3. Descreva a importância de implementar pipelines de desenvolvimento e produção numa solução de aprendizado de máquinas.

Em conjunto com o TDSP, nesse projeto utilizaremos os pipelines de desenvolvimento e produção, que é uma técnica fundamental quando queremos garantir a aplicação do TDSP e sua eficiência e confiabilidade em um modelo de aprendizado de máquina. Isso porque ela garante a eficiência e a consistência na criação e implantação de modelos de aprendizado de máquina, ajuda a economizar tempo e recursos reduzindo a necessidade de retrabalho, aumenta a confiabilidade e a estabilidade do modelo em um ambiente de produção, além de ajudar a garantir que o modelo esteja atualizado, e seja mantido conforme necessário, e permitir a integração com outros sistemas de software e infraestrutura.

4. Como as ferramentas Streamlit, MLFlow, PyCaret e Scikit-Learn auxiliam na construção dos pipelines descritos anteriormente? A resposta deve abranger os seguintes aspectos:

Para a construção desses pipelines utilizaremos diferentes ferramentas, sendo algumas opções Streamlit, MLFlow, PyCaret e Scikit-Learn. Abaixo estão descritos como cada uma delas auxilia considerando diversos aspectos:

1. Quando temos a necessidade de realizar rastreamento de experimentos para acompanhar os resultados de cada um deles, o MLFlow é a ferramenta ideal, uma vez que trabalha incluindo os parâmetros, métricas, artefatos e código. Além disso, nele também é possível comparar e visualizar os resultados dos experimentos de forma interativa, utilizando gráficos e tabelas.
2. Se temos como objetivo o treinamento de modelos, as ferramentas PyCaret e Sickit-Learn, que é conhecida por ser uma das bibliotecas mais populares para aprendizado de máquina em Python, são ótimas opções, visto que fornecem conjuntos de funções de treinamento de modelos prontas para uso e fáceiS de implementar, as quais incluem etapas como pré-processamento de dados, seleção de recursos, ajuste de hiperparâmetros, treinamento e avaliação de modelos.
3. Agora, quando precisamos monitorar a saúde de um modelo, temos como opções os recursos presentes nas ferramenta MLFlow e Streamlit. A MLFlow permite monitorar a saúde do modelo em produção com recursos de monitoramento de modelos, isso inclui acompanhamento de métricas de desempenho em tempo real, alertas de anomalias e uma interface de usuário para visualizar e comparar o desempenho do modelo. Enquanto a Streamlit possibilita a criação de interfaces de usuário interativas para monitorar o desempenho do modelo em tempo real, permitindo a visualização dos resultados de previsão em diferentes conjuntos de dados e acompanhamento da performance do modelo.
4. Caso haja necessidade de atualização de um modelo, temos como opções as ferramentas MLFlow e Sickit-Learn que fazem essa tarefa de maneiras diferentes, uma vez que uma, respectivamente, permite atualizar o modelo em produção com recursos de registro de modelo, incluindo o registro de diferentes versões do modelo, comparação de desempenho de diferentes versões e a implantação da nova versão do modelo em um ambiente de produção, enquanto a outra fornece uma biblioteca fácil de usar para treinamento de modelos que permite a atualização do modelo por meio da reexecução do pipeline com novos dados.
5. Por fim, quando a necessidade é de provisionar modelos de diferentes maneiras, todas as quatro ferramentas são capazes de realizar esse trabalho, isso pelo fato de todos possuirem recursos de implantação, permitindo que os modelos treinados sejam facilmente implantados em um ambiente de produção. Algumas diferenças são notadas quando comparamos funções específicas, por exemplo, a Streamlit permite construir interfaces de usuário para modelos de aprendizado de máquina, facilitando a visualização e a interação com os resultados do modelo. O MLFlow, por sua vez, inclui recursos de monitoramento e registro, permitindo que os usuários rastreiem o desempenho do modelo em tempo real e registrem todos os experimentos de treinamento. Enquanto isso, o Scikit-Learn, por ser uma biblioteca tão popular, tem o diferencial de ser compatível com outras ferramentas e plataformas, tornando mais fácil integrar modelos treinados em diferentes ambientes de produção.

5. Com base no diagrama realizado na questão 2, aponte os artefatos que serão criados ao longo de um projeto. Para cada artefato, indique qual seu objetivo.

Como descrito anteriormente, o TDSP é uma metodologia de trabalho em etapas, e cada uma delas pode ser representada por um artefato. Por isso, na etapa de Aquisição e Compreensão serão gerados artefatos a partir de funções de limpeza, exploração e preparação de dados, onde podemos incluir manipulações de colunas além de criação de novas variáveis a partir das originais e tratar da separação dos dados entre treino e teste. Na etapa de Modelagem criaremos artefatos com todos os modelos treinados, bem como os resultados, métricas de avaliação e gráficos de visualização de cada um deles. Já na etapa de Implantação serão criados artefatos que contém os testes de aderência, além de funções de monitoramento de desempenho do modelo e documentação da implantação.

6 c.

Algumas técnicas como aumento de dados de treino, regularização para evitar sobreajuste e balanceamento de classes (com reamostragem, pesos e etc), podem ser utilizadas para minimizar o efeito de viés dos dados.

7 c.

A escolha do modelo de Árvores de Decisão foi feita com base na natureza dos dados e da tarefa de classificação em questão, como não foi realizada a análise descritiva, ocorre a possibilidade de existir relações não lineares entre as variáveis preditoras e a variável resposta. Além disso, as árvores de decisão são relativamente fáceis de interpretar e permitem identificar as variáveis mais importantes para a classificação.

8

a.

Sabendo que o valor do log loss foi maior com a base nova do que com a base inicial e que o valor do F1 score do modelo com a nova base foi menor do que o resultado do modelo com a base inicial é possível afirmar que a nova base não é aderante ao modelo treinado.

b.

Quando temos a variável resposta disponível, é possível monitorar a saúde do modelo usando métricas de desempenho como o log loss, o F1 score e outras métricas relevantes para o problema em questão. Já quando não temosn a variável resposta disponível, o monitoramento da saúde do modelo pode ser mais desafiador, uma vez que não há uma métrica de desempenho clara para avaliar a qualidade das previsões do modelo. Nesse caso, é necessário monitorar outras métricas relevantes, como a distribuição dos dados de entrada e a taxa de previsões incorretas ou ambíguas do modelo.

c.

A estratégia reativa de retreino envolve a recriação do modelo inteiro sempre que novos dados são adicionados à base de dados. Isso significa que o modelo é treinado novamente usando toda a base de dados atualizada, sem levar em consideração o modelo anterior. Já a estratégia preditiva de retreino, envolve a atualização incremental do modelo à medida que novos dados são coletados. Nessa abordagem, o modelo é treinado usando apenas os novos dados e o modelo anterior, em vez de toda a base de dados.

Produtos pagos do Colab - Cancelar contratos