Nota Conceitual e Plano de Implementação do Projeto Capstone

Título do projeto: Previsão de Cancelamento de Reservas em Hotel e Resort usando Machine Learning

Membros da Equipe:

Júlio César

Alberto Pereira

Serena Lopes

Nota Conceitual

1. Visão Geral do Projeto

Este projeto final visa construir um modelo de Machine Learning capaz de **prever a probabilidade de cancelamento de uma reserva** em hotel ou resort. A solução é crucial para permitir que a gestão hoteleira **antecipe a demanda e o fluxo de visitantes** com antecedência , otimizando o planejamento de recursos.

Relevância para os ODS: O projeto se alinha aos Objetivos de Desenvolvimento Sustentável (ODS) ao permitir o **planejamento de recursos com precisão**, evitando o excesso de uso ou a criação de infraestrutura desnecessária, promovendo assim o uso mais sustentável dos recursos.

Problema e Impacto Potencial: O problema principal é a perda de receita e os custos operacionais associados a cancelamentos de última hora. O impacto potencial da solução é **minimizar as perdas de receita** ao permitir intervenções proativas da equipe (e-mails, chamadas, ofertas de retenção) para "salvar" o máximo de reservas reais possível, mesmo que isso resulte em mais alarmes falsos.

2. Objetivos

O objetivo principal do projeto é desenvolver um modelo de classificação robusto que cumpra as métricas de negócio estabelecidas:

> **Minimizar Reservas Canceladas Falsas (Precision):** Quando o modelo prever que a reserva será cancelada, ele deve acertar na maioria das vezes para otimizar os recursos da equipe.
>
> **Capturar o Máximo de Cancelamentos Reais (Recall):** Identificar a maioria dos clientes que realmente irão cancelar (Positivos Reais), o que é a métrica mais crucial para evitar perdas de receita.

O projeto alcançou um **Recall de 0.67** , indicando que ele é eficaz em capturar a maioria dos cancelamentos reais, o que contribui diretamente para resolver o problema identificado, dando à equipe operacional a chance de reter o cliente.

3. Histórico

O problema da previsão de cancelamento de reservas é complexo, pois fatores como o *lead time* (tempo de antecedência da reserva) influenciam diretamente. O projeto descobriu que, para reservas com um *lead time* superior a 113 dias, as chances de cancelamento estão quase nas mesmas proporções que as chances de não-cancelamento, necessitando de uma análise preditiva mais apurada.

A abordagem de **Aprendizado de Máquina (ML)** é benéfica e necessária porque permite que o modelo:

Extraia padrões complexos e não lineares nos dados.

Lide com a natureza empírica da otimização de parâmetros.

Faça previsões mais próximas da realidade, com o mínimo de erro, com base no conhecimento adquirido no conjunto de treinamento.

4. Metodologia

O projeto seguiu uma metodologia de experimentação com foco no refinamento do modelo e seleção de features:

Aspeto	**Técnica Aplicada**	**Detalhes**
Modelos Testados	K-Neighbors (KNN), XGBoost, LogisticRegression	O **KNN** foi o modelo escolhido por apresentar o melhor **F1-Score** de 0.716 (média harmónica de Precision e Recall)1515151515.
Ajuste de Hiperparâmetros	**GridSearchCV** e Validação Cruzada	Utilizado para capturar os melhores hiperparâmetros. Para o KNN, os parâmetros ótimos encontrados foram: {'n_neighbors': 11, 'weights': 'distance'}16161616.
Seleção de Features	Seleção **Passo a Passo (Stepwise Selection)**	Usada para colunas numéricas, sendo menos robusta, mas permitindo que features menos relevantes ajudem o modelo a generalizar melhor17171717.
Seleção de Features	**Filtro Estatístico Bivariado**	Aplicado em colunas categóricas, excluindo aquelas com um p-valor

Aspeto	Técnica Aplicada	Detalhes
Categóricas		≥ 0.05 em relação à variável alvo18181818.
Validação	**Cross-Validation** com $cv=3$	Utilizado cross_val_score(cv=3) por ter obtido um desempenho superior, garantindo dados suficientes em cada grupo para extrair padrões19191919.

5. Diagrama de Projeto de Arquitetura

A arquitetura do projeto segue um fluxo padrão de Machine Learning que abrange desde a aquisição dos dados até a implantação em produção.

Componentes Chave e Funcionalidades:

Fonte de Dados (Kaggle): Os dados são extraídos em formato .csv.

Pré-Processamento (Python/Scikit-learn): Nesta etapa são realizados:

Limpeza de Dados: Remoção de colunas irrelevantes ("agent", "company").

Engenharia de Recursos: Seleção de features (Passo a Passo e Filtro Estatístico Bivariado).

Transformação: Uso de **LabelEncoder** para variáveis categóricas e **Transformação Logarítmica** para variáveis com distribuição exponencial (como 'adr').

Normalização: Aplicação do **StandardScaler** para padronizar valores.

Treinamento do Modelo (K-Neighbors): O modelo é treinado nos dados processados, com otimização dos hiperparâmetros via **GridSearchCV**.

Serialização (Joblib): O modelo final treinado é salvo (**dump**) no formato .joblib para garantir salvamento e carregamento rápidos, especialmente para grandes *arrays* de dados Numpy, utilizando mapeamento de memória.

Implantação (Streamlit Community): O modelo serializado é carregado no ambiente de produção do Streamlit Community (que fornece hospedagem gratuita e interface).

Serviço de Predição: A aplicação recebe novas *features* de entrada através da interface e chama a função predict() do modelo carregado para fazer as novas previsões.

6. Fontes de Dados

Os dados foram extraídos diretamente da plataforma **Kaggle**. São dados do tipo texto com extensão **.csv**.

Relevância: O conjunto de dados contém informações sobre reservas de hotel/resort, permitindo a identificação de padrões correlacionados à variável alvo (cancelamento). A Análise Exploratória de Dados (EDA) revelou que a variável arrival_date_day_of_month possui uma distribuição normal, enquanto a maioria das outras variáveis possui uma distribuição exponencial (ex: 'adr').

Etapas de Pré-processamento:

Colunas irrelevantes ("agent", "company", "country") foram descartadas.

Colunas categóricas foram tratadas com **LabelEncoder** para transformação numérica.

Variáveis com distribuição exponencial foram submetidas à **transformação logarítmica** para normalização.

7. Revisão de Literatura

A metodologia do projeto é apoiada por pesquisas existentes na área de Machine Learning aplicada à gestão hoteleira. A pesquisa inicial incluiu:

Uso de AutoML para previsão de demanda e cancelamento de hospedagem: Um estudo que demonstra a previsão de demanda de hotel para antecipar o desempenho futuro das principais métricas do hotel/resort.

Previsão da demanda turística com base em um modelo de rede neural temporal híbrido: Apresenta um novo modelo híbrido (BiLSTM e Transformer) para previsão de demanda turística sustentável.

Este projeto se baseia nesse trabalho ao focar na previsão de cancelamento, mas utiliza uma abordagem inicial mais focada em modelos clássicos (KNN, XGBoost) e técnicas robustas de seleção de features (Passo a Passo) para obter um equilíbrio entre interpretabilidade e performance

Plano de Implementação

1. Pilha de Tecnologia

A implementação do projeto utiliza a seguinte pilha de tecnologia:

Componente	Tecnologia / Ferramenta	Uso
Modelos de ML	KNeighborsClassifier (KNN), XGBoost, LogisticRegression	Algoritmos testados41.
Frameworks/Bibliotecas ML	Scikit-learn, joblib, Numpy	GridSearchCV e cross_val_score para treinamento/avaliação4242424 2. joblib para serialização e carregamento eficiente do modelo43.
Normalização de Dados	**StandardScaler**	Para padronizar os valores das *features* no mesmo intervalo444444.
Implantação	**Streamlit Community**	Escolhido para hospedagem gratuita e criação de uma interface interativa para as previsões45.
API (Futuro)	Flask ou FastAPI	Considerado para disponibilização do modelo como um serviço de API46.

2. Linha do Tempo e 3. Marcos

Embora um gráfico de Gantt detalhado não tenha sido fornecido, o projeto foi estruturado nas seguintes fases com marcos claros:

Fase	Tarefas Gerenciáveis	Marco de Conclusão
Fase 1: Preparação de Dados e Engenharia de Recursos	Coleta, Limpeza (drop de colunas), Transformação (LabelEncoder/Logarítmica), Normalização (StandardScaler), Seleção de Features (Passo a Passo, Filtro Bivariado)47474747474747.	Conjunto de dados final processado e pronto para o treinamento48.
Fase 2: Desenvolvimento, Treinamento e	Teste de Modelos (KNN, XGBoost, LogReg)49. Otimização de Hiperparâmetros	Seleção do Modelo Vencedor (KNN) com melhor

Fase	Tarefas Gerenciáveis	Marco de Conclusão
Avaliação	(GridSearchCV)50.	F1-Score (0.716)51515151.
Fase 3: Refinamento e Teste Final	Validação Cruzada (cv=3)52. Teste com dados de validação53.	Obtenção de métricas de teste (Precisão 73%, Recall 67%, F1-Score 70%)54.
Fase 4: Implantação	Serialização do modelo (**joblib**)55. Implementação da interface e carregamento do modelo no serviço (Streamlit Community)56.	Modelo em produção e realizando novas previsões através da interface57.

4. Desafios e Mitigações

Desafio Potencial	Estratégia de Mitigação
Overfitting e Viés nos Dados: O modelo apresentava *overfitting* devido a uma coluna ("estado_reserva") que era uma cópia da variável alvo5858.	**Qualidade dos Dados (Mitigação Imediata):** Eliminação manual da coluna ofensiva que a seleção Passo a Passo deixou escapar, corrigindo o viés59.
Desempenho Subótimo do Modelo: O desempenho obtido, embora bom, ainda não era o "melhor possível"60.	**Qualidade dos Dados (Mitigação a Longo Prazo):** Acredita-se que tratar *ainda mais* os dados melhorará a qualidade e levará a resultados extraordinários61.
Seleção Agressiva de Features: O GridSearchCV demonstrou ser muito agressivo, selecionando apenas três *features*62.	**Ajuste Metodológico:** Optou-se pela seleção Passo a Passo para garantir que *features* menos relevantes, que podem ajudar o modelo a generalizar, fossem incluídas63.

5. Considerações Éticas

O projeto de Machine Learning apresenta considerações éticas e operacionais, especialmente no que tange à segurança e ao impacto das previsões.

Viés e Preconceito: Foi crucial identificar e remover colunas que poderiam introduzir viés de forma acidental (como a coluna que era uma cópia da variável alvo).

Impacto no Cliente (Falsos Positivos): O modelo escolhido tem um Recall alto, mas perde na Precision. Isso implica mais alarmes falsos (previsão de cancelamento para quem não cancelaria). A mitigação aceita é que o impacto de um **Falso Positivo** (tempo gasto em e-mails/chamadas desnecessárias) é financeiramente menor do que o impacto de um **Falso Negativo** (perda de receita por um cancelamento real não previsto).

Segurança e Monitoramento (Lacunas Atuais): Não foi aplicada qualquer medida de segurança ou parâmetro de **monitoramento de registros** do desempenho do modelo em produção. Em uma próxima iteração, será crucial explorar técnicas de monitoramento como **Grafana e Prometheus**.

6. Referências

Usando AutoML para previsão de demanda e cancelamento de hospedagem: https://matheus-buniotto.medium.com/usando-automl-para-previs%C3%A3o-de-demanda-e-cancelamento-de-hospedagem-hotel-e-resort-f2ac289ec7ac

Previsão da demanda turística com base em um modelo de rede neural temporal híbrido para turismo sustentável: https://www.mdpi.com/2071-1050/17/5/2210

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/discussion/565098