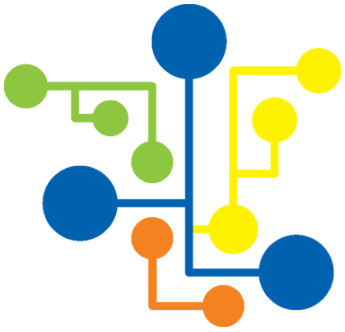

Data Science Academy

www.datascienceacademy.com.br

Microsoft Power BI Para Data Science

Projeto

Prevendo a Inadimplência de Clientes

O uso de cartão de crédito é bastante comum nos dias de hoje e um uso correto dele realmente ajuda na construção de uma boa pontuação de crédito. Quão importante é a pontuação de crédito? Sempre que você precisa realizar uma compra a prazo, sua pontuação de crédito é consultada e pode ser o fator decisivo na aprovação da sua compra. No entanto, a inadimplência pode fazer com que a pontuação de crédito caia. Não apenas a pontuação de crédito cai, como também haverá um efeito adverso sobre o limite de crédito e empréstimos futuros de qualquer tipo.

Neste projeto, estaremos lidando com um histórico de pagamentos de clientes em Taiwan. O conjunto de dados tem 24 atributos e um rótulo de classe e existem 30000 instâncias. Criaremos um modelo preditivo para permitir que o banco possa prever se o seu cliente será inadimplente no próximo pagamento ou não.

Objetivo: construir um classificador e usá-lo para prever se o cliente do cartão de crédito será inadimplente no próximo pagamento.

Conjunto de dados: Dados bancários do repositório de Machine Learning da UCI:

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#

Antes de iniciar nossa análise, precisamos nos familiarizar com o conjunto de dados. O conjunto de dados está desequilibrado, ou seja, a maioria dos rótulos de classe são *não-inadimplentes*. Das 30000 instâncias, 78% são não-inadimplentes e 22% restantes são inadimplentes.

A seguir estão os atributos presentes no conjunto de dados: ID, equilíbrio de crédito, gênero, educação, estado civil e idade são auto-explicativos. Pay_0, Pay_2, Pay_3, Pay_4, Pay_5, Pay_6 são o estado do pagamento nos meses de abril a setembro, respectivamente. O status do pagamento é definido como o atraso no pagamento. Exemplo: se o valor de Pay_0 for -1, então significa que o cliente foi devidamente pago, se o valor for 2, isso significa que o pagamento está atrasado por dois meses. Bill_Amt1 a Bill_Amt6 são os montantes das faturas do cartão de crédito para o mês de abril até setembro. Pay_Amt1 a Pay_Amt6 são o valor que o cliente pagou na conta do cartão de crédito no mês de abril até setembro.

Abaixo, a tabela com uma descrição de cada um dos atributos:

Atributo	Descrição
ID	ID único de cada registro
Credit Balance	Quantidade de crédito no cartão de crédito
Gender	Sexo do cliente (masculino/feminino)
Education	Nível de Escolaridade, i.e. Pos-graduado, Graduado, Ensino Médio, Outros
Marital Status	Estado Civil, i.e. casado, solteiro, outros
Age	Idade do cliente
Pay_0	Status de Pagamento em Setembro
Pay_2	Status de Pagamento em Agosto
Pay_3	Status de Pagamento em Julho
Pay_4	Status de Pagamento em Junho
Pay_5	Status de Pagamento em Maio
Pay_6	Status de Pagamento em Abril
Bill_Amt1	Valor da Conta do cartão em Setembro
Bill_Amt2	Valor da Conta do cartão em Agosto
Bill_Amt3	Valor da Conta do cartão em Julho
Bill_Amt4	Valor da Conta do cartão em Junho
Bill_Amt5	Valor da Conta do cartão em Maio
Bill_Amt6	Valor da Conta do cartão em Abril
Pay_Amt1	Valor pago em Setembro
Pay_Amt2	Valor pago em Agosto
Pay_Amt3	Valor pago em Julho
Pay_Amt4	Valor pago em Junho
Pay_Amt5	Valor pago em Maio
Pay_Amt6	Valor pago em Abril
Default_payment_next_month	Valor 0 ou 1 - 0 significa não-inadimplência e 1 significa inadimplência

Todos os atributos do conjunto de dados são inteiros, então convertemos - gênero, educação, estado civil, idade e status de reembolso em fator. Isso ajudaria na classificação dos clientes. Convertemos o rótulo da classe, ou seja, default_payment_next_month para fator também e renomeamos o nome da coluna. O valor da conta (Bill_Amt) e o valor do pagamento (Pay_Amt) permanecem inteiros.

Nós excluímos a coluna ID, pois não é útil em nossa análise, sendo apenas um número sequencial que identifica cada instância de forma exclusiva. Nós também excluímos as linhas que tinham valores faltantes (missing). Remover essas poucas linhas não é relevante para esta análise em particular.

Análise de Dados

Parece haver uma forte correlação negativa entre saldo de crédito e status de pagamento (de abril a setembro). O status de pagamento representa o atraso no pagamento, como mencionado anteriormente. Isso pode ser investigado em mais detalhes posteriormente.

Em seguida, dividimos o conjunto de dados em conjunto de dados de treinamento e conjunto de dados de teste. Para construir nosso modelo, usamos método de amostragem estratificada e uso de validação cruzada de 10 folds para treinar e avaliar o nosso modelo. Construímos nosso conjunto de dados de treinamento com 45% dos dados para processar nosso classificador mais rápido. Esses parâmetros podem ser alterados e outros valores podem ser testados.

Construção e Avaliação do Modelo

Utilizamos o RandomForest como algoritmo para a construção do classificador e obtivemos 81% de acurácia na primeira versão do modelo. Na sequência listamos os atributos mais relevantes no conjunto de variáveis preditoras. Essas relações devem ser investigadas em mais detalhes.

Usamos o coeficente Mean Decrease Gini a fim de medir como cada variável contribui para a homogeneidade dos nós e as folhas na floresta aleatória (conjunto de árvores de decisão) resultante. Em termos simples, classificamos a importância das variáveis de cima para baixo. Menos importantes são os dados demográficos, isto é, escolaridade, idade, estado civil e sexo.

De acordo com este problema de negócio, um classificador seria considerado satisfatório se tivesse um valor de recall mais alto porque o banco estaria mais interessado em conhecer os aspectos positivos reais, ou seja, o número de clientes que provavelmente seria inadimplente no próximo mês.

A pontuação F1 (também chamada F-score ou F-measure) é uma medida da precisão de um teste. Considera tanto a precisão p como o recall r do teste para calcular a pontuação. Existem duas outras medidas, a saber, F2, que dá um peso mais alto para o recall em relação a precisão (colocando mais ênfase em falsos negativos) e F0.5, que dá um peso menor para o recall em relação a precisão (atenuando a influência de falsos negativos).

Publicação do Modelo

Existem diversas formas de publicar o modelo criado, que deverá ser aplicado a novos conjuntos de dados a fim de realizar as previsões.