistic Regression significantly outperformed KNN in identifying actual cancellations, achieving a precision of 0.81 and recall of 0.69 for the cancellation class, versus KNN’s much lower precision and recall of 0.33. The F1-score for cancellations was 0.74 for Logistic Regression, indicating a better balance between false positives and false negatives, compared to KNN’s 0.33.

From a business perspective, correctly identifying cancellations is critical to avoid lost revenue and optimize room management. Therefore, despite its lower overall accuracy, Logistic Regression is the preferred model due to its stronger performance on the minority (cancellation) class, better interpretability, and robustness to imbalanced data.