

Reduzindo prejuízos financeiros em hotéis: uma abordagem de análise de dados para prever cancelamentos de reservas

Gilson Francisco de Oliveira Castro¹

¹Instituto de Computação – Universidade Federal de Alagoas (UFAL)

{gilson.castro}@feac.ufal.br

Abstract. *This article presents a data analysis approach to predict reservation cancellations and reduce financial losses in hotels. The research objective was to develop machine learning models capable of accurately predicting whether a reservation would be canceled or not. Techniques such as exploratory data analysis, feature engineering, and relevant variable selection were utilized. The results show that the random forest technique had the best performance in predicting reservation cancellations, followed by decision tree, logistic regression, and CatBoost techniques. With this data analysis approach, hotels can identify reservations with a higher likelihood of being canceled in advance and adopt strategies to reduce the financial losses resulting from these cancellations.*

Resumo. *Este artigo apresenta uma abordagem de análise de dados para prever cancelamentos de reservas e reduzir os prejuízos financeiros dos hotéis. O objetivo da pesquisa foi desenvolver modelos de aprendizado de máquina capazes de prever com precisão se uma reserva seria cancelada ou não. Para isso, foram utilizadas técnicas de análise exploratória de dados, engenharia de recursos e seleção de variáveis relevantes. Os resultados mostram que a técnica de floresta aleatória obteve o melhor desempenho na previsão de cancelamentos de reservas, seguida pelas técnicas de árvore de decisão, regressão logística e CatBoost. Com essa abordagem de análise de dados, os hotéis podem identificar antecipadamente as reservas com maior probabilidade de serem canceladas e adotar estratégias para reduzir os prejuízos financeiros decorrentes desses cancelamentos.*

1 . Introdução

A indústria hoteleira enfrenta constantemente o desafio de prever a demanda e garantir que os quartos estejam sempre ocupados. Um dos principais obstáculos nesse processo é a imprevisibilidade dos cancelamentos de reservas, que podem afetar significativamente a receita dos hotéis e causar prejuízos financeiros. Para minimizar esses riscos, a análise de dados e machine learning têm sido cada vez mais utilizados por empresas do setor hoteleiro para prever as chances de um cliente cancelar sua reserva com antecedência e tomar medidas preventivas para reduzir os prejuízos financeiros.

Este estudo tem como objetivo desenvolver um modelo preditivo capaz de prever com precisão as chances de um cliente cancelar sua reserva em um hotel e identificar quais fatores influenciam essa decisão. A metodologia incluiu a coleta de dados sobre reservas e clientes, a preparação dos dados para análise, o uso de técnicas de análise de dados e machine learning para desenvolver o modelo preditivo e a validação do modelo por meio de testes e comparações com dados reais de cancelamentos de reservas em hotéis. Os resultados indicam que o modelo proposto apresenta uma alta precisão na previsão de cancelamentos de reservas em hotéis e sugere que fatores como preço, datas da reserva e localização do hotel são importantes determinantes na tomada de decisão do cliente.

Este estudo tem como objetivo principal desenvolver um modelos preditivos capaz de prever com precisão as chances de um cliente cancelar sua reserva em um hotel. Para isso, serão utilizadas técnicas de análise de dados e machine learning para analisar dados de reservas e clientes, identificar padrões e fatores relevantes e desenvolver um modelo que possa prever com precisão os cancelamentos de reservas em hotéis.

2. Problema de negócio

A demanda hoteleira é gerada por uma variedade de viajantes, incluindo aqueles em viagens de negócios e turistas em busca de lazer. Com base nas informações das reservas de hotel, nosso objetivo é prever se um cliente cancelará sua reserva ou não. Dessa forma, podemos responder duas perguntas importantes para os negócios do hotel:

1. Existe a possibilidade de um cliente cancelar sua reserva? Essa é uma questão crucial para que o hotel possa se preparar adequadamente para acomodar reservas futuras e evitar perdas financeiras devido a quartos vazios.
2. Se houver cancelamento, quais são os fatores determinantes que levam os clientes a cancelarem suas reservas? Compreender esses fatores pode ajudar o hotel a ajustar suas políticas de cancelamento e estratégias de preços para minimizar as taxas de cancelamento e maximizar a receita.



3. Importância do problema

A previsão da demanda hoteleira é um aspecto crucial da gestão eficaz de um hotel. Ela permite que os gerentes aloquem adequadamente a equipe para cobrir todas as áreas essenciais do hotel, como a recepção, o restaurante e a limpeza, permite também ter uma compreensão clara da demanda esperada ajuda os gerentes a estimarem a demanda dos negócios subsidiários, tais como lojas de presentes, academias e espaços localizados dentro do hotel.

Com a popularização das políticas de cancelamento flexíveis oferecidas pelos serviços online, tornou-se cada vez mais importante estimar as taxas de cancelamento. Isso evita que o hotel fique com quartos vazios que precisam ser vendidos a preços mais baixos para evitar perdas financeiras. Além disso, ter uma compreensão precisa da taxa de cancelamento esperada pode ajudar a determinar a taxa de overbooking ideal do hotel, levando em consideração a sazonalidade. Ao prever a probabilidade de cancelamento, é possível alimentar uma análise de demanda mais completa do hotel para o período de tempo considerado.

4. Descrição dos dados

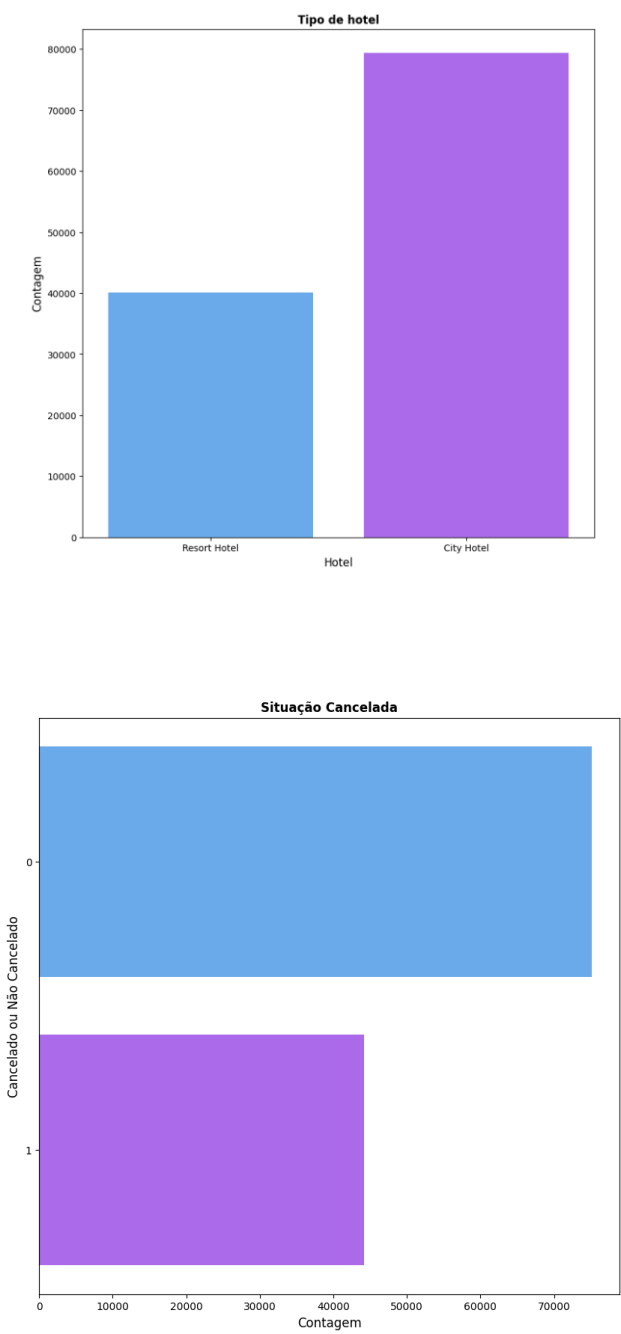
A base de dados "hotel booking demand" disponível no Kaggle contém informações sobre reservas de hotéis em Portugal, incluindo dados demográficos dos clientes, informações sobre a estadia no hotel e se a reserva foi cancelada ou não. A base de dados foi coletada entre 2015 e 2017 e contém 119.390 entradas.

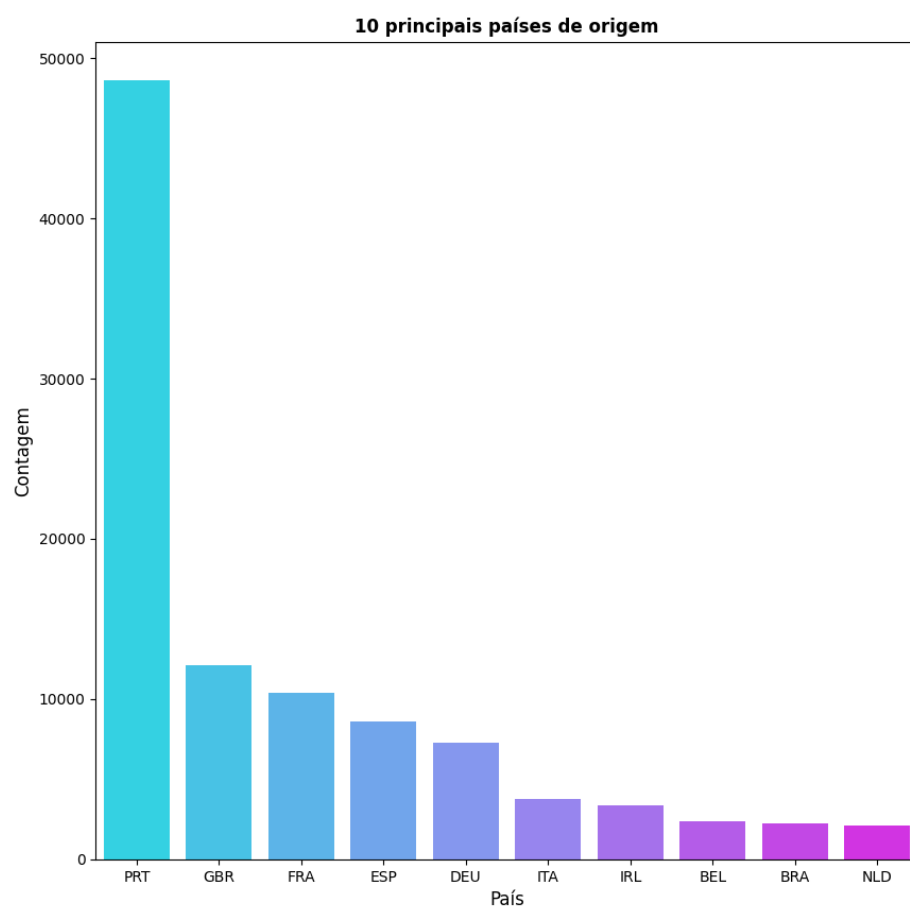
 dados_hotel.describe().T									
		count	mean	std	min	25%	50%	75%	max
is_canceled	119390.0	0.370416	0.482918	0.00	0.00	0.000	1.0	1.0	
lead_time	119390.0	104.011416	106.863097	0.00	18.00	69.000	160.0	737.0	
arrival_date_year	119390.0	2016.156554	0.707476	2015.00	2016.00	2016.000	2017.0	2017.0	
arrival_date_week_number	119390.0	27.165173	13.605138	1.00	16.00	28.000	38.0	53.0	
arrival_date_day_of_month	119390.0	15.798241	8.780829	1.00	8.00	16.000	23.0	31.0	
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.00	0.00	1.000	2.0	19.0	
stays_in_week_nights	119390.0	2.500302	1.908286	0.00	1.00	2.000	3.0	50.0	
adults	119390.0	1.856403	0.579261	0.00	2.00	2.000	2.0	55.0	
children	119386.0	0.103890	0.398561	0.00	0.00	0.000	0.0	10.0	
babies	119390.0	0.007949	0.097436	0.00	0.00	0.000	0.0	10.0	
is_repeated_guest	119390.0	0.031912	0.175767	0.00	0.00	0.000	0.0	1.0	
previous_cancellations	119390.0	0.087118	0.844336	0.00	0.00	0.000	0.0	26.0	
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.00	0.00	0.000	0.0	72.0	
booking_changes	119390.0	0.221124	0.652306	0.00	0.00	0.000	0.0	21.0	
agent	103050.0	86.693382	110.774548	1.00	9.00	14.000	229.0	535.0	
company	6797.0	189.266735	131.655015	6.00	62.00	179.000	270.0	543.0	
days_in_waiting_list	119390.0	2.321149	17.594721	0.00	0.00	0.000	0.0	391.0	
adr	119390.0	101.831122	50.535790	-8.38	69.29	94.575	126.0	5400.0	
required_car_parking_spaces	119390.0	0.062518	0.245291	0.00	0.00	0.000	0.0	8.0	
total_of_special_requests	119390.0	0.571363	0.792798	0.00	0.00	0.000	1.0	5.0	

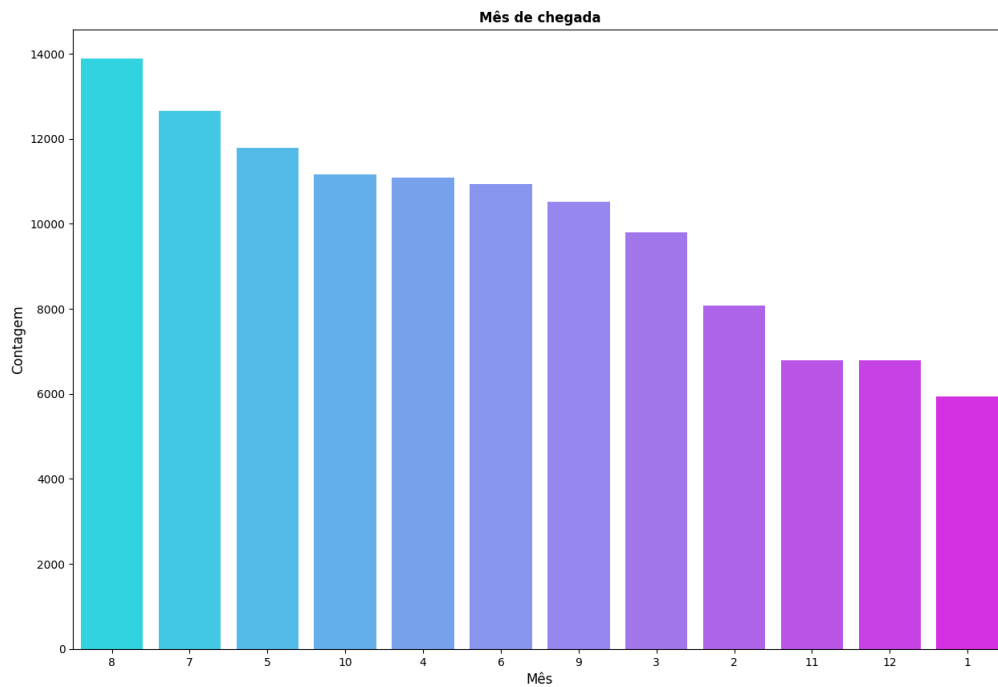
Na figura acima temos um sumário estatístico de todas as colunas do conjunto de dados desse modo, vemos medidas de tendência central e dispersão dos dados.

Abordagem e Metodologia

Foi realizada uma visualização dos dados utilizando para compreender a relação entre as variáveis preditoras. Isso nos ajudou a desenvolver hipóteses sobre se as características poderiam ser fatores influenciadores para o cancelamento.







5.2 Pré-processamento do dados

Enfrentamos desafios com valores ausentes, multicolinearidade e desequilíbrio de classes em nosso conjunto de dados. Para abordá-los, eliminamos variáveis altamente correlacionadas, descartamos variáveis com alta porcentagem de valores ausentes e agrupamos valores ausentes na coluna do país em uma nova categoria. Também imputamos o recurso de crianças com o modo dos valores. Além disso, projetamos novos recursos, como a data de chegada, o dia da semana de chegada, o sinalizador de família, o sinalizador de mudanças no tipo de quarto, o total de noites hospedadas e o sinalizador de não reembolso. Usamos a codificação one-hot para variáveis categóricas e lidamos com o desequilíbrio de classes usando downsampling. Seleccionamos o melhor conjunto de recursos usando a Eliminação Recursiva de Atributos (RFE) com um modelo de floresta aleatória e validação cruzada de 3 dobras. O resultado foi 43 recursos selecionados em um conjunto de dados de 61.914 linhas.

5.4 Construção de modelos

O objetivo é prever a probabilidade de cancelamentos, tratamos isso como um problema de classificação e escolhemos modelos de acordo. Construímos os seguintes modelos pelos seguintes motivos:

- Regressão Logística – A função logarítmica limita o valor entre 0 e 1, tornando-o adequado para um problema de classificação. É um modelo simples com uma equação interpretável e os coeficientes nos dão a direção da influência na variável dependente.
- Árvore de decisão - Este modelo de árvore é muito explicável. O passo de precaução é que as previsões às vezes são grosseiras, pois uma previsão é feita para uma subdivisão inteira do espaço de recursos e também é suscetível a sobreajuste.
- Floresta aleatória - Este modelo de árvore fornece resultados mais precisos, pois é um conjunto de muitos modelos individuais. A seleção aleatória de recursos para construir cada modelo torna as árvores construídas menos correlacionadas e isso pode melhorar os resultados. Também fornece os recursos importantes usando ganho de informação.
- CatBoost - Este modelo de árvore funciona melhor com informações categóricas. Como nosso conjunto de dados tinha muitos recursos categóricos, este modelo foi criado para melhorar os resultados. Este modelo também fornece os recursos importantes.

4.5 Escolha da métrica de avaliação

A métrica adotada para avaliar o desempenho do modelo foi o Recall, tal métrica é calculada através da fórmula:

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

Fonte: Autor(2023)

Onde verdadeiro Positivo é quando o modelo prevê corretamente a classe positiva real (no caso, a classe de cancelamento) e falso negativo é quando o modelo classifica erroneamente uma instância como negativa, sendo que ela deveria ser positiva.

Considerando a possibilidade de prejuízos para os proprietários de hotéis em decorrência de excesso de reservas, foi priorizado um alto valor de Recall na avaliação do modelo. Isso porque o custo de prever incorretamente um cancelamento é muito maior do que o de prever erroneamente que uma reserva será cancelada. Entretanto, também foi levada em conta a precisão do modelo para garantir a qualidade do serviço prestado aos clientes.

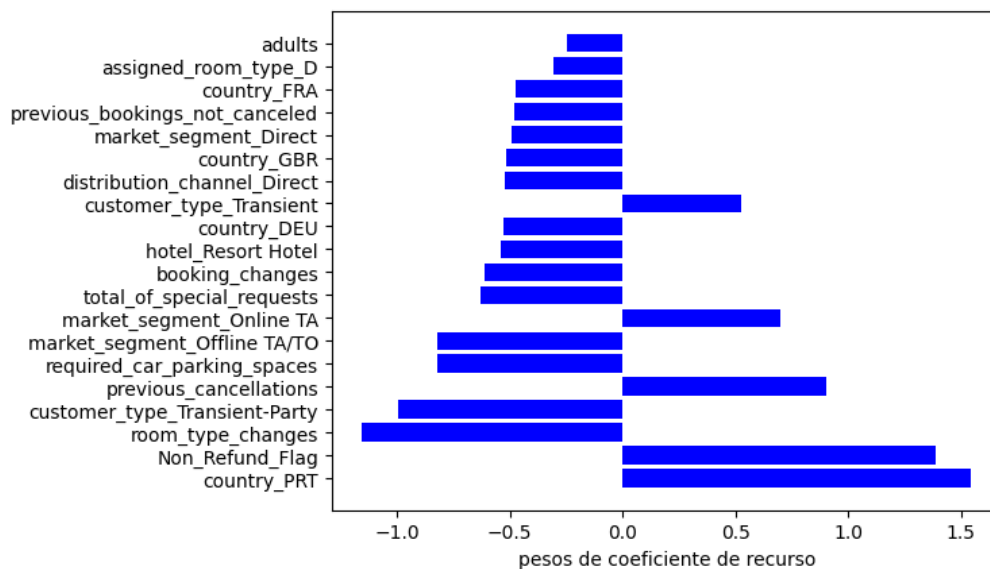
6. Resultados

6.1 Regressão Logística

Para que este modelo tenha um bom desempenho, é necessário que as variáveis independentes não apresentem multicolinearidade ou sejam altamente correlacionadas. Além disso, assume-se a linearidade das variáveis independentes e das chances logarítmicas.

Por meio da regularização L2, a melhor versão deste modelo obteve um recall de 0,86, que foi usada como referência para comparar o desempenho de outros modelos.

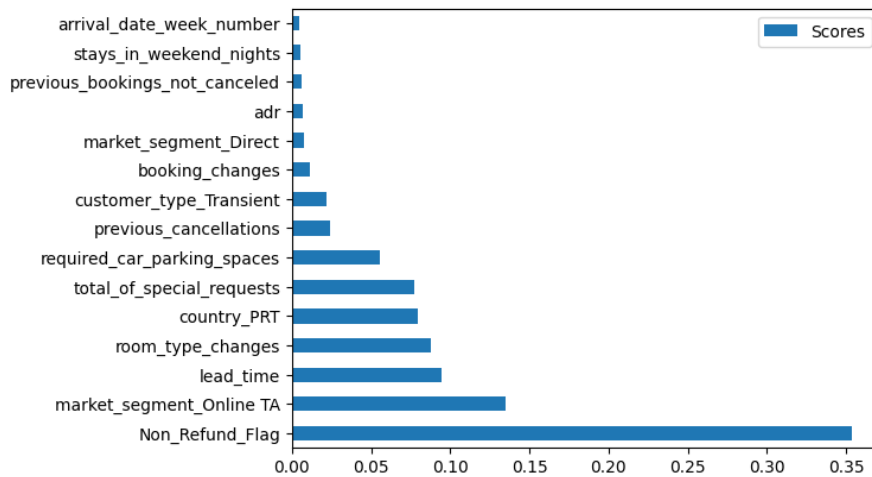
Para entender o impacto das características na rotulagem, foram analisados os coeficientes do modelo de regressão logística. O gráfico abaixo mostra as vinte características mais importantes, tanto positivas quanto negativas. Isso indica que se não houver reembolso, a probabilidade de cancelamento é alta.



6.2 Decision Tree

Usamos o algoritmo Decision Tree para classificar casos de cancelamento e não cancelamento. O Decision Tree usa particionamento recursivo para dividir o espaço de recursos com base em limites de decisão e identificar recursos importantes usando ganho de informação do recurso. Tanto dados categóricos quanto numéricos podem ser tratados pelo Decision Tree e, para esses dados, uma árvore com profundidade máxima de 10 nos deu uma taxa de recall de 0,93.

A importância dos recursos determinou que as reservas têm maior chance de serem canceladas se forem não reembolsáveis, se a reserva for feita por um Agente de Viagens online e se o tempo de espera for alto, ou seja, se a reserva for feita com muita antecedência.

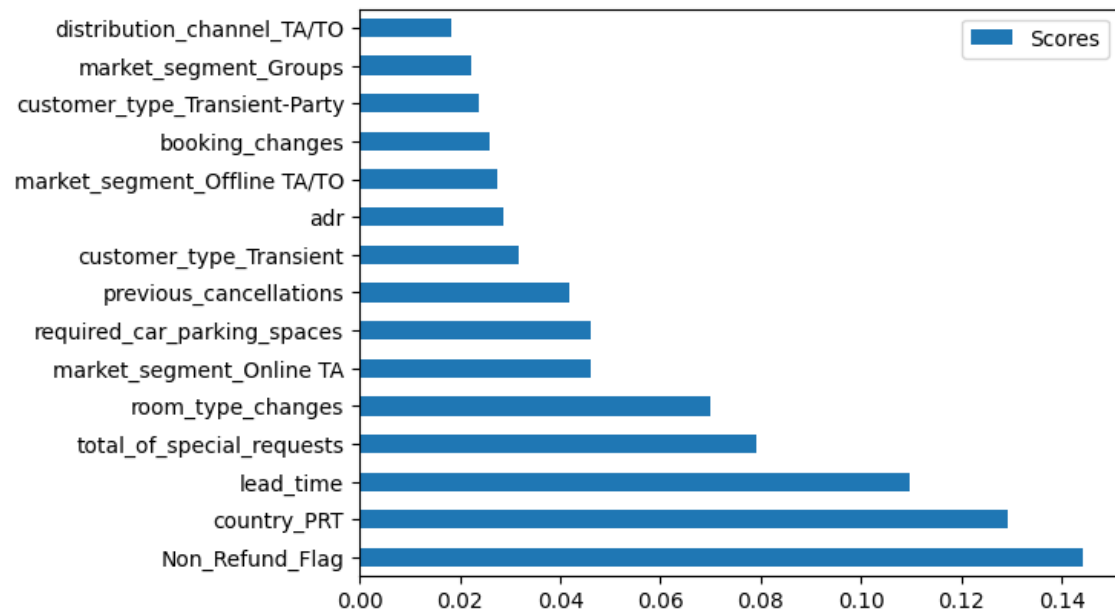


6.3 Random Forest

Random Forest é um algoritmo baseado em árvores de decisão que cria várias árvores a partir de subconjuntos aleatórios de dados do conjunto de treinamento. Essas árvores são agregadas com base nos votos das diferentes árvores para selecionar a melhor. O modelo padrão não ajustado apresentou uma pontuação de recall de 0,99 no conjunto de treinamento e 0,91 no conjunto de validação, mostrando claramente sobreajuste.

Após ajustar o modelo usando GridsearchCV com os melhores parâmetros de profundidade 13, número de árvores como 500 e amostra mínima dividida por 2, obteve-se um valor de recall de 0,93, que corrigiu o overfitting e melhorou os resultados.

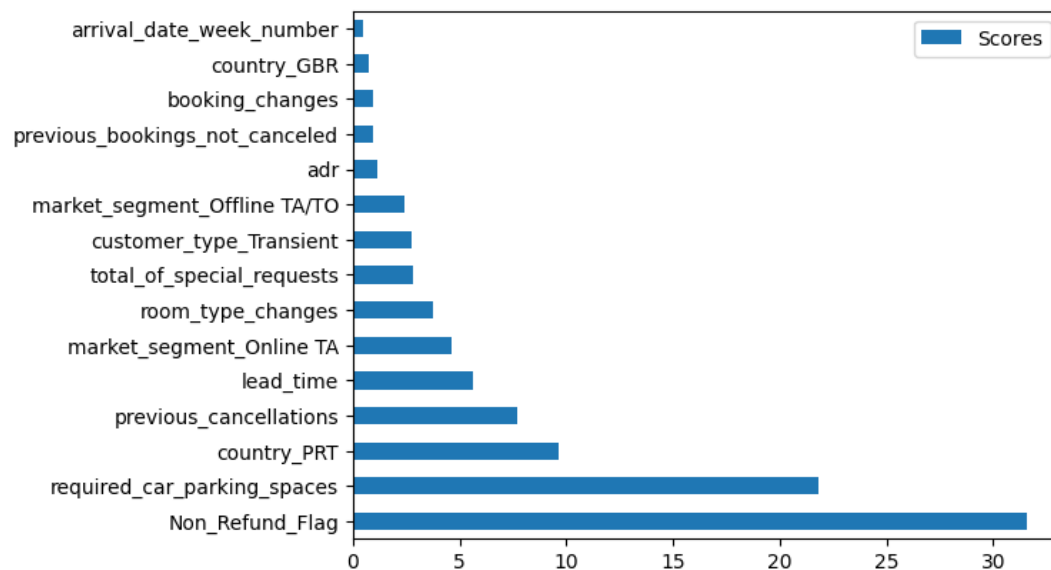
A importância das características derivadas desse modelo determinou que as reservas têm maior probabilidade de serem canceladas em Portugal quando o tipo de depósito é não-reembolsável e se o tempo de espera é longo. Essas características são semelhantes às derivadas usando árvores de decisão.



6.4 Catboost

O CatBoost significa "Category Boosting" e é um algoritmo de conjunto que usa o princípio do Gradient Boosting. O modelo pode trabalhar com diferentes tipos de dados enquanto oferece um desempenho de classe mundial. Duas das principais vantagens do pacote CatBoost são sua capacidade de lidar com tipos de dados, como variáveis categóricas, sem a necessidade de treinamento extensivo dos dados.

As características importantes são novamente semelhantes às da Árvore de Decisão e Random Forest. O gráfico abaixo mostra as características importantes.



7 . Comparação dos modelos selecionados

Tabela 1: resultado dos scores

Modelo	Precision	Recall	F1-Score
Logistic Regression	0.61	0.86	0.72
Decision Tree	0.63	0.93	0.75
<u>Random Forest</u>	<u>0.66</u>	<u>0.93</u>	<u>0.77</u>
CatBoost	0.68	0.90	0.78

Fonte: Autor (2023)

Ao comparar as pontuações dos vários modelos no conjunto de validação, podemos observar que tanto a Árvore de Decisão quanto o Random Forest obtiveram a maior pontuação de recall, com um valor de 0,93. No entanto, o Random Forest também apresentou uma pontuação mais alta de precisão. Sendo assim, selecionamos o modelo Random Forest como base para nossas recomendações.

8 . Conclusão

Com base na análise e no treinamento dos modelos, concluímos que o modelo Random Forest é o mais adequado para a classificação de cancelamentos de reservas de hotéis. Além disso, nossas descobertas sugerem que reservas não reembolsáveis, reservas feitas através de agentes de viagens online e um tempo de antecedência maior são os fatores mais importantes na predição de cancelamentos. Com essas informações, as empresas de hotelaria podem tomar medidas para minimizar a taxa de cancelamentos e aumentar a satisfação do cliente.

Uma aplicação futura é a integração com sistemas de gerenciamento de reservas, permitindo que os hotéis monitorem automaticamente a probabilidade de cancelamento de uma reserva e tomem as medidas adequadas para minimizar os prejuízos financeiros. Isso pode incluir a oferta de incentivos aos clientes para que mantenham a reserva, ou o ajuste da política de cancelamento de acordo com as previsões de demanda.

Apesar dos resultados promissores obtidos com a abordagem, é importante considerar algumas limitações. A qualidade dos dados utilizados pode afetar a precisão das previsões e a abordagem baseia-se em modelos estatísticos que não levam em conta fatores imprevisíveis. Além disso, a aplicação da abordagem pode exigir um investimento significativo em termos de tempo, recursos e expertise técnica.

8 . Referências

ANTÓNIO, Nuno. Predictive models of hotel booking cancellation: a semi-automated analysis of the literature. **Tourism & Management Studies**, v. 15, n. 1, p. 7-21, 2019.

CHEN, Yiyi et al. Comparison and analysis of machine learning models to predict hotel booking cancellation. In: **2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)**. Atlantis Press, 2022. p. 1363-1370.

JUBA, Brendan; LE, Hai S. Precision-recall versus accuracy and the role of large data sets. In: **Proceedings of the AAAI conference on artificial intelligence**. 2019. p. 4039-4048.

GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Alta Books, 2019.