Nanodegree Engenheiro de Machine Learning

Proposta de Projeto Final

José Carlos Bezerra Filho

17/01/2018

Histórico do assunto

A formação de preço de um seguro é um desafio para as companhias de seguros ao redor do mundo, pois para fazer um preço é necessário entender qual o risco que cada cliente representa para a carteira da companhia. Em minha experiência de 7 anos como corretor de seguros eu aprendi que não existe um valor padrão para um seguro, principalmente de automóvel, algumas vezes um detalhe, que muitos acham pequeno, como o CEP pode agravar o prêmio em mais de 100%.

Ao longo dos anos as seguradoras ao redor do mundo têm investido quantias gigantes de dinheiro para evoluir seus departamentos de estatística, tudo isso para conseguir prever cada vez com mais precisão qual o risco que cada cliente representa para a sua carteira. Por ser um assunto com o qual trabalhei por muitos anos e sempre me deixou curioso eu escolhi fazer meu projeto baseado no desafio proposto pela Porto Seguro no Kaggle, onde o objetivo é encontrar qual a probabilidade de um determinado cliente usar o seguro no próximo ano. O desafio pode ser visualizado neste link: https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data.

Descrição do problema

A grande questão é: como prever um cliente de alto risco? A resposta para essa pergunta é o motivo de as companhias de seguro ao redor do mundo investirem tanto dinheiro para melhorar seus modelos. Veja que para elas é interessante atrair cliente bons e afastar clientes ruins e para fazer isso é preciso classificar cada cliente de acordo com o seu perfil. As informações que elas utilizam são as mais variadas como: idade, sexo, CEP, CPF, etc. Prevendo quem são os clientes de baixo risco (clientes bons) e os clientes de alto risco (cliente ruins) no momento da cotação elas são capazes de oferecer vantagens como descontos e serviços extras para clientes de baixo risco e agravar o valor do prêmio para clientes de alto risco. Para este problema, como se trata de um problema de classificação, o ideal é utilizar um modelo de aprendizagem supervisionada.

Conjuntos de dados e entradas

Para desenvolver esse projeto eu irei utilizar o dataset disponibilizado pela Porto Seguro na página do desafio. Ele é composto por um arquivo de treino e um arquivo de teste que já estão separados.

O arquivo de treino conta com 595.212 registros com 58 atributos, já o arquivo de teste conta com 892.816 registros com os mesmos 58 atributos. O que cada atributo representa não foi disponibilizado, o nome de cada um foi substituto por um aliás.

Descrição da solução

A minha intenção com este projeto é desenvolver um modelo que quando receba os atributos de um novo cliente em potencial ele possa fazer uma predição precisa de qual a probabilidade deste cliente vir a utilizar o seguro no próximo ano.

Modelo de referência

A minha referencia para a realização deste projeto é a submissão que o Felipe Antunes fez para o desafio, que pode ser encontrada neste link https://github.com/felipeeeantunes/kaggle-porto-seguro.

Design do projeto

Primeiramente eu vou importar os dados e analisa-los um pouco para entende-los melhor. Em um segundo momento vou fazer a limpeza dos dados. Outro ponto importante será a classificação da importância dos atributos através de um algoritmo de random forest, assim será possível saber quais atributos são irrelevantes para a analise e exclui-los.

Quando eu já estiver com a base carregada e limpa eu vou treinar 3 modelos diferentes com as configurações padrão, aquele que apresentar o melhor resultado será o que irei refinar para buscar um resultado melhor. Os 3 modelos que pretendo treinar são:

KNN – Este modelo me interessa pois a analise mais assertiva feita até hoje pelas seguradoras é pelo perfil. Este modelo consiste em agrupar os indivíduos semelhantes e quando entramos com um novo individuo ele faz uma predição baseada no resultado para os indivíduos que estão próximos dele. É mais ou menos como aquele ditado “me diga com andas que direi quem é”. Acredito que este é o modelo que deve se sair melhor.

Arvore de decisão – Este modelo é basicamente uma arvore binaria, em que analisa qual o peso de cada atributo para o resultado final. Quando entramos

um com um novo individuo pelo caminho percorrido no encadeamento dos atributos ele é capaz de nos retornar uma predição.

Aquele com o melhor resultado será refinado com uso de GridSearch, ou busca em matriz em português. Está técnica consiste em treinar e avaliar o modelo com diversas configurações diferentes para assim encontrar as configurações ideais.

Naive Bayes – Este modelo trabalha puramente com probabilidade, e através desta análise ele é capaz de retornar uma predição.

Métrica de avaliação

Para a avaliação do modelo será utilizado o Coeficiente de Gini Normalizado, a mesma métrica proposta no desafio original.