Data Science Hotel Booking Cancellation Exploration

Jasin Doughmani

April 22, 2024

Contents

1 Datenexploration

1.1 Einleitung

1.1.1 Vorhersage von Buchungsstornierungen in der Hotelbranche

Die Analyse von Buchungsdaten und die Anwendung von Vorhersagemodellen zur Stornierung von Reservierungen bieten Hotels wertvolle Möglichkeiten, um ihre Ressourcen effektiv zu planen und das Risiko von Leerständen zu minimieren. Im vorliegenden Data-Exploration-Projekt wird untersucht, wie diverse Merkmale von Hotelbuchungen zur Vorhersage von Stornierungen herangezogen werden können. Darüber hinaus wird die Optimierung des Vorhersagemodells erörtert und erste Konzepte für eine praktische Umsetzung in der Hotelbranche werden diskutiert.

1.1.2 Beschreibung der Daten

Der vorliegende Datensatz umfasst 36.285 individuelle Datensätze von Hotelbuchungen aus den Jahren 2015 bis 2018. Diese Daten wurden direkt aus realen Hotelbuchungen erhoben, um sicherzustellen, dass die daraus abgeleiteten Vorhersagemodelle authentisch und glaubwürdig sind. Der Datensatz ist vielfältig und enthält eine breite Palette von Merkmalen, darunter folgende:

Merkmal	**Beschreibung**
number of week nights	Anzahl der Nächte unterhalb der Woche
number of weekend nights	Anzahl der Nächte am Wochenende
average price	Durchschnittlicher Zimmerpreis
number of adults	Anzahl der Erwachsenen
lead time	Vorlaufzeit der Buchung
market segment type	Buchungskanal (Offline oder Online)
special requests	Anzahl Sonderwünsche
booking status	Status der Buchung (Stornierung Ja/Nein)

Table 1: Beschreibung der Merkmale

Der Datensatz wurde am 16.02.2024 von Kaggle heruntergeladen und kann unter folgendem Link gefunden werden:
https://www.kaggle.com/datasets/youssefaboelwafa/hotel-booking-cancellation-prediction.

1.2 Charakterisierung des Datensatzes

1.2.1 Analyse der Datenqualität

Die folgenden Datenaufbereitungsschritte wurden durchgeführt, um sicherzustellen, dass der Datensatz für die Modellierung geeignet ist und ein gutes Vorhersagemodell erstellt werden kann:

- Umwandlung von Text in Binärwerte, da viele Machine-Learning-Algorithmen nur numerische Eingaben verarbeiten können.
- Umwandlung von Gleitkommazahlen in Ganzzahlen, um sicherzustellen, dass die Preise als ganze Zahlen vorliegen und das Modell einfacher zu interpretieren ist.
- Umwandlung des Datumsformats in separate Spalten: Daten werden in separate Spalten für Tag, Monat und Jahr aufgeteilt, um das Datum besser analysieren zu können und möglicherweise saisonale Trends zu identifizieren.

- Prüfen und Korrektur fehlender oder nicht-interpretierbarer Werte.

Die Qualitätsanalyse zeigte keine fehlenden oder nicht-interpretierbaren Werte auf.

1.2.2 Analyse der Balanciertheit

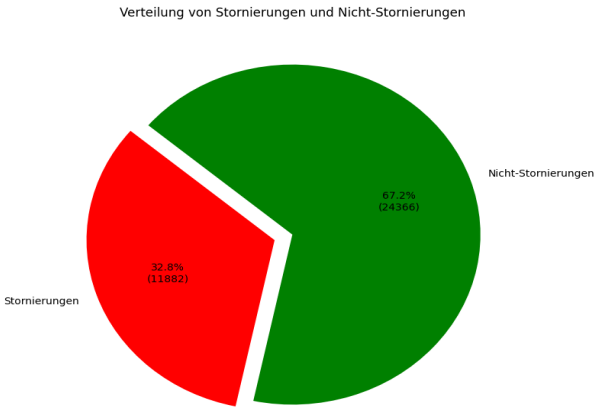
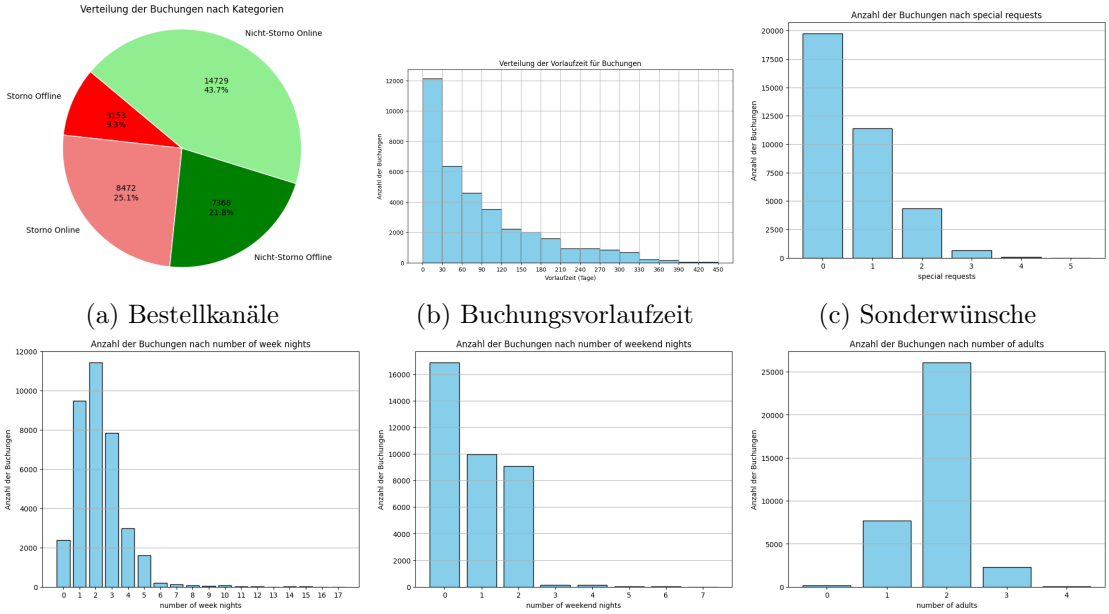


Figure 1: Balanciertheit der Stornierungen

Die Ungleichheit in der Verteilung zwischen Stornierungen und Nicht-Stornierungen deutet darauf hin, dass der Datensatz nicht vollständig ausbalanciert ist. Eine solche Ungleichheit könnte Auswirkungen auf die Leistung von Vorhersagemodellen haben, insbesondere wenn das Modell dazu neigt, die dominierende Klasse (hier Nicht-Stornierungen) über die unterrepräsentierte Klasse (hier Stornierungen) zu bevorzugen. Es ist wichtig, diese Ungleichheit zu berücksichtigen, da ein unausgewogener Datensatz dazu führen kann, dass das Modell Schwierigkeiten hat, seltene Ereignisse wie Stornierungen korrekt vorherzusagen. Dies kann zu einer Verzerrung der Vorhersagen führen, insbesondere wenn das Ziel darin besteht, Stornierungen genauer zu identifizieren. In solchen Fällen können verschiedene Techniken wie beispielsweise das Oversampling der unterrepräsentierten Klasse, das Undersampling der überrepräsentierten Klasse oder die Anpassung von Gewichtungen verwendet werden, um die Auswirkungen der Ungleichheit zu mildern und die Leistung des Modells zu verbessern. In diesem Projekt werden Upsampling und Gewichtung vorgenommen. Es ist hierbei wichtig zu beachten, dass beim Upsampling lediglich die Trainingsdaten betroffen sind. Die Testdaten bleiben unberührt, damit die Vorhersagen auf der realitätsnahen Verteilung basieren. Beide Techniken sollten den gleichen Effekt aufweisen.



(a) Bestellkanäle (b) Buchungsvorlaufzeit (c) Sonderwünsche

(d) Anzahl Wochentage (e) Anzahl Wochenendtage (f) Anzahl Erwachsene

Figure 2: Analyse der Verteilungen

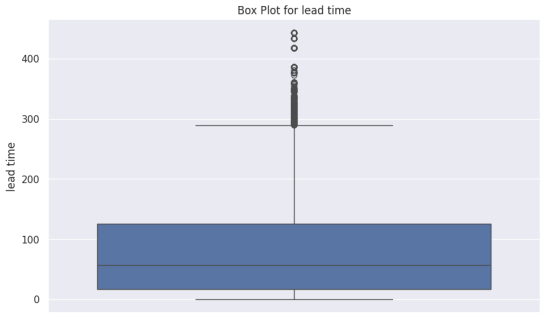
Die Diagramme zeigen hauptsächlich zu erwartende Verteilungen. Die Mehrheit der Buchungen erfolgen kurz vor Beginn des Aufenthalts, für wenige Tage und wenige Erwachsene mit keinen bis wenigen Sonderwünschen. Dies ist typisch für Hotelbuchungen, da die meisten Menschen ihre Reisen kurzfristig planen und buchen. Die wenigen Ausreißer mit längeren Aufenthalten scheinen hier durchaus plausibel.
Eine interessante Erkenntnis, die aus dem Diagramm (a) gezogen werden kann, ist die bedingte Wahrscheinlichkeit von Stornierungen für beide Kanäle.

	$Y = 0$ (Offline)	$Y = 1$ (Online)
$X = 1$ (Storno)	30%	36.52%
$X = 0$ (Kein Storno)	70%	63.48%

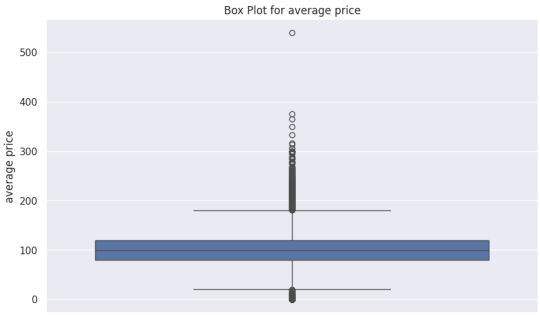
Table 2: Bedingte Wahrscheinlichkeiten für Stornierungen

Die berechneten bedingten Wahrscheinlichkeiten zeigen, dass Online-Buchungen wahrscheinlicher storniert werden als Offline-Buchungen.

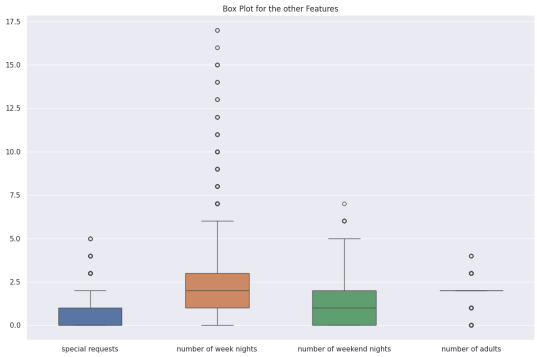
1.2.3 Analyse der Ausreißer



(a) Lead Time



(b) Price



(c) Other Features

Figure 3: Boxplots

Es sind nahezu keine auffallende Ausreißer zu erkennen. Die Entscheidung, Buchungen mit einer Vorlaufzeit von über 1 Jahr für das Training des Modells nicht zu berücksichtigen, basiert auf der Annahme, dass solche langfristigen Buchungen möglicherweise unrealistisch sind. In der Regel planen Kunden ihre Reisen nicht so weit im Voraus, und Buchungen mit einer derart langen Vorlaufzeit könnten ungewöhnliche oder unvorhersehbare Umstände widerspiegeln, die nicht repräsentativ für das normale Buchungsverhalten sind. Durch das Entfernen dieser Ausreißer kann sichergestellt werden, dass das Modell auf Daten trainiert wird, die realistische und typische Buchungsmuster widerspiegeln, was zu einer genaueren und zuverlässigeren Vorhersage der Stornierungswahrscheinlichkeit führt.

1.3 Feature Engineering

Mittels One-Hot-Encoding werden textuelle Werte in separate Spalten umgewandelt, um kategoriale Variablen in einem Format zu präsentieren, das für Machine-Learning-Algorithmen geeignet ist.
Durch die Bestimmung der Feature-Importance können wir verstehen, welche Merkmale oder Variablen in einem Datensatz am stärksten zur Vorhersage des Zielwerts beitragen. Dies ermöglicht es uns, die relevantesten Informationen zu identifizieren und die Modellkomplexität zu reduzieren, indem wir uns auf die wichtigsten Merkmale konzentrieren.

Feature	**Importance**
lead time	0.4262
average price	0.2242
special requests	0.1060
number of week nights	0.0661
number of weekend nights	0.0434
number of adults	0.0268
market segment type_Online	0.0241
market segment type_Offline	0.0195
Summe Feature-Abdeckung	0.9362

Table 3: Feature Importance

In diesem Projekt wird die Feature-Importance mithilfe eines Random Forest Classifiers berechnet. Random Forest ist ein leistungsstarkes Ensemble-Lernverfahren, das auf der Aggregation mehrerer Entscheidungsbäume basiert. Die Berechnung der Feature Importance im Random Forest basiert auf dem Gini-Index. Dieser misst die Unreinheit der Daten und wird verwendet, um zu bestimmen, wie gut ein bestimmtes Merkmal die Klassen trennen kann. Wenn ein Merkmal häufig verwendet wird, um die Klassen zu trennen, wird es als wichtiger angesehen. Es eignet sich gut für die Bestimmung der Feature-Importance, da es natürlicherweise die Relevanz verschiedener Merkmale während des Trainingsprozesses bewertet.
Die festgelegte Anforderung einer Feature-Abdeckung von mindestens 93% für dieses Projekt zielt darauf ab, sicherzustellen, dass unser Modell einen Großteil der relevanten Informationen im Datensatz berücksichtigt. Diese Entscheidung könnte in der Realität nach Rücksprache mit Fachleuten aus der Hotellerie getroffen werden, um die wirtschaftlichen Ersparnisse bzw. Kosten einer richtigen bzw. falschen Vorhersage einer Stornierung zu bewerten und somit sicherzustellen, dass keine wichtigen Merkmale übersehen werden. Dennoch dient diese pragmatische Festlegung dazu, die Komplexität des Projekts zu reduzieren und klare Richtlinien für die Auswahl der relevanten Merkmale festzulegen.

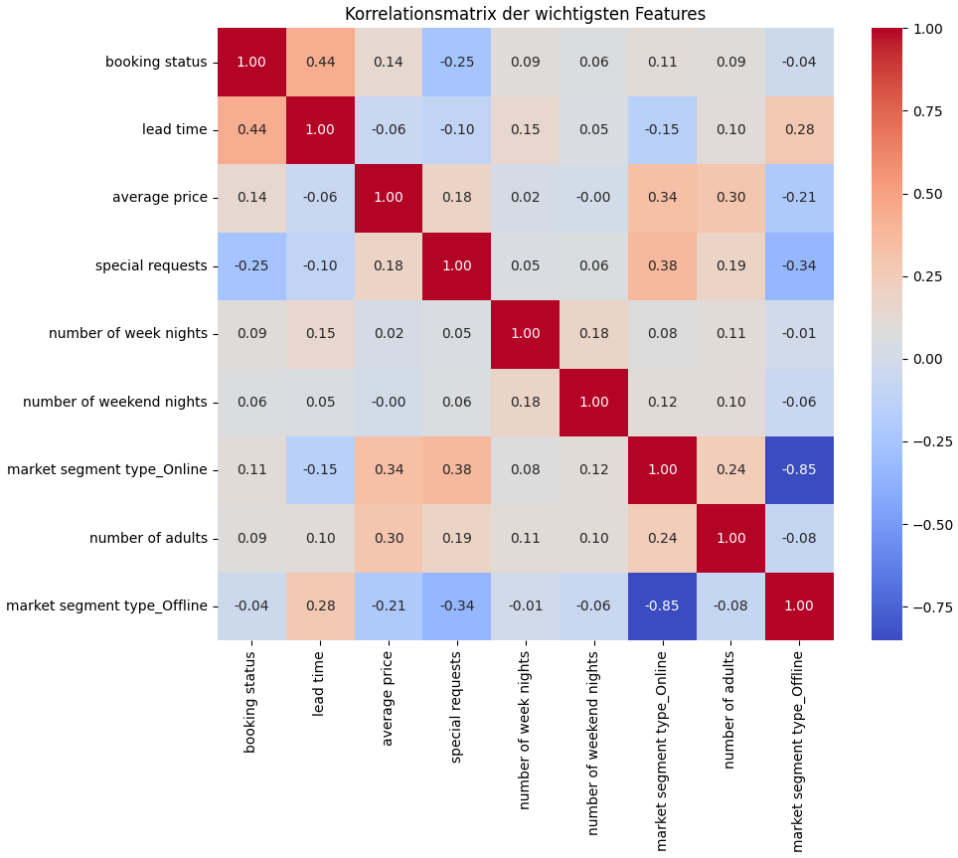


Figure 4: Korrelationsmatrix

Es ist zu beobachten, dass längere Vorlaufzeiten, höhere Preise, mehr gebuchte Nächte und mehr erwachsene Personen alle mit einer erhöhten Stornierungswahrscheinlichkeit einhergehen. Dies legt nahe, dass Kunden tendenziell eher stornieren, wenn sie frühzeitig, teuer, für längere Aufenthalte und für mehr Personen buchen. Eine bemerkenswerte Erkenntnis ist zudem die negative Korrelation zwischen der Anzahl der Sonderwünsche und der Stornierungswahrscheinlichkeit. Dies deutet darauf hin, dass Kunden, die viele Sonderwünsche haben, tendenziell weniger wahrscheinlich stornieren. Möglicherweise liegt dies daran, dass Kunden mit spezifischen Anfragen bereits eine feste Absicht haben, im Hotel zu übernachten, und sich daher weniger wahrscheinlich für eine Stornierung entscheiden.

1.4 Split des Datensatzes

Die ursprünglichen Daten werden in 80% Trainingsdaten und 20% Testdaten aufgeteilt, um eine ausgewogene Verteilung zwischen der Modelllernfähigkeit und der Evaluierung der Modellleistung sicherzustellen. Es ermöglicht eine robuste Modellentwicklung und -bewertung, ohne das Risiko von Überanpassung oder Unteranpassung zu erhöhen.

1.5 Auswahl der Metriken

Für die Optimierung des Modells wird der F1-Score als Metrik gewählt, da er ein ausgewogenes Maß für Precision und Recall bietet. Bei der Vorhersage von Buchungsstornierungen ist es entscheidend, sowohl wahre Stornierungen korrekt zu identifizieren (Recall), um potenzielle Umsatzverluste zu minimieren, als auch falsche Stornierungsvorhersagen zu reduzieren (Precision), um unnötige Kosten zu vermeiden. Die ausschließliche Optimierung von Precision oder Recall könnte zu einem Ungleichgewicht führen, das entweder zu vielen falsch

vorhergesagten Stornierungen oder zu vielen nicht erkannten Stornierungen führt. Der F1-Score ermöglicht eine ausgewogene Bewertung. Der ROC (Receiver Operating Characteristic) und der zugehörige AUC (Area Under Curve) bieten eine alternative Sichtweise auf die Leistung des Modells und sind besonders nützlich, um die Balance zwischen True Positive Rate (Recall) und False Positive Rate (FPR) zu verstehen.

1.6 Auswahl der ML-Methode

Für die Ermittlung der am geeignetesten Methode für das Vorhersagemodell werden mehrere Methoden implementiert und mittels der ausgewählten Metrik gegenübergestellt.

ML-Methode	**F1-Score**
Logistic Regression	0.76
KNeighbors	0.77
Decision Tree	0.83
GBM	0.80
Random Forest	0.86

Figure 5: F1-Scores der ML-Methoden

Jeder Baum im Random Forest wird auf einer zufälligen Teilmenge der Trainingsdaten trainiert, ein Prozess bekannt als Bootstrapping. Während des Trainings wählt jeder Baum zufällig eine Teilmenge der verfügbaren Merkmale aus, um die beste Trennung für jeden Knoten zu finden. Dies führt zu einer Vielzahl von Bäumen, die individuell schwach sein können, aber durch ihre Kombination ein robustes und präzises Vorhersagemodell bilden.

1.7 Hyperparameter-Tuning

Mittels Grid-Search und der systematischen Variation von Parametern wie der Anzahl der Bäume im Wald, maximaler Tiefe jedes Baumes und der Mindestanzahl von Beispielen in einem Baumknoten wird die bestmögliche Hyperparameter-Kombination ermittelt. Dies trägt dazu bei, Overfitting zu reduzieren und die Generalisierungsfähigkeit des Modells zu verbessern, was letztendlich zu präziseren und zuverlässigeren Vorhersagen führt.

1.8 Vorhersage-Demo

Durch die zufällige Auswahl eines Datenpunktes, der Vorhersage dessen Labels und der Ausgabe des vorhergesagten und realen Buchungsstatus wird eine Simulation des Vorhersagemodells umgesetzt.

2 Evaluation und Ergebnisse

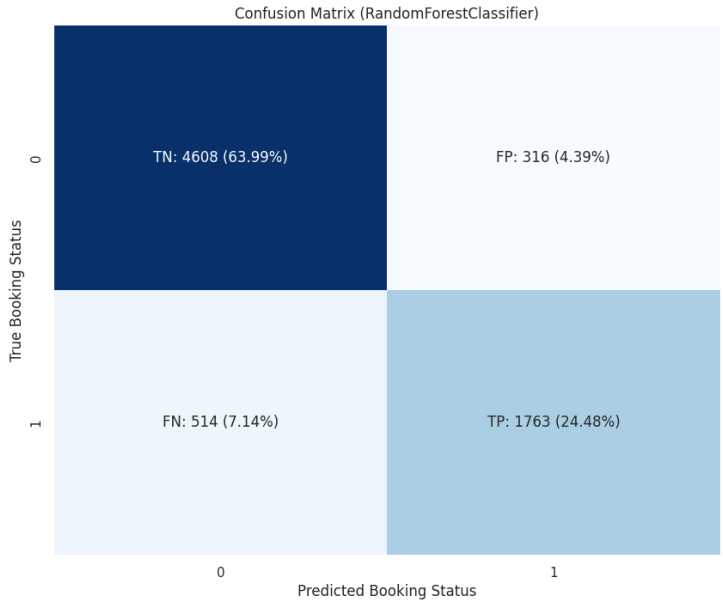
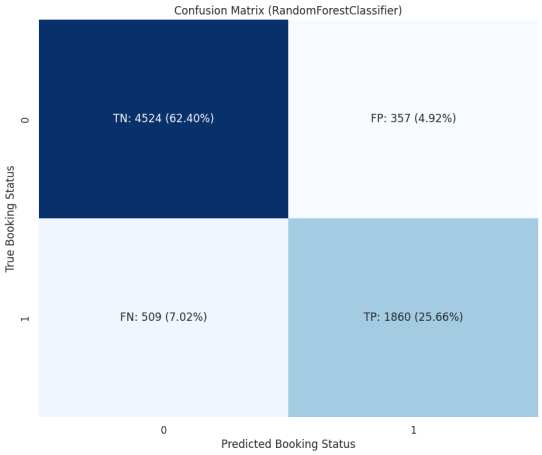
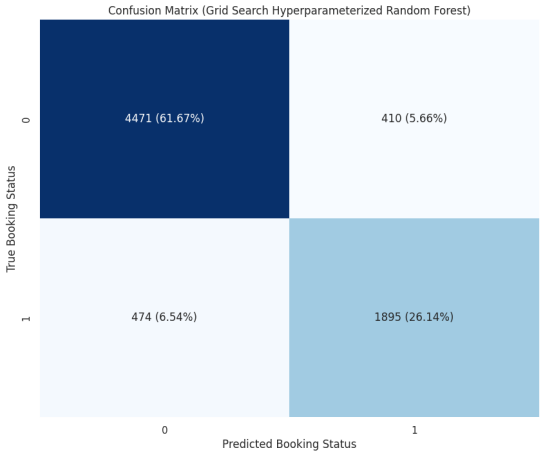


Figure 6: Confusion Matrix für den Random Forest (Nicht ausbalanciert)

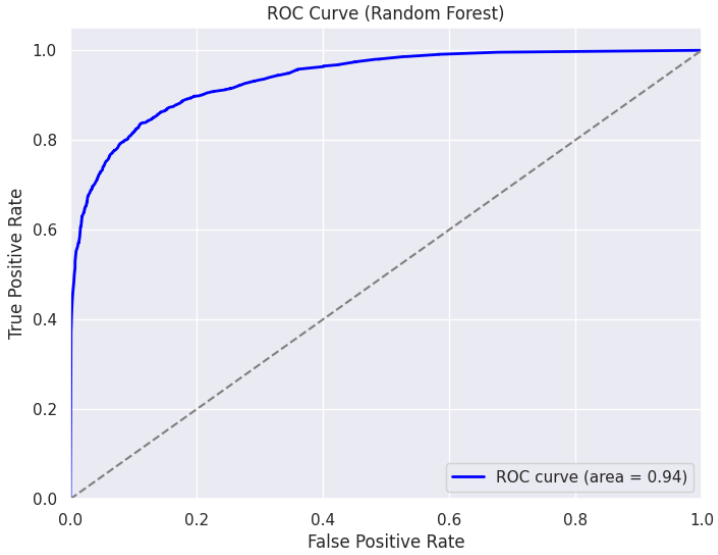
Der F1-Score für Stornierungen beträgt bei dem unbalancierten Datensatz 0.81, während er für Nicht-Stornierungen 0.91 erreicht. Der binäre F1-Score dieses Modells beträgt somit ca. 0.86. Der ROC-Wert liegt hier bei 0.94.



(a) Random Forest (Upsampled)



(b) Random Forest (Tuned)



(c) ROC-Kurve für den Random Forest (Upsampled)

Figure 7: Confusion Matrices und ROC-Kurve für den Random Forest

Insgesamt zeigt die Analyse der Vorhersage von Hotelbuchungsstornierungen mithilfe verschiedener Machine-Learning-Modelle wie Random Forest, Logistic Regression und k-NN einen vielversprechenden Ansatz. Das Random Forest-Modell erreichte von allen Modellen den besten F1-Score von ca. 0.86, was auf eine gute Leistung bei der Vorhersage von Stornierungen hinweist. Der ROC-Wert von 0.94 deutet auf eine hohe Trennschärfe zwischen den Klassen hin. Das Hyperparameter-Tuning sowie das Korrigieren der Balanciertheit mittels Upsampling und Gewichtung konnten kaum bis keine Optimierungen erzielen. Dies könnte daran liegen, dass der Random Forest durch seine Eigenschaften des Bootstrapping und einer entsprechenden Menge an Trainingsdaten bereits sehr robust gegenüber Overfitting ist und somit eine hohe Generalisierungsfähigkeit aufweist. Verglichen mit dem Kaggle-Benchmark von ca. 0.85 (F1-Score) konnte eine minimale Verbesserung trotz einer geringeren Feature-Anzahl erzielt werden. Dies liegt mit hoher Wahrscheinlichkeit an den entfernten Datensätzen der definierten Ausreißer. Zukünftige Erweiterungen dieses Modells könnten einen Vergleich mit hyperparameterisierten Classifiern beinhalten sowie eine detaillierte Betrachtung, welche Features zur Verbesserung des Vorhersagemodells beitragen und die Ermittlung der optimalen Feature-Anzahl.

Um das Modell erfolgreich in der Praxis anzuwenden, müssten jedoch einige Schritte unternommen werden. Zunächst ist es wichtig, das Modell an die spezifischen Gegebenheiten und Bedürfnisse des Hotels anzupassen, indem es auf die eigene Buchungsdatenbank trainiert wird. Dabei ist eine sorgfältige Evaluation der Ergebnisse und Anpassungen an das Modell erforderlich, um eine hohe Vorhersagegenauigkeit sicherzustellen. Ein weiterer entscheidender Aspekt bei der Anwendung des Modells ist die Berücksichtigung der wirtschaftlichen Kosten und Folgen der einzelnen Fälle. Eine falsche Vorhersage von Stornierungen kann zu erheblichen Umsatzverlusten führen, da das Hotel möglicherweise Zimmer blockiert und Ressourcen für Gäste bereitstellt, die letztendlich nicht erscheinen. Auf der anderen Seite kann auch eine falsche Vorhersage von Nicht-Stornierungen zu Problemen führen, da das Hotel möglicherweise nicht genügend Zimmer oder Dienstleistungen für tatsächliche Gäste zur Verfügung hat. Dies kann zu unzufriedenen Kunden, negativem Feedback und langfristigen Schäden für das Hotel führen. Daher ist eine sorgfältige Abwägung der Kosten und Nutzen sowie eine kontinuierliche Überwachung und Optimierung des Modells unerlässlich, um eine erfolgreiche Implementierung in der Hotelbranche zu gewährleisten.