ROTEIRO PARA ELABORAÇÃO DO MINI-PROJETO
BOOTCAMP em Machine Learning

Como elaborar o seu mini-projeto:

No ato da inscrição, o candidato deverá preencher um formulário com o mini projeto de aplicação de conhecimentos prévios em análise de dados. O objetivo principal é verificar se o candidato possui capacidade analítica em resolução de problemas a partir de análise de dados. Para tanto, deve apresentar o esboço de um problema que enfrenta em seu ambiente profissional ou acadêmico e possíveis caminhos que vislumbra para sua solução, que venha se beneficiar de aplicações de machine learning, em poucas linhas.

Os elementos que deve conter esse pré-projeto são:

1. Título;

2. Problema de negócio a ser resolvido, destacando a relevância para o órgão ou Administração ou a generalidade do problema, cuja solução poderia beneficiar outras instituições;

3. Solução considerada, destacando possível abordagem, viabilidade e referências a soluções similares, deixando explícito exatamente aquilo que se deseja prever por meio de modelagem e quais as informações disponíveis para realizar essa predição; e

4. Fontes de dados que serão utilizadas para o projeto, destacando disponibilidade, qualidade, nível de publicidade ou sigilo da base e familiaridade do candidato com esses dados. Dados em formato tabular (planilha) com colunas de números, categorias ou texto. Não poderão ser utilizados dados de outros tipos, como imagens, vídeos, áudios, séries temporais ou banco de dados normalizados.

5. Mini amostra de dados (Opcional, recomendado): Anexar no formulário de inscrição uma amostra desses dados em formato tabular (planilha excel ou arquivo csv) explicitando quais são as variáveis independentes que servirão de base para prever a variável dependente, alvo da modelização. Cada linha será uma observação independente das demais linhas, composta de valores (numéricos, categóricos ou textuais) para cada variável independente, representadas nas colunas. Prever uma amostra com algumas centenas de linhas e explicitar o tamanho do conjunto de dados já disponíveis (ou que possa ser extraído até o início do curso) em termos de número de linhas e colunas. Explicitar qual a coluna

com a variável dependente que se deseja prever e a origem desses valores, seja por anotações manuais realizadas por especialistas ou por dados históricos.

Exemplo um conjunto de dados para uma tarefa de detecção/classificação de transações fraudulentas em cartão de crédito:

500 linhas de transações de cartão autorizadas por uma operadora no passado; cada linha/transação é caracterizada por um conjunto de informações que estão disponíveis no momento da autorização da transação, que correspondem às colunas da tabela, por exemplo: valor da transação, data da transação, antiguidade do cliente, saldo em conta do cliente, percentual do limite do cartão já utilizado, categoria do vendedor (ramo de operação, tipo de produto ou serviço), localização do vendedor (país, estado, município), etc. A variável dependente, nesse caso, seria se cada uma dessas transações foi reportada como fraudulenta ou não até seis meses após a sua realização. Assim, a tarefa consistiria em prever se uma nova transação seria ou não fraudulenta em função das características anteriores, para conceder ou não uma autorização. Neste exemplo, os valores da variável dependente são extraídos do histórico de transações passadas, mas esses rótulos também poderiam ter sido gerados por anotações manuais por especialistas de negócio.

Exemplos de projetos de esperado:

Exemplo 1:

1. Classificação automática do objeto da reclamação de passageiros do transporte aéreo.
2. Passageiros reclamam de problemas relacionados às suas viagens aéreas no site consumidor.gov.br em campo de texto aberto. A ANAC precisa identificar os problemas específicos de cada reclamação para fins estatísticos e de controle das empresas aéreas. Esse trabalho foi feito durante anos manualmente, acumulando um histórico de dezenas de milhares de reclamações classificadas nos temas mais relevantes. Deseja-se automatizar essa tarefa.
3. Acredita-se que os temas considerados pela ANAC estejam associados a frequência relativa de palavras-chave empregadas na descrição da reclamação. Assim, seria possível, por exemplo, distinguir uma reclamação sobre bagagem extraviada de um atraso de vôo em função do número de ocorrências de algumas poucas palavras-chave sem precisar realmente entender o texto.
4. O site consumidor.gov.br é transparente e os dados de relato das reclamações já foram extraídos e associados à classificação manual realizada nos últimos anos. Os relatos são muito variados em nível de linguagem e vocabulário empregado, mas parece haver um claro padrão na frequência de palavras-chave.

(Referência:
http://www.ipea.gov.br/sites/images/mestrado/turma3/esa-pekka-tapani-horttanainen.pdf)

Exemplo 2

1. Identificação das melhores escolas no Enem.
2. Tradicionalmente consideram-se melhores escolas aquelas com melhores resultados médios de seus alunos. No entanto, muito do resultado de um aluno está fortemente relacionado às suas características socio-econômicas, que também estão associadas à escola que cursa. Deseja-se um método objetivo para distinguir escolas com alunos com desempenhos abaixo ou acima do que se espera deles, em função dessas características socio-econômicas, que possam ser dadas como contribuição da escola.
3. Por exemplo, se considerarmos apenas o fator renda dos pais, poderia-se agregar as escolas em patamares de renda média dos pais e assim, para cada segmento, identificar as escolas com melhores resultados o que limitaria a influência do fator renda dentro de cada segmento. Idealmente, procuraremos generalizar esse isolamento de fatores, treinando um modelo preditivo para que estime o desempenho de um aluno em função de todas as suas características, excluindo apenas a escola frequentada. Depois compraríamos esse desempenho estimado com o desempenho real e ordenaríamos às escolas pela contribuição ao desempenho além do esperado.
4. Os microdados anonimizados do Enem encontram-se publicados pelo Inep (inep.gov.br/dados) com boa qualidade de dados em múltiplos anos, para milhões de alunos. Os dados socioeconômicos e de desempenho na prova encontram-se presentes para cada aluno individualmente, viabilizando assim a modelagem desejada.
(Referência: https://exame.com/brasil/7-rankings-mais-realistas-do-desempenho-das-escolas-no-enem/)

Fontes para elaboração do pré-projeto: recomenda-se explorar o painel de projetos já realizados por alunos de turmas anteriores para inspiração e ajuste do nível de complexidade do projeto de ML que pode ser desenvolvido durante o curso:
https://datastudio.google.com/reporting/fc8c5035-8650-4478-ac85-2e7d551bd2a9