UNIVERSIDADE REGIONAL INTEGRADA DO ALTO URUGUAI E DAS MISSÕES
PRÓ-REITORIA DE ENSINO, PESQUISA, EXTENSÃO E PÓS-GRADUAÇÃO
CÂMPUS ERECHIM – RS
CURSO DE CIÊNCIA DA COMPUTAÇÃO

Felipe Rodighero Zarichta

TÓPICOS ESPECIAIS EM COMPUTAÇÃO I

ERECHIM - RS
2024

Sumário

1- Introdução

O dataset escolhido para o desenvolvimento deste projeto é denominado **Hotel Reservations**, disponível na plataforma Kaggle. O objetivo principal consiste em prever se um hóspede irá **cancelar ou não uma reserva em um hotel**, utilizando variáveis preditivas disponíveis no conjunto de dados. A crescente popularização das plataformas de reservas online transformou significativamente o comportamento dos clientes, trazendo tanto benefícios quanto desafios para o setor hoteleiro. Um dos principais problemas enfrentados atualmente pelos hotéis é o alto número de **cancelamentos e não comparecimentos**, os quais impactam diretamente na gestão e na receita. Entre os principais motivos que levam os clientes a cancelarem suas reservas estão mudanças de planos, imprevistos, conflitos de agenda e a facilidade oferecida por muitas plataformas, que permitem cancelamentos gratuitos ou com baixo custo. Embora essa política seja vantajosa para os clientes, representa um desafio para os hotéis, que precisam lidar com a incerteza na ocupação e possíveis prejuízos financeiros. Este projeto visa, portanto, aplicar técnicas de **aprendizado de máquina** para construir um modelo capaz de **prever a probabilidade de cancelamento de uma reserva**, auxiliando os hotéis na tomada de decisões estratégicas e na mitigação de riscos.

Dataset: Hotel Reservations
Projeto no GitHub: https://github.com/feliperodighero/hotel-reservations

2- Desenvolvimento

O desenvolvimento do projeto teve início com a escolha criteriosa do dataset, uma etapa fundamental para garantir a qualidade dos resultados. Para isso, foram

selecionados dados provenientes de uma fonte confiável e amplamente utilizada pela comunidade de ciência de dados.

2.1- Análise dos Dados

A análise exploratória dos dados (EDA) foi realizada para compreender a estrutura do dataset, identificar padrões, anomalias e relações entre as variáveis. O dataset Hotel Reservations.csv contém 36.275 entradas e 19 colunas, abrangendo informações como:

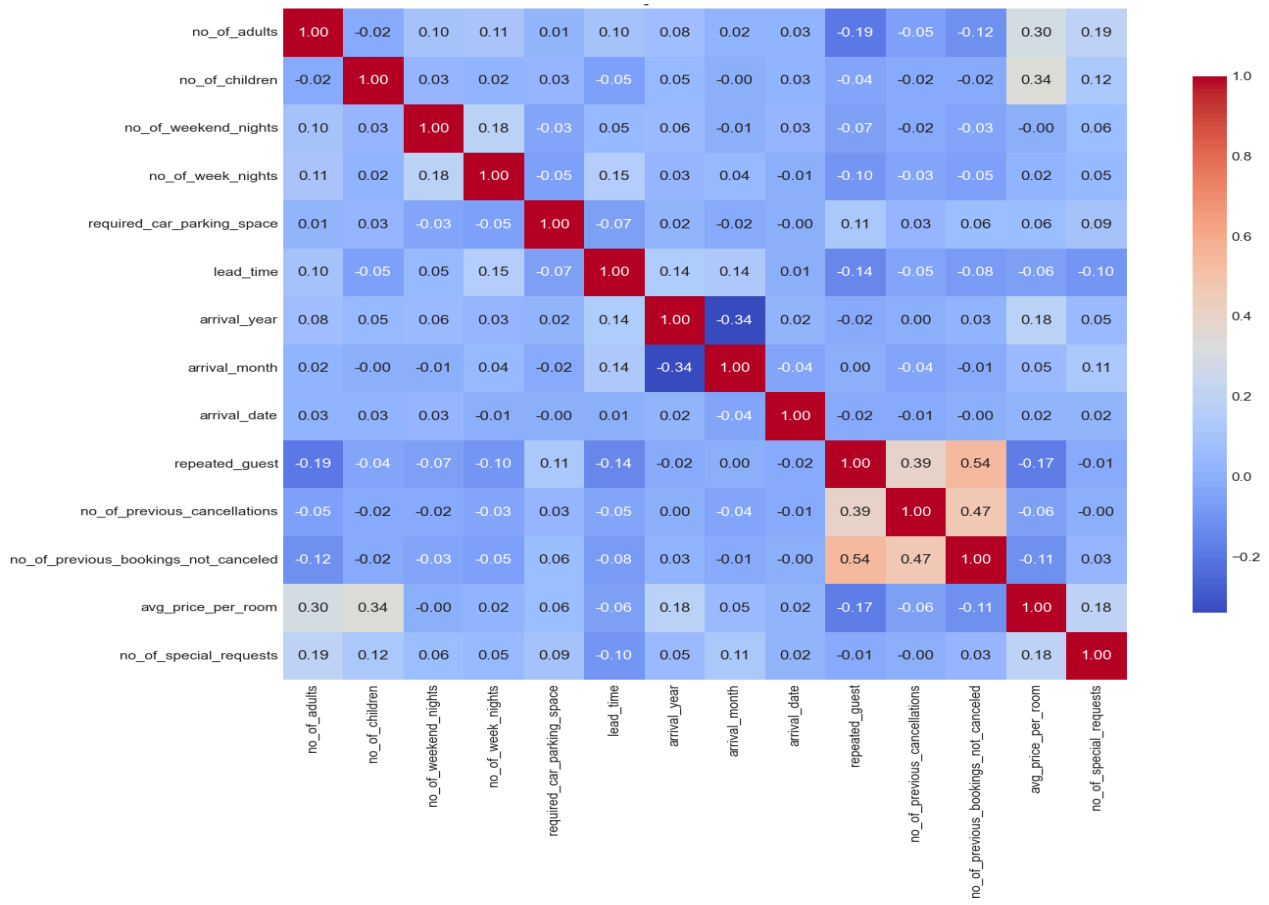
•Booking_ID: Identificador único de cada reserva
•no_of_adults: Número de adultos
•no_of_children: Número de crianças
•no_of_weekend_nights: Número de noites de fim de semana (sábado ou domingo) que o hóspede ficou ou reservou para ficar no hotel
•no_of_week_nights: Número de noites de dias úteis (segunda a sexta) que o hóspede ficou ou reservou para ficar no hotel
•type_of_meal_plan: Tipo de plano de refeição reservado pelo cliente
•required_car_parking_space: O cliente precisa de vaga de estacionamento? (0 - Não, 1 - Sim)
•room_type_reserved: Tipo de quarto reservado pelo cliente. Os valores estão codificados pelos hotéis INN
•lead_time: Número de dias entre a data da reserva e a data de chegada
•arrival_year: Ano da data de chegada
•arrival_month: Mês da data de chegada
•arrival_date: Dia do mês da chegada
•market_segment_type: Segmento de mercado designado
•repeated_guest: O cliente é um hóspede recorrente? (0 - Não, 1 - Sim)
•no_of_previous_cancellations: Número de reservas anteriores canceladas pelo cliente antes da reserva atual
•no_of_previous_bookings_not_canceled: Número de reservas anteriores não canceladas pelo cliente antes da reserva atual
•avg_price_per_room: Preço médio por dia da reserva; os preços dos quartos são dinâmicos (em euros)

•no_of_special_requests: Total de pedidos especiais feitos pelo cliente (por exemplo, andar alto, vista do quarto, etc)
•booking_status: Indicador que informa se a reserva foi cancelada ou não.

As bibliotecas pandas, matplotlib.pyplot e seaborn foram utilizadas para a manipulação e visualização dos dados. A inspeção inicial do dataset revelou a ausência de valores nulos, o que simplifica a etapa de pré-processamento. As variáveis foram analisadas individualmente e em conjunto para identificar distribuições, correlações e a relevância de cada uma para o problema de previsão de cancelamento.

A análise dos dados também envolveu a plotagem de gráficos, para melhorar a visualização do dataset. O gráfico de matriz de correlação (figura 1) entre variáveis numéricas mostra que não existe uma correlação forte entre as variáveis. A maior correlação é entre as variáveis repeated_guest e no_of_previous_bookings_not_canceled.

Figura 1 - Matriz de Correlação



Fonte: Autoria Própria

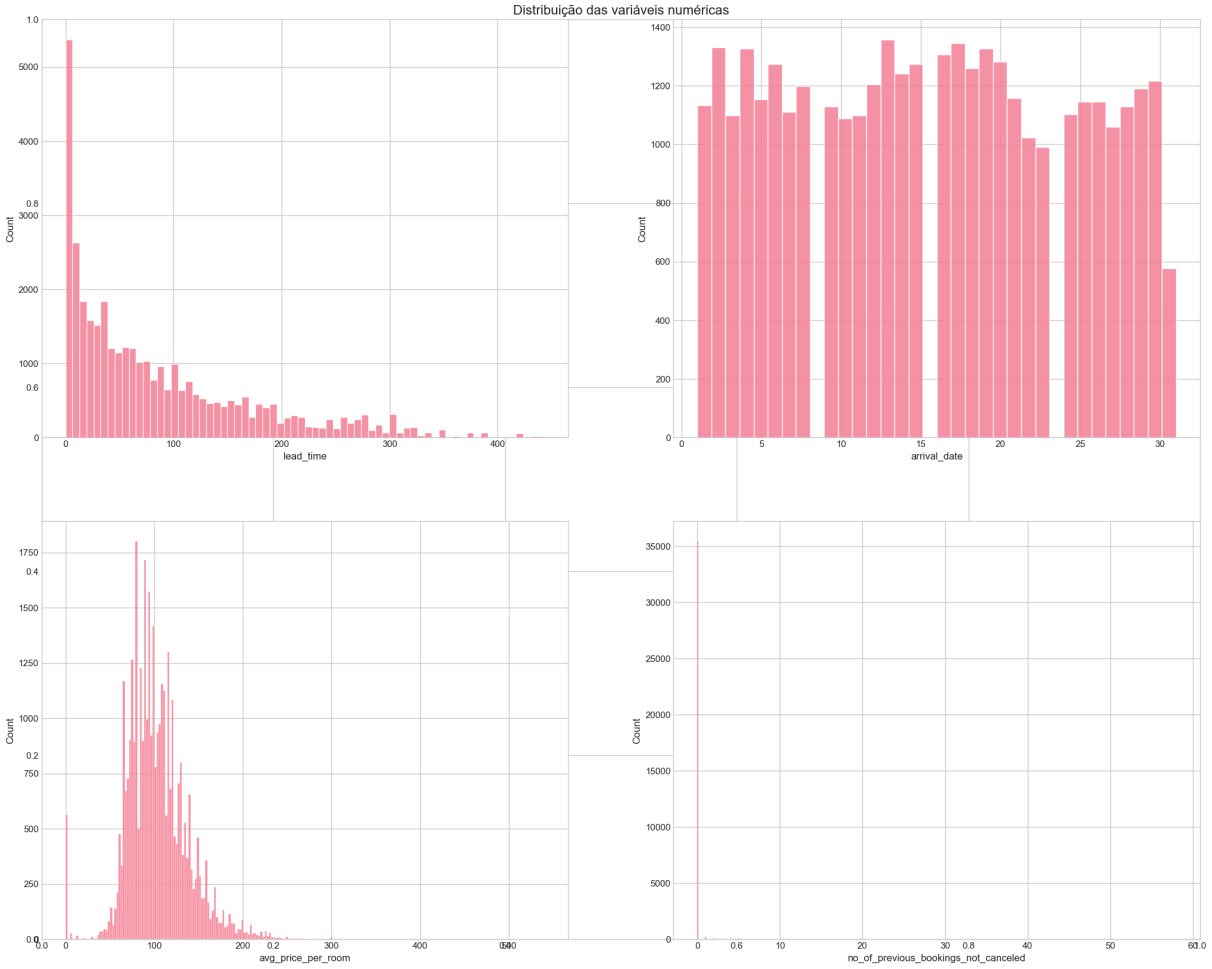
Em seguida o grid de gráficos (figura 2) envolvendo variáveis específicas, mostra pontos importantes. A maioria das reservas são para 2 adultos, 0 crianças, durante a semana, sem vaga de garagem para carros, o quarto tipo 1, em 2018, em outubro, com novos hóspedes, sem pedidos especiais e sem cancelamento prévio.

Figura 2 - Grid de Gráficos

Fonte: Autoria Própria

As variáveis numéricas também foram avaliadas, para verificação de outliers. Neste gráfico de distribuição (figura 3) a gente consegue visualizar a distribuição de valores, com alguns valores fora da curva padrão (outiliers).

Figura 3 - Gráfico de Distribuição



Fonte: Autoria Própria

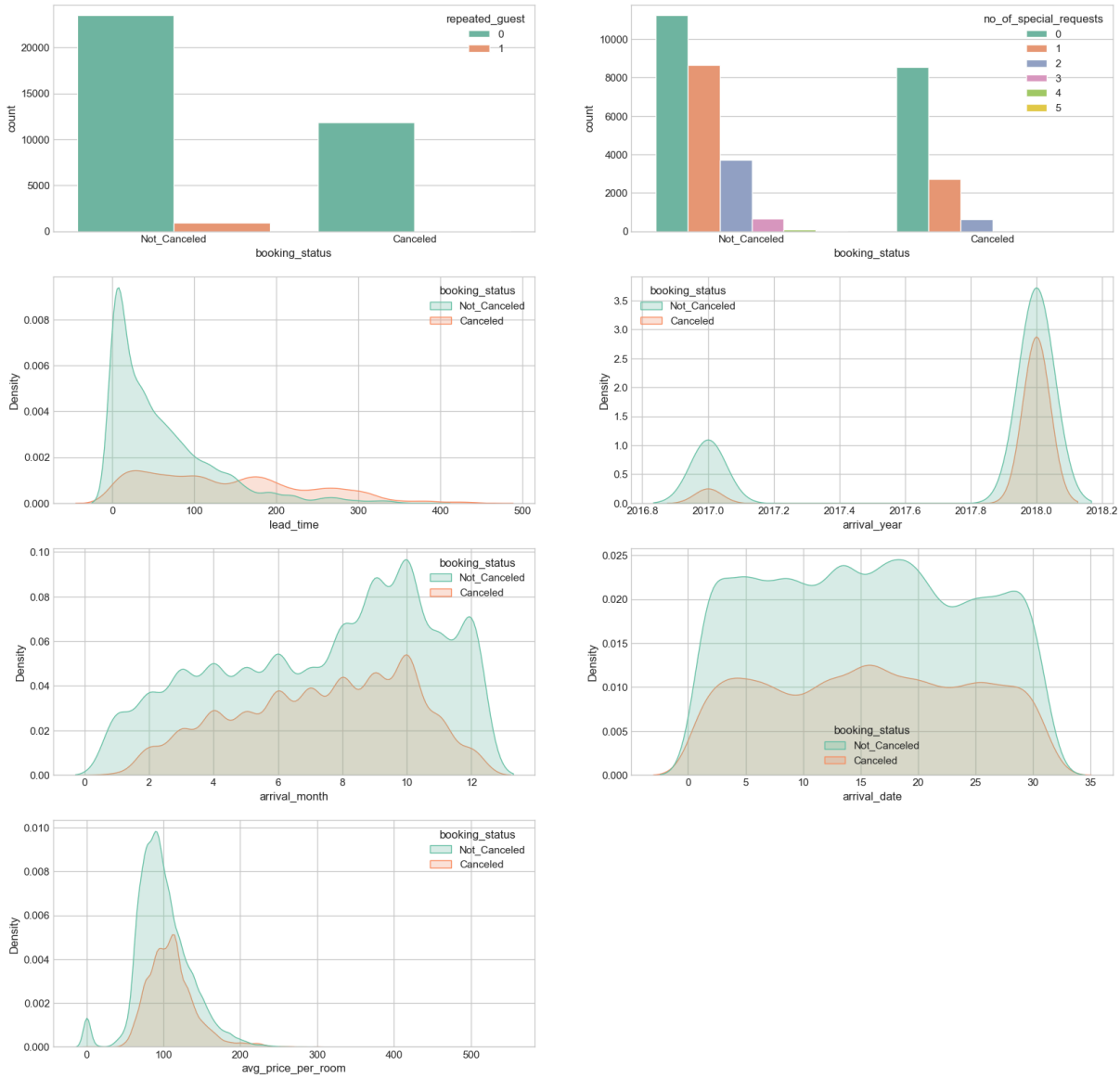
E para completar é interessante também a gente visualizar a correlação do status da reserva (figura 4 e 5) com as demais variáveis, para visualizar se alguma variável tem grande influência para o cancelamento ou não da reserva. A maioria das variáveis não possuem uma influência nítida, porém tem alguns pontos importantes. Reservas com 2 ou 9 crianças, foram 50% canceladas. Quanto mais a quantidade de finais de semana, maior a porcentagem de cancelamento. Com o plano de alimentação 2 possui uma alta taxa de cancelamento. Reservas que solicitaram vaga de estacionamento possuem menor taxa de cancelamento. Reservas com o quarto tipo 6 tiveram uma alta taxa de cancelamento.

Figura 4 - Status da Reserva

Fonte: Autoria Própria

Os gráficos criados na segunda análise (figura 5) mostram que Hóspedes antigos geralmente não cancelam a reserva. Reservas mais próximas têm uma taxa menor de cancelamentos. Reservas em 2018 tiveram uma taxa muito alta de cancelamento. Outubro tem o maior número de cancelamentos

Figura 5 - Status da Reserva 2



Fonte: Autoria Própria

2.2- Tratamento dos Dados

Antes do treinamento dos modelos, foi realizado um pré-processamento dos dados para garantir a qualidade e a adequação para os algoritmos de Machine Learning. As etapas incluíram:

1. **Label Encoding:** Variáveis categóricas como type_of_meal_plan, room_type_reserved, market_segment_type e booking_status foram convertidas em valores numéricos utilizando LabelEncoder. Esta técnica é

crucial, pois a maioria dos algoritmos de aprendizado de máquina exige entradas numéricas.

2. **Over Sampling (SMOTE):** Para lidar com o desbalanceamento de classes na variável booking_status (cancelado vs. não cancelado), foi aplicada a técnica de Over Sampling utilizando o algoritmo SMOTE (Synthetic Minority Over-sampling Technique). Isso aumentou o número de amostras da classe minoritária, equilibrando o dataset e melhorando a capacidade do modelo de aprender padrões de ambas as classes.
3. **Standard Scaler:** As variáveis numéricas foram padronizadas utilizando StandardScaler. Esta técnica transforma os dados para que tenham média zero e desvio padrão um, garantindo que todas as variáveis contribuam igualmente para o treinamento do modelo e evitando que variáveis com valores maiores dominem o processo de aprendizado.

Após o pré-processamento, o dataset foi dividido em conjuntos de treinamento e teste (over_X_train, over_X_test, over_y_train, over_y_test) para avaliação do desempenho dos modelos.

2.3- Treinamento do Modelo

Diversos modelos de classificação foram treinados e avaliados para prever o status de cancelamento das reservas. Os modelos utilizados incluem:

- AdaBoostClassifier
- BaggingClassifier
- DecisionTreeClassifier
- ExtraTreesClassifier
- GradientBoostingClassifier
- KNeighborsClassifier
- RandomForestClassifier
- XGBClassifier

O treinamento foi realizado com validação cruzada (5-fold) e as métricas de avaliação consideradas foram acurácia, F1-score e AUC-ROC. Os resultados dos modelos no conjunto de teste foram (figura 6):

Figura 6 - Resultados Modelos

model	train_accuracy	train_roc_auc	train_f1	test_accuracy	test_roc_auc	test_f1	time_taken
RandomForestClassifier	0.994651	0.999545	0.994651	0.924559	0.924559	0.925203	14.418795
ExtraTreesClassifier	0.994657	0.999922	0.994652	0.916052	0.916052	0.916198	13.444553
BaggingClassifier	0.989103	0.999167	0.989079	0.910722	0.910722	0.909956	3.608899
XGBClassifier	0.925751	0.981741	0.926044	0.906724	0.906724	0.907520	0.878652
DecisionTreeClassifier	0.994657	0.999922	0.994652	0.891554	0.891554	0.891175	0.563290
KNeighborsClassifier	0.899844	0.969980	0.899642	0.867159	0.867159	0.866226	4.592118
GradientBoostingClassifier	0.840143	0.925602	0.840736	0.841533	0.841533	0.842148	12.503818
AdaBoostClassifier	0.780334	0.874430	0.784235	0.784338	0.784338	0.787432	3.403917

Fonte: Autoria Própria

O modelo RandomForestClassifier apresentou o melhor desempenho em termos de acurácia de teste, AUC-ROC de teste e F1-Score de teste, indicando sua superioridade na previsão de cancelamentos de reservas neste dataset. Embora alguns modelos como DecisionTreeClassifier e ExtraTreesClassifier tenham apresentado alta acurácia de treino, o RandomForestClassifier demonstrou um melhor equilíbrio entre desempenho no treino e no teste, sugerindo menor overfitting. O XGBClassifier também obteve um bom desempenho, com um tempo de treinamento significativamente menor.

Buscando aprimorar ainda mais os resultados, foi aplicada a técnica de Stacking Ensemble, combinando os modelos RandomForestClassifier, ExtraTreesClassifier e DecisionTreeClassifier. Essa abordagem resultou em um desempenho ainda mais robusto, conforme demonstrado (figura 7).

Figura 7 - Resultados Stacking

	Model	Train Accuracy	Test Accuracy	Train F1	Test F1	Train ROC AUC	Test ROC AUC	Time taken for tuning (s)
0	RandomForestClassifier	0.994414	0.923842	0.994415	0.924469	0.999539	0.976687	6.981227
1	ExtraTreesClassifier	0.994286	0.918717	0.994285	0.918941	0.999806	0.974176	6.981227
2	DecisionTreeClassifier	0.962228	0.893091	0.962083	0.892174	0.995047	0.915538	6.981227

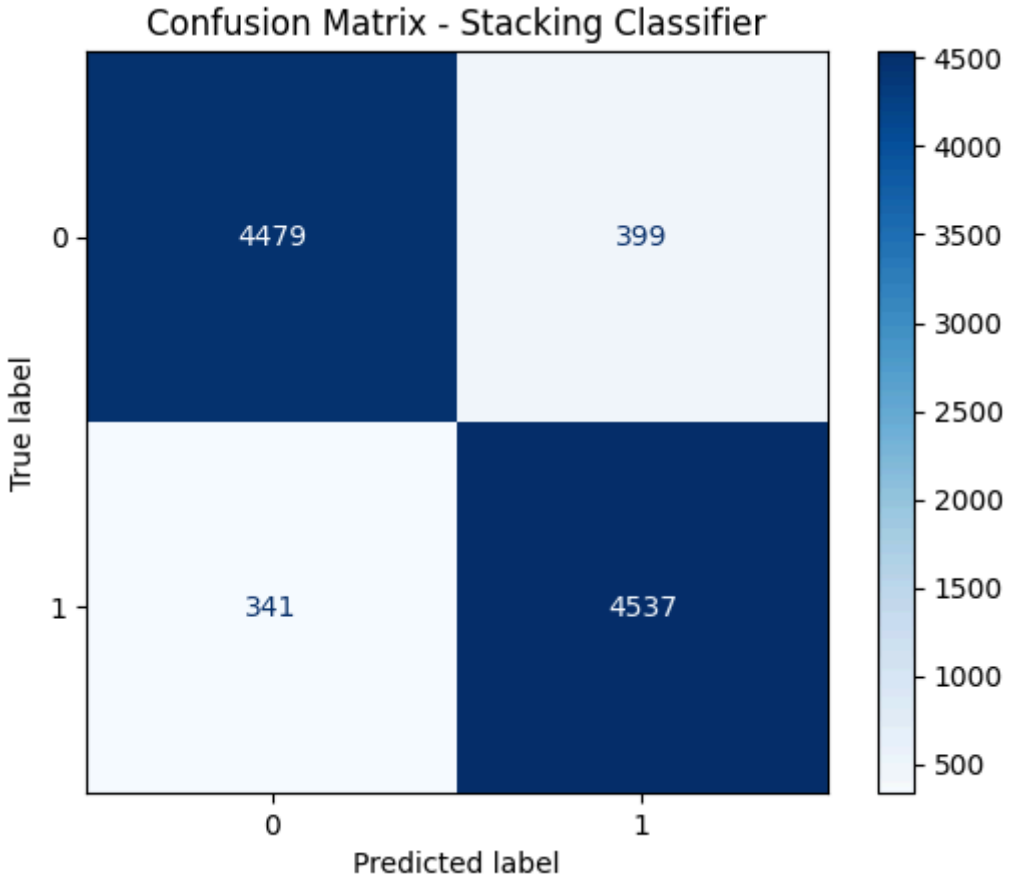
Fonte: Autoria Própria

2.4- Visualização dos Resultados

As visualizações são cruciais para entender as relações entre as variáveis e o impacto no status de cancelamento. A matriz de confusão (figura 8) demonstra que

os modelos tiveram uma alta performance, conseguindo acertar grande parte das previsões.

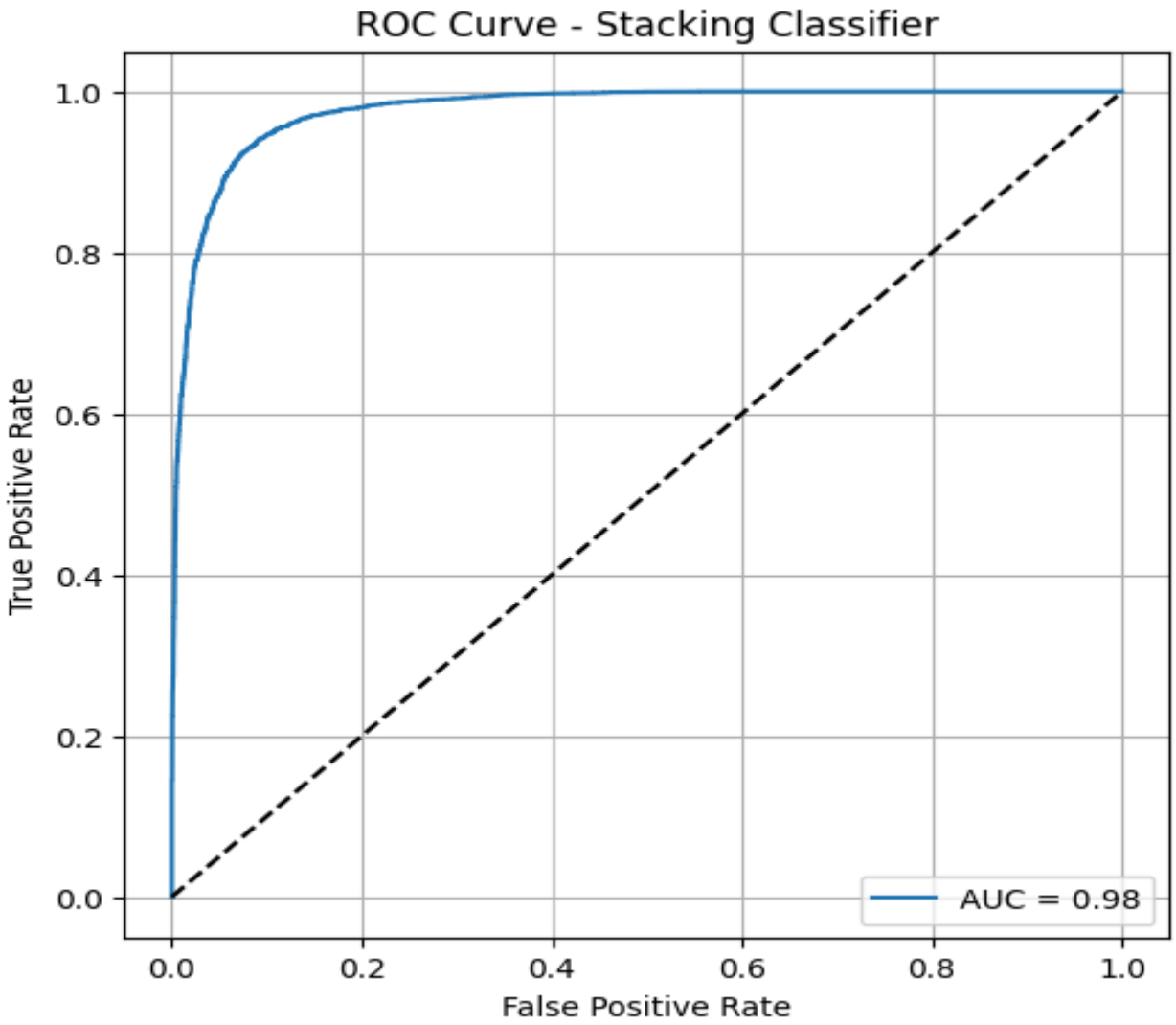
Figura 8 - Matriz de Confusão



Fonte: Autoria Própria

Além disso, a curva ROC (figura 9), demonstra que o modelo possui uma excelente capacidade discriminativa, apresentando uma área sob a curva (AUC) elevada com 98%, o que indica alta performance na diferenciação entre reservas que foram canceladas e aquelas que não foram.

Figura 9 - Curva de ROC



Fonte: Autoria Própria

3- Conclusão

O desenvolvimento deste projeto permitiu compreender de forma prática como modelos de aprendizado de máquina podem ser aplicados para resolver problemas reais do setor hoteleiro, como a previsão de cancelamento de reservas.

Por meio de uma análise detalhada dos dados, foi possível entender o comportamento dos clientes e identificar características relevantes que influenciam no cancelamento das reservas. O pré-processamento dos dados, incluindo técnicas como o balanceamento das classes com SMOTE e a padronização dos atributos, foi essencial para garantir a qualidade dos modelos. Na etapa de treinamento, diversos algoritmos foram testados, sendo que o RandomForestClassifier se destacou,

apresentando o melhor equilíbrio entre os conjuntos de treino e teste, com ótimos resultados nas métricas de acurácia, F1-score e AUC-ROC. A aplicação da técnica de Stacking, combinando os modelos RandomForestClassifier, ExtraTreesClassifier e DecisionTreeClassifier, proporcionou uma melhora significativa no desempenho, demonstrando que abordagens baseadas em ensemble podem ser extremamente eficazes em problemas de classificação com 98% de AUC. As visualizações, por meio da matriz de confusão e da curva ROC, reforçaram a alta performance dos modelos desenvolvidos, evidenciando sua capacidade de generalização e sua utilidade prática. Diante disso, conclui-se que a utilização de modelos de machine learning, aliada a uma boa análise e tratamento dos dados, é uma estratégia eficaz para auxiliar hotéis na previsão de cancelamentos, permitindo que adotem medidas preventivas e estratégias comerciais mais assertivas. Este trabalho também proporcionou um grande aprendizado prático sobre todo o ciclo de desenvolvimento de um projeto de ciência de dados, desde a análise exploratória até a avaliação dos modelos preditivos.