Propostas de aplicação de técnicas de machine learning

1. título

2. problema de negócio a ser resolvido

3. solução considerada

4. fontes de dados

5. descreva seu nível de conhecimento do problema de negócio e experiência com o uso das fontes de dados citadas.

Planilha para contribuições de ideias de projeto:

<https://docs.google.com/spreadsheets/d/1oEjyRTojVAqRikoPWRKFU5banPgQuFr4BEEtIf9HGB4/edit#gid=0>

Exemplos de projetos de machine learning

**Exemplo 1**

1. Classificação binária de textos de deliberações em acórdãos do TCU sobre atos de pessoal quanto ao tema "suspensão de pagamentos"

2. A Sefip do TCU analisa um grande número de atos de pessoal e precisa monitorar que as determinações sobre suspensão de pagamentos sejam repercutidas no Siape. Para tanto precisa classificar os itens de deliberações individuais de cada acórdão sobre atos de pessoal se trazem determinação quanto à suspensão de pagamentos. Essa classificação poderia ser realizada automaticamente para ganhos de produtividade.

3. Há um padrão na frequência das palavras empregadas na descrição de determinações sobre suspensão de pagamentos que diferencia esse tema dos demais relacionados a atos de pessoal. A frequência relativa desses termos pode então ser utilizada para construir um classificador automático.

4. Os acórdãos sobre atos de pessoal emitidos pelo TCU são documentos públicos disponíveis no portal do órgão. Já as anotações de quais itens deliberam sobre suspensão de pagamento foram realizadas manualmente pelo autor do trabalho citado.

Ref:<https://portal.tcu.gov.br/biblioteca-digital/mineracao-de-deliberacoes-para-monitoramento-de-atos-de-pessoal.htm>

**Exemplo 2**

1. Priorização de fiscalização de licitações no ComprasNet em função de padrão histórico de irregularidades

2. O ComprasNet registra a maior parte das licitações federais, agrupando metadados que as descrevem. Deseja-se estimar o nível de risco e suas razões para uma licitação ainda não concluída.

3. A base histórica do ComprasNet está disponível (publicamente ou reservada aos órgãos de controle?) e poderia ser cruzada com uma base de condenações por irregularidades associadas a licitações, que hoje encontra-se dispersa. O resultado do julgamento pelo TCU de algumas dessas licitações poderia ser a anotação/target a ser previsto por um modelo de ML.

4. Base do ComprasNet (existente, mas disponibilidade incerta). Base de licitações consideradas irregulares, dispersa em acórdãos.

**Exemplo 3**

1. Prevendo a Taxa de Mortalidade Infantil em Municípios Brasileiros

2. Identificar o que priorizar para reduzir a taxa de mortalidade infantil para um município específico, comparando suas características com as de outros municípios para suporte à decisão de que medidas gerariam o maior impacto.

3. O uso de mais de 200 identificadores demográficos dos 5.565 municípios brasileiros disponíveis no Atlas do Desenvolvimento Humano no Brasil (http://www.atlasbrasil.org.br/2013/pt/) permitiria prever algum outro indicador de difícil mensuração direta. Aqui, no caso, procura-se estimar a taxa de mortalidade infantil com base nos demais indicadores, avaliar a contribuição de cada indicador no modelo para então decidir quais desses indicadores procuraria-se alterar para promover uma redução na taxa de mortalidade. ALERTA: o modelo de previsão aponta correlações entre indicadores e não relações de dependência causal. Assim, alterações induzidas em um indicador por alguma política pública não necessariamente surtirá o efeito esperado pelo modelo correlacional de alteração no indicador alvo.

4. <http://www.atlasbrasil.org.br/2013/pt/>

Ref: <https://www.kaggle.com/c/icict-fiocruz-taxa-mort-brasil/overview> (encontra-se sem os dados preparados, mas viável de ser replicado a partir do Atlas)

**Exemplo 4**

1. Classificação de Risco de Fraude em Benefícios Trabalhistas

2. Como distinguir casos de fraude nos requerimentos do seguro desemprego, para focar esforço de fiscalização?

3. Temos uma base de requerimento ao seguro desemprego, que traz diversos dados correlacionados à probabilidade de ser uma fraude. Existe uma base consolidada pelo Min. Trabalho que catalogava em 2018, cerca de 2000 requerimentos cujas fraudes foram confirmadas. Essa base de fraudes pode servir de anotação (target) para o processo de treinamento do modelo de previsão.

4. Base de requerimento seguro desemprego disponível no LabContas mas a base de fraudes é sigilosa e foi descontinuada.

Ref: Análise de Aloísio Dourado (TCU): <http://www.brasildigital.gov.br/brasil-digital/eventos-anteriores/brasil-100-digital-4/programa/>

**Exemplo 5**

1. Priorização de denúncias à ouvidoria do TCU

2. O Sistema S é um dos principais alvos de denúncias à ouvidoria do TCU, deseja-se priorizar as denúncias recebidas para alocação de força investigativa em função da probabilidade de relevância da denúncia.

3. A base histórica de denúncias registra quais foram analisadas e as conclusões após análise. Seria então viável construir um modelo para prever qual seria a posição do TCU sobre uma nova denúncia, com base na similaridade da nova denúncia com as denúncias já analisadas no passado. Target: posição do TCU quanto à denúncia.

4. A Ouvidoria do TCU tem uma base de denúncias, com alguns dados estruturados que em conjunto com o texto livre da denúncia poderia ser mapeada no resultado histórico após análise humana. Grau de acesso aos dados incertos.

**Exemplo 6**

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 sócio-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 socioeconômicas, que possa ser imputado 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, procuramos 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 comparamos esse desempenho estimado com o desempenho real e ordenamos as 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.

Ref: <https://exame.com/brasil/7-rankings-mais-realistas-do-desempenho-das-escolas-no-enem/>

**Exemplo 7**

1. Prevendo irregularidades no Seguro Safra

2. O Seguro Safra indeniza agricultores quando da perda de sua produção. Deseja-se identificar os casos mais prováveis de fraude para suspender o seguro antes do pagamento.

3. A base histórica do Seguro Safra indica metadados associados à produção segurada e a posterior detecção de fraudes em pagamentos irrecuperáveis. Seria viável mapear os padrões de similaridade das fraudes já identificadas, para suspender os pedidos de indenização antes do pagamento.

4. A disponibilidade da base do Seguro Safra, com registro de fraudes passadas, é incerta.

**Exemplo 8**

1. Prevendo valor de venda de imóveis da Secretaria de Patrimônio da União

2. Forte incerteza quanto aos valores de venda de imóveis da SPU leva a falhas em seu registro contábil e imóveis paralisados, sem uso nem disposição à venda

3. Utilizar uma base de valores de vendas de imóveis, seja por exemplo, com a base de financiamentos de bancos públicos como a Caixa Econômica, ou por raspagem de preços solicitados em sites de imobiliárias, como o WImoveis, para associar preços de venda com características de imóveis (localização, área total, número de cômodos, acabamento, etc.). O modelo criado sobre o mercado privado poderia ser assim validado com resultados de pregões da SPU (ex. Concorrência Pública SPU/MP 01/2017)

4. Dados de imóveis da SPU estão disponíveis no site. Já os valores de imóveis no Brasil precisariam ser extraídos por web scraping de sites de imobiliárias ou com convênio com a Caixa (sigiloso). Possível alternativa: https://www.kaggle.com/mpwolke/brazil-real-state-listings

Ref: sugestão de Marcelo Silva (TCU) e Thiago Marzagão (CGU) em trabalhos ainda não publicados

**Exemplo 9**

1. Prever a classificação de assunto a partir de sumário de acórdãos do TCU

2. O mapeamento de assunto por texto livre do sumário permitiria realizar uma busca por assunto sobre parágrafos de texto em geral, encontrando referências fortemente correlacionadas ao assunto de interesse

3. Temos uma base de acórdãos com texto completo, sumário e assunto anotado

4. A base de acórdãos é facilmente acessível pelo Kaggle: <https://www.kaggle.com/ferraz/acordaos-tcu>

**Exemplo 10**

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.

Ref: <http://www.ipea.gov.br/sites/images/mestrado/turma3/esa-pekka-tapani-horttanainen.pdf>

**Exemplos de onde encontrar bases de dados**, que podem ser utilizados no treinamento de modelos:

- Painel colaborativo com fontes de dados: <https://funretro.io/publicboard/bJT50t6W4pgLc0bu29AtFDHroVf1/ec9a6b50-4513-4853-9ac0-a433989e49cc>

Alguns dos exemplos do painel acima

- Dados Abertos: <http://dados.gov.br/>

- Kaggle: Brasil: <https://www.kaggle.com/search?q=brazil>

- 500 Datasets clássicos para ML: <https://archive.ics.uci.edu/ml/datasets.php>

- Pesquisa por dataset: <https://datasetsearch.research.google.com/>

- Google Public Datasets: <https://cloud.google.com/public-datasets/>

- Plataforma +Brasil: <http://plataformamaisbrasil.gov.br/download-de-dados>

- Corpus 927: consolida decisões vinculantes do STF e do STJ, e a jurisprudência do STJ: <http://corpus927.enfam.jus.br/>

- Dados imobiliários Brasil: <http://www.properati.com.br/dados>

- Dados gerais do Brasil: <https://pt.knoema.com/atlas/Brasil>