Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Pedro Cesar Barros Loureiro

13 de Março de 2018

Uso de Machine Learning na Priorização de Ordens de Serviço na Compesa

Histórico do assunto

Empresas públicas, em geral, que prestam serviço à população recebem milhares de solicitações e reclamações mensalmente. A Compesa, empresa de distribuição de água e saneamento de Pernambuco, e na qual eu trabalho, não foge a esta regra. Apenas no ano de 2017 foram 1,579,817 aberturas de Registro de Atendimento (RAs), as quais geram Ordens de Serviço, onde 33.617 são relativos a vazamentos

Infelizmente boa parte destes problemas tem origem na infraestrutura antiga da cidade, onde são utilizados canos de ferro que enferrujam com o tempo e causam os vazamentos e faltas de água em bairros ou cidades inteiras. Como não há orçamento suficiente para atender todas as demandas o mais rápido possível é necessário priorizá-las baseado em fatores como o local em que ocorreu o problema e o número de pessoas afetadas.

Descrição do problema

Apesar da grande capacidade humana em resolver problemas de priorização, este processo pode tornar-se lento e até ineficaz dependendo da quantidade de informações necessárias para resolver o problema. No caso da priorização das Ordens de Serviço (OSs) há características não quantitativas, como o local da ocorrência, fonte da reclamação e tempo que está aberto que não podem ser calculadas com softwares tradicionais, necessitando intervenção humana. Utilizando técnicas de classificação é possível calcular um fator de priorização para cada uma

destes reclamações, assim, no futuro, quando uma nova OS for criada, será possível calcular este fator e ordená-las, priorizando sempre aquelas com o maior fator calculado. Existe a possibilidade de OSs sempre ficarem para trás, entretanto, estas OSs ficarão mais antigas, aumentando sua chance de serem classificadas como prioritárias devido a data de abertura mais antiga.

Conjuntos de dados e entradas

A Compesa possui um software comercial, o GSAN, no qual são armazenados todos as Ordens de Serviço. O software permite filtrar estas informações por tipo, incluindo vazamentos, que será trabalhado. Iremos trabalhar com estes dois tipos e que foram registrados no ano de 2017 fornecendo as seguintes características:

- **Data Abertura da OS:** Data em que a Ordem de Serviço foi aberta. Quanto maior o tempo de aberto mais chances há de ser priorizada.
- **Meio de Solicitação:** Forma como o RA foi aberto gerando as OSs.
    - 1: BALCAO
    - 4: INTERNO
    - 6: TELEFONE
    - 8: INTERNET
    - 9: OUTROS
    - 10: AGENCIA MOVEL
    - 11: MOBILE

    A empresa mostra interesse em priorizar RAs abertas pelo mobile, pois é mais barato e fideliza o cliente.

- **Cidade da Ocorrência:** Cidade onde a ocorrência acontece.
- **Bairro da Ocorrência:** Bairro onde a ocorrência acontece.
- **Quantidade de Reiterações :** Reiterações são aberturas de novas RAs relativas a RA original informando que o cliente reportou novamente sobre o mesmo problema em datas diferentes.
- **Diâmetro da rede:** O diâmetro da rede é um bom indicador do tamanho do vazamento. Logo, quanto maior o diâmetro maiores as chances de priorização.
- **Data de Encerramento da OS:** Data em que a Ordem de Serviço foi concluído. Este atributo não será considerado no treinamento, existindo aqui apenas para referência.
- **Priorizado:** Esta coluna não existe no banco de dados e será calculada. Serão consideradas prioritárias aquelas OSs cujo tempo entre a data de abertura e a data de encerramento seja menor que a mediana entre estas datas. É uma coluna binária, 0 ou 1,onde 1 são as OS priorizadas.

Descrição da solução

Serão utilizados os dados importados do GSAN para, inicialmente, será realizada uma análise PCA e separação de cluster para identificar a relação entre as características. Em seguida utilizarei um algoritmo de classificação, como Logistic Regression ou KNN calculando o grau de afinidade de cada RA em relação a sua priorização. Com este grau de afinidade será feita uma ordenação decrescente e os primeiros que se encaixarem na capacidade de resolução de RAs serão os priorizados.

Modelo de referência (benchmark)

O assunto em questão está mais restrito a empresas de saneamento ou prefeituras que são responsáveis pela distribuição de água, por isso não foi encontrado nenhum artigo relativo a este assunto para comparar diretamente. Baseado em outros problemas de priorização, em áreas diferentes, como a de manutenção de máquinas industriais, serão utilizados algoritmos que permitam visualizar a probabilidade de determinada Ordem de Serviço ser executada ou não (Logistic Regression parece ser o mais adequado).

Atualmente a classificação se uma ordem de serviço é prioritária ou não é realizada manualmente e dura algumas horas. O algoritmo será considerado válido se sua acurácia for de ao menos 80%, ou seja, há 80% de chance dele acertar que aquela OS é realmente prioritária. Este valor foi escolhido porque acredito que valores como 90% poderiam causar overfitting deixando OSs que deveriam ser priorizadas fora do ranking de priorização. É possível que este valor possa mudar ao longo do tempo caso seja necessário adicionar mais ou menos OSs para serem executadas.

Métricas de avaliação

Para validar a solução irei comparar o resultado do treinamento com o resultado de uma priorização real realizada pelas equipes de atendimento da Compesa. Serão escolhidas ao menos 100 Ordens de Serviço, prioritárias e não prioritárias, onde o algoritmo deverá classificá-las e o resultado final deve ser ao menos 95% similar à classificação manual.

Design do projeto

Inicialmente devem ser coletados os dados, alguns destes dados não estão na base de dados do GSAN, mas em outros projetos. Serão utilizados os dados das RAs

abertas de Janeiro de 2017 até Janeiro de 2018 que foram concluídas com realização do serviço, ou seja, não foram concluídas por cancelamento.

O segundo passo será fazer uma limpeza dos dados, removendo possíveis outliers, os dados que estão fora do intervalo interquartil. Para evitar overfitting utilizarei a análise PCA para verificar se há características correlacionadas, como o Bairro e a dimensão da rede talvez. Neste caso, considero como forte relação aquelas cujo fator de correlação seja acima de 0.7, entretanto, este valor pode ser alterado para se ajustar ao algoritmo final.

Com os dados prontos para análise utilizarei a função GridSearchCV para comparar o desempenho de alguns algoritmos de Classificação. Os algoritmos a serem testados serão: SVM e Logistic Regression.

O algoritmo que obtiver melhor desempenho em termos de precisão será utilizado nos testes finais. Com o resultado de cada registro, irei ordenar pela acurácia e comparar com a priorização manual realizada pela equipe da Compesa.