Nome: Jordan Butkenicius Malheiros

1. Quais passos você realizaria para solucionar este problema? Por favor descreva da forma mais completa e clara possível todos os passos que você enxerga como essenciais para a resolução do problema.

Para a solução do problema utilizamos o método CRIPS-DS, que envolve o desenvolvimento cíclico de projeto, com versões End-To-End e velocidade na entrega de valor e mapeamento de possíveis problemas e soluções. Os passos tomados durante os ciclos podem ser definidos pelas seguintes etapas:

1. Entendimento do Negócio.
2. Leitura / Coleta dos Dados.
3. Descrição dos Dados.
4. Featuring Engineering.
5. Data Filtering.
6. Análise Exploratória dos Dados (EDA).
7. Preparação dos Dados.
8. Seleção das Features.
9. Aplicação e Avaliação do Modelo de Machine Learning.
10. Hiper Parametrização dos parâmetros do modelo.
11. Aplicação Modelo Final .
12. Resultados de Negócio.
13. Modelo em Produção.

2. Qual métrica técnica de data science você utilizaria para solucionar este desafio? Ex: erro absoluto, rmse, etc.

Para solucionar esse desafio, nós utilizamos como principais métricas o recall e a AUC PR.

O recall é um score que usa como fundamento o cálculo de todas as previsões positivas do modelo, quais delas são realmente verdadeiras. Ou seja, quanto maior for o recall, maior é a quantidade de previsões da classe positiva (‘pos’) que nosso modelo acerta. Que é o que mais impacta o nosso resultado de negócio final.

A métrica AUC PR está mais relacionada à performance do modelo como um todo, não só a classe positiva como na métrica de Recall.

3. Qual métrica de negócio você utilizaria para solucionar o desafio?

A métrica de negócio utilizada para solução do problema foi a KPI de Custo($), por meio dessa métrica é que avaliamos se as estratégias estão funcionando de modo eficaz. Ao final de cada ciclo calculamos o custo final gerado($) pelo modelo e avaliamos seu desempenho quanto ao objetivo final de reduzir os gastos do sistema de manutenção de ar da empresa.

4. Como as métricas técnicas se relacionam com a de negócio?

A métrica de Recall está diretamente relacionada ao custo da empresa. Pois nesse caso, estamos dando prioridade a classe positiva “pos”, visto que é a que gera mais impacto no nosso negócio. O maior problema no nosso projeto é quando erramos as previsões de classes positivas pois ela tem um custo mais elevado se comparado a outros erros, por isso ela é definida como prioridade no nosso projeto.

5. Quais tipos de análises você gostaria de realizar na base de dados do cliente?

Durante o ciclo do projeto foram feitas algumas análise da base de dados, como: verificação das colunas, dimensões dos dados, tipo e transformação desses tipos de dados, checagem de valores faltantes, checagem de linhas e colunas duplicadas e também análise estatísticas descritiva.

Na seção de 4.0 (EDA) foram analisadas também a correlação e a distribuição das variáveis dos dados.

Um futuro trabalho interessante seria solicitar ao time de negócio a decodificação das colunas, para podermos ver o significado delas no mundo real e assim gerar algumas hipóteses de negócio que irão fornecer Insights ao time de negócio.

6. Quais técnicas você utilizaria para reduzir a dimensionalidade do problema?

No ciclo03 foi implementada a redução da dimensionalidade dos dados usando o PCA (Principal Component Analysis). Essa técnica consiste em transformar o espaço de dados originais (espaço das features) em um outro espaço de menor dimensão cujo os dados são projeções desses dados originais. É importante salientar que ao utilizar essa técnica de redução de dimensionalidade, nós perdemos a interpretabilidade das features originais.

7. Quais técnicas você utilizaria para selecionar variáveis para seu modelo preditivo?

Para a seleção das features utilizamos a abordagem “feature importance” gerado pelo modelo de XGBoost. O modelo em si estima a importância de cada variável para o modelo preditivo. Essa importância é uma pontuação que indica o quão útil ou valioso cada feature foi na construção da árvore de decisão, quanto mais um atributo é utilizado para tomar decisões importantes nas árvores de decisão, maior é sua importância relativa.

Também fizemos alguns testes ao utilizar “feature importance” de outros algoritmos como RandomForest e ExtraTrees e uma abordagem que envolve uma análise de variância (ANOVA).

8. Quais modelos preditivos você utilizaria ou testaria para este problema? Por favor indique pelo menos 3.
No nosso projeto utilizamos diferentes tipos de modelos, foram testados:

- KNN: que utiliza uma abordagem de vizinhos mais próximos (cálculo de distâncias) para classificar os dados nas classes.

- LogisticRegression: utiliza uma função que retorna a probabilidade de um label, e essa probabilidade é comparada a um limite predefinido, e o objeto é atribuído de acordo com o label.

- RandomForest: utiliza abordagem de árvores de decisão.

- LGBM E XGBoost: utilizar abordagem de árvores de decisão com *gradient boosting.*

9. Como você avaliaria qual dos modelos treinados é o melhor?

Para a avaliação dos modelos nós realizamos então a implementação de uma validação cruzada (cross_validation) que utiliza a abordagem de folds, no qual ele separa os dados de treino em folds e treina e retorna as métricas baseada em cada fold gerado. Essa abordagem é muito útil para se evitar o overfitting e para que possamos escolher o melhor modelo treinado baseado nas métricas definidas.

10. Como você explicaria o resultado de seu modelo? É possível saber quais variáveis são mais importantes?

Os resultados do modelo podem ser explicados por meio das métricas geradas pelo modelo (principalmente o recall) e Matriz de Confusão que pode ser traduzida para o âmbito de negócio, no qual podemos avaliar o impacto financeiro gerado pelo modelo a cada ciclo.

As variáveis mais importantes são aquelas que tiveram a maior score de “importance” gerada pelo modelo na seção seleção de features. No nosso caso, as 2 principais variáveis do nosso dataset são as features (ci_000, bj_000). Portanto seria interessante analisarmos o que elas realmente significam no contexto do negócio, para que assim, melhores decisões possam ser tomadas quanto ao tratamento dessas features na vida real.

11.Como você avaliaria o impacto financeiro do modelo proposto?

O impacto financeiro gerado foi avaliado por meio da comparação entre os modelos de cada ciclo por meio de tabelas e gráficos. O objetivo final é reduzir os custos, então nossa meta final era reduzir o custo para os dados do presente ano abaixo de $37.000. Esse objetivo foi alcançado com sucesso, pois nosso modelo final sugerido teve como custo um valor de $31.225, gerando uma economia de $5.775 doláres para os dados do ano presente.

12. Quais técnicas você utilizaria para realizar a otimização de hiperparâmetros do modelo escolhido?

Existem diversas técnicas que podem ser utilizadas para hiper parametrização do modelo, algumas delas:

Grid_Search: É uma técnica de tunagem de modelos que tenta calcular os valores ótimos de hiper parametrização utilizando-se da combinação de todos possíveis parâmetros definidos pelo usuário. O problema principal é que ele demanda muito tempo visto que modelos podem ter muitos parâmetros e isso acaba causando uma longa e exaustiva demora em sua execução.

Random Search: utiliza a mesma abordagem do Grid Search porém os parâmetros são aleatórios e não definidos pelo usuário. Isso significa que temos que definir um número máximo de iterações, e mesmo usando um número muito alto não é garantia que ele irá encontrar o melhor parâmetro possível.

Bayesian Optimization: ao contrário do Grid e RandomSearch, esse método acompanha os resultados de avaliações anteriores no qual eles usam um modelo probabilístico de mapeamento desses parâmetros para uma probabilidade de pontuação final.

Neste projeto utilizamos então o *optuna*, que é um *framework* de automatização de procura por hiperparâmetros que utiliza a implementação da Otimização Bayesiana.

13. Quais riscos ou cuidados que você levaria para o cliente antes de colocar este modelo em produção?

Um dos cuidados ao se levar em consideração para esse projeto é a escolha do threshold final do modelo. Podemos ajustar ele conforme o contexto de negócio, porém temos que ter consciência que há um tradeoff desse ajuste, no qual o nosso modelo ao mesmo tempo que ‘acerta’ mais as predições positivas, ele erra mais também predições da classe negativa ‘neg’. Essa decisão então é tomada baseada no objetivo final da resolução do problema de negócio.

14. Caso o seu modelo preditivo seja aprovado, como você colocaria ele em produção?

O modelo final foi colocado em produção por meio do Render que é um serviço para aplicações em Cloud gratuito. Esse modelo está disponível por meio de uma API que recebe os dados de teste e realiza a classificação dos dados nas classes ‘pos’ e ‘neg’.

Para melhor visualização foi implementado também um script para realizar as predições por meio de planilhas no Google Spreadsheets. No qual só é necessário preencher os dados que desejam ser preditos e utilizar um botão que roda um script que consulta a API e retorna a classificação daqueles dados.

15. Caso o modelo esteja em produção, como você faria seu acompanhamento?

O monitoramento do modelo é o mesmo que verificar se a performance avaliada no momento da construção é a mesma com o passar do tempo. Logo, se os dados forem mudando, seja em distribuição ou comportamento, o seu modelo também será afetado, geralmente caindo sua performance.

Para conseguir identificar as possíveis alterações de comportamento é comum utilizar algumas métricas de performance, como por ex:

- Métricas de performance do modelo.
- Métricas de estabilidade de variáveis.

Uma das maneiras mais fácil de realizar esse acompanhamento são:

- Gerar relatórios exibindo gŕafico das métricas.
- Gerar alertas sempre que uma métrica sair do range pré definido.
- Enviar email indicando que há algo fora do esperado na performance.

Esse processo pode ser manual ou então utilizar algumas ferramentas já conhecidas no mercado para essa automatização.

16. Caso o modelo esteja em produção, como você saberia o momento de retreinar o modelo?

Após o deploy do modelo em produção, as diferenças dos dados do mundo real resultaram em “model drift”, que na verdade é a degradação da capacidade preditiva de um modelo ao longo do tempo como resultado de uma mudança no ambiente. Portanto, o retreinamento e redeploy devem ser tratados como processos contínuos.

Para definirmos quando retreinar o nosso modelo, podemos adotar algumas estratégias:

- Retreino Periódico: decidindo sobre um intervalo de tempo em que o retreinamento para o modelo será realizado.

- Retreino baseado em triggers de performance: retreino baseado na performance do modelo, nessa abordagem um novo deploy talvez seja necessário visto que a performance do modelo abaixou.

- Retreino baseado em mudança dos dados: retreino por meio de detecção de alterações nas distribuições dos dados originais se comparados com os dados utilizados nos dados de treino do último deploy.

Então para saber o momento certo para retreinar é preciso acompanhar as métricas do modelo e comportamento dos dados de produção ao longo tempo.