Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Vinicius Ferreira Santos
28 de setembro de 2018

Prevendo o risco de inadimplência do crédito residencial

O Brasil atualmente passa por uma grave crise financeira que culminou em mais de 13 milhões de desempregados. Muitos desses desempregados estavam pensando em adquirir seu primeiro imóvel, mas, sem um contracheque para comprovar a renda, acabaram perdendo o poder de financiamento e se juntando a outros milhares que não conseguem adquirir um financiamento de imóvel, devido a falta de crédito no mercado financeiro.
Pensando no problema de falta de crédito para milhares de pessoas ao redor do mundo e no poder de previsão que pode ser alcançado com machine learning, a Home Credit liberou seus dados e pediu ajuda para montar um modelo capaz de ajudar as pessoas com históricos de crédito insuficientes ou inexistentes a conseguirem empréstimos.

Acredito que a solução que será desenvolvida para este projeto, pode ser levada em consideração por todos os tipos de negócios que realizam a liberação de crédito para futuros clientes, somente se baseando em comprovação de renda por vínculo empregatício.

Histórico do assunto

Muitas pessoas lutam para obter empréstimos devido aos históricos de crédito insuficientes ou inexistentes e, infelizmente, essa população é frequentemente aproveitada por credores não confiáveis.

A Home Credit se esforça para ampliar a inclusão financeira para a população sem banco, proporcionando uma experiência de empréstimo positiva e segura. Para garantir que essa população carente tenha uma experiência de empréstimo positiva, a Home Credit utiliza uma variedade de dados alternativos - incluindo informações de telecomunicações e transacionais - para prever as capacidades de reembolso de seus clientes.

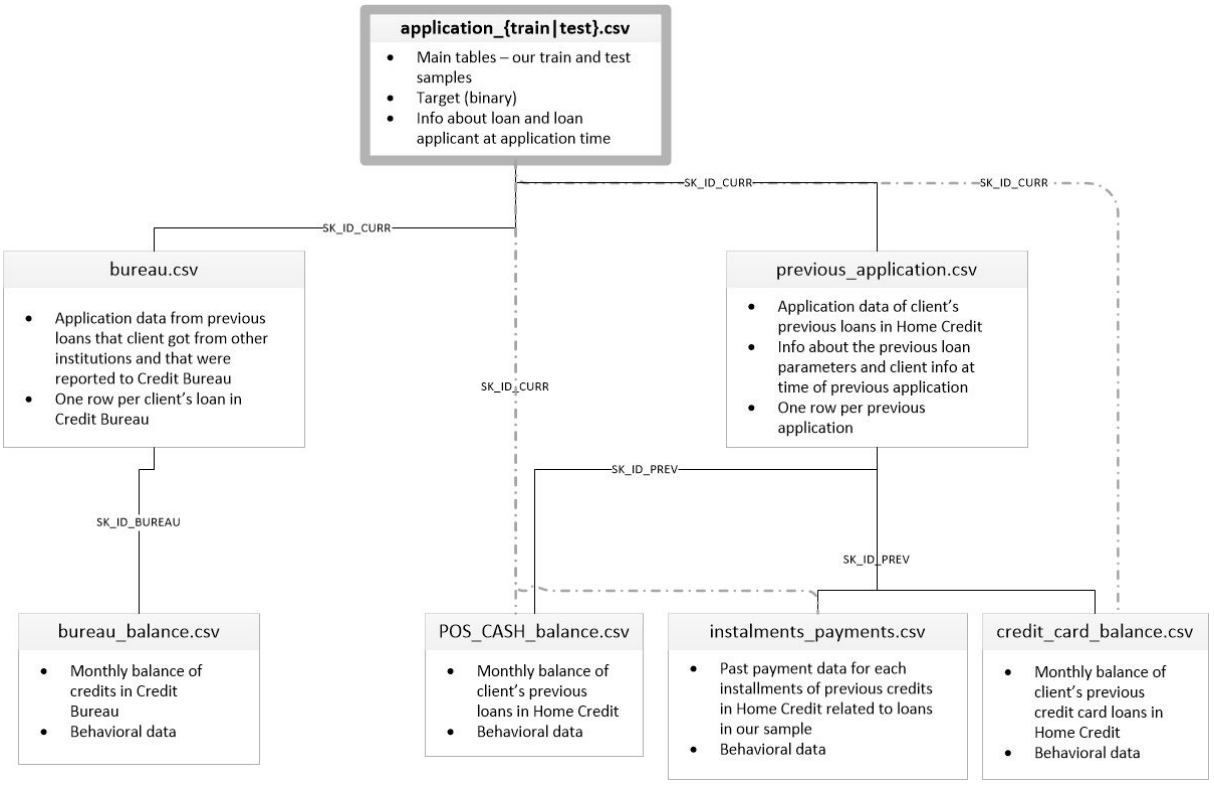
A Home Credit está atualmente usando vários métodos estatísticos e de aprendizado de máquina para fazer essas previsões. Fazer isso garantirá que os clientes com capacidade de pagamento não sejam rejeitados e que os empréstimos sejam concedidos com um calendário principal, de vencimento e de reembolso, que capacitará seus clientes a serem bem-sucedidos.

Descrição do problema

O principal objetivo desse projeto é prever quais pessoas sem crédito bancário que terão capacidade de pagar seus empréstimos. Como esse é um problema clássico de aprendizado supervisionado de classificação, utilizarei os dados para treinar um modelo que informará se o cliente está apto ou não a receber um empréstimo. Após o modelo treinado, o mesmo poderá ser utilizado para prever clientes futuros, com a possibilidade de acompanhar o desempenho das previsões através de métricas.

Conjuntos de dados e entradas

Como pode ser visto no diagrama de dados abaixo, os dados possuem diversas origens como, por exemplo, os dados de outras instituições(bureau.csv), mas para treinar o modelo que será utilizado no projeto, utilizarei somente os dados do arquivo application.csv de treino e teste.
Os disponibilizados pela Home Credit estão disponíveis no kaggle.



Descrição de dados

- application_{train|test}.csv
    - This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
    - Static data for all applications. One row represents one loan in our data sample.
- bureau.csv
    - All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).
    - For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.
- bureau_balance.csv

    - Monthly balances of previous credits in Credit Bureau.
    - This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.
- POS_CASH_balance.csv
    - Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
    - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows.
- credit_card_balance.csv
    - Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
    - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows.
- previous_application.csv
    - All previous applications for Home Credit loans of clients who have loans in our sample.
    - There is one row for each previous application related to loans in our data sample.
- installments_payments.csv
    - Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
    - There is a) one row for every payment that was made plus b) one row each for missed payment.
    - One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.
- HomeCredit_columns_description.csv
    - This file contains descriptions for the columns in the various data files.

Descrição da solução

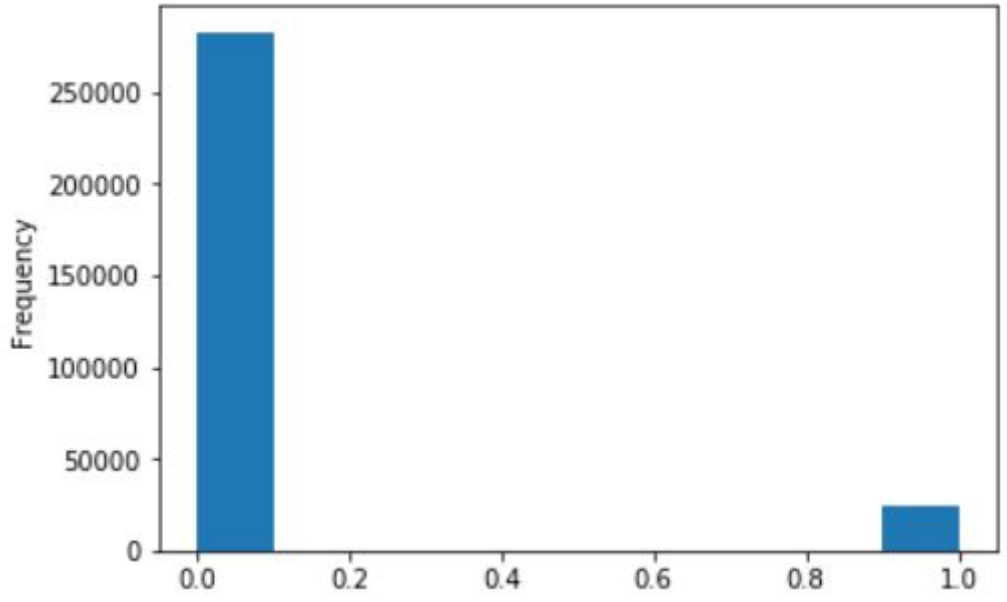
Para criar o modelo que fará as previsões, primeiro será feito uma análise exploratória para ter um maior entendimento sobre os dados. Com a análise em mãos, saberemos muitas informações sobre os dados e será realizado um pré-processamento utilizando várias técnicas como: feature engineering para criação de novas features a partir das existentes, Scale, feature importances para verificar a importância das features através do treinamento de uma RandonForest, regressão logística para inferir os dados categóricos faltantes e etc. Após realizar todo o pré-processamento necessário, será criado uma pipeline utilizando algoritmos de aprendizagem supervisionada especializados em classificação, após isso será selecionado o modelo que apresentar o melhor desempenho e em seguida o tunning para melhorar os valores das previsões.

Modelo de referência (benchmark)

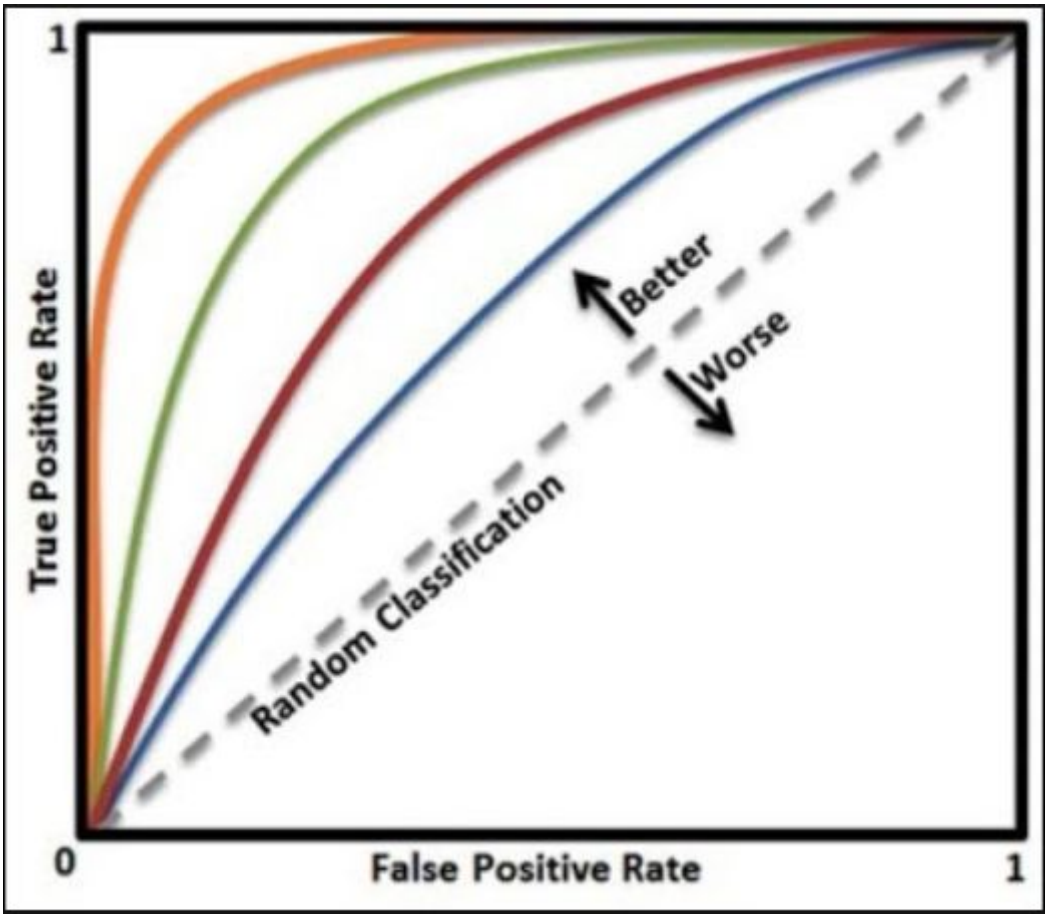
Como modelo de referência, usaremos um random classifier com valor base de 0.5 e para comparar os resultados, o modelo final selecionado deverá ultrapassar o valor base do random classifier utilizando a métrica selecionada.

Métricas de avaliação

A escolha da métrica que será utilizada neste projeto é baseado em um problema muito comum quando estamos trabalhando com classificação, **classes desbalanceadas**.



Dentro dos dados de treinamento que serão utilizados, existem uma coluna chamada target, que indica se um empréstimo foi pago a tempo (valor = 0) ou se o se o cliente teve dificuldade para realizar os pagamentos(valor = 1). Como podemos perceber na imagem acima, nossos dados possuem mais informações de bons pagadores do que mal pagadores. Para esse problema em particular, é recomendado utilizar a métrica **ROC AUC**, que basicamente mede a taxa de exemplos positivos, quando eles realmente são positivos **TP** e a taxa de exemplos positivos, quando na verdade eles são negativos **FP.**



A imagem acima mostra como a métrica funciona, basicamente os resultados que estão acima da linha pontilhada estão obtendo resultados melhores que o random classifier, caso contrário, estão prevendo pior que uma seleção aleatória.

Design do projeto

A construção do modelo preditivo seguirá um processo geral com as fases abaixo:

1. **Coletar os dados**
2. **Explorar os dados:** as fases podem, em alguns momentos, ser intercaladas entre si:
    - Explorar as características dos dados;
    - Explorar os dados estatisticamente;
    - Identificar valores ausentes;
    - Identificar ruídos nos dados(outliers);
    - Explorar os dados visualmente;
    - Estudar as distribuições dos atributos;
    - Estudar possíveis correlações entre os atributos.
2. **Preparar os dados:**
    - Fixar ou remover ruídos;
    - Engenharia de atributos;
    - Seleção de características(atributos);
    - Preencher valores ausentes(com média, zero...) ou removê-los.
3. **Escolha e Afinamento de um Modelo:**
    - Treinar uma variedade de classificadores: como nosso problema é de natureza supervisionada, os algoritmos usados deverão atender essa característica;
    - Medir e comparar as performances;
    - Selecionar o melhor modelo;
    - Afinar e Tunar o modelo;
4. **Avaliar e Validar:**
    - Feita em paralelo com a fase anterior.
5. **Apresentar a solução para o problema:**
    - Documentar a solução;
    - Explicar porque a solução atinge o objetivo desejado;
    - Explicar o que funcionou e não funcionou

- Apresentar os resultados

Referências:

http://gim.unmc.edu/dxtests/roc3.htm

https://en.wikipedia.org/wiki/Receiver_operating_characteristic

https://www.kaggle.com/c/home-credit-default-risk

http://www.chioka.in/class-imbalance-problem/

http://www.homecredit.net/

Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn & TensorFlow: 1 ed. O' REILLY