Tainá Moura Nogueira

Modelos de Aprendizado de Máquina para Previsão de Default

Brasil

2022, v-1.9.7

Tainá Moura Nogueira

Modelos de Aprendizado de Máquina para Previsão de Default

Dissertação apresentada ao Curso de Mestrado Profissional em Economia, Universidade de Brasília, como requisito parcial para a obtenção do título de Mestre em Economia

Universidade de Brasília - UnB

Faculdade de Administração Contabilidade e Economia - FACE

Departamento de Economia - ECO

Programa de Pós-Graduação

Orientador: Daniel Oliveira Cajueiro

Brasil

2022, v-1.9.7

Tainá Moura Nogueira
Modelos de Aprendizado de Máquina para Previsão de Default / Tainá Moura Nogueira. – Brasil, 2022, v-1.9.7-
56p. : il. (algumas color.) ; 30 cm.

Orientador: Daniel Oliveira Cajueiro

Dissertação (Mestrado) – Universidade de Brasília - UnB
Faculdade de Administração Contabilidade e Economia - FACE
Departamento de Economia - ECO
Programa de Pós-Graduação, 2022, v-1.9.7.

1. Palavra-chave1. 2. Palavra-chave2. 3. Palavra-chave3. II. Universidade de Brasília. III. Faculdade de Administração, Contabilidade e Economia - FACE. IV. Departamento de Economia IV. Modelos de Aprendizado de Máquina para Previsão de Default

Tainá Moura Nogueira

Modelos de Aprendizado de Máquina para Previsão de Default

Dissertação apresentada ao Curso de Mestrado Profissional em Economia, Universidade de Brasília, como requisito parcial para a obtenção do título de Mestre em Economia

---

Daniel Oliveira Cajueiro
Orientador

Herbert Kimura
Convidado 1

Regis Augusto Ely
Convidado 2

Brasil
2022, v-1.9.7

Agradecimentos

Agradeço ao meu orientador, Daniel Oliveira Cajueiro, pela sua orientação, competência e por todos os ensinamentos ao longo do mestrado.

À todos os professores que estiveram conosco e contribuíram para nosso aprendizado e crescimento durante todo o período.

À minha mãe, Arlete, heroína que sempre me deu apoio incondicional e encheu a minha vida de muito amor e carinho . Agradeço aos incontáveis sacrifícios que levaram à minha formação como pessoa e profissional.

À Raíssa por todo o amor, companheirismo, apoio e incentivo durante o período, compartilhando inúmeros momentos de ansiedade e estresse.

Agradeço à toda a minha família e amigos que sempre estiveram presentes em todos os momentos.

Aos amigos Leila e Leonardo pela compreensão necessária durante o curso, Sheilla e Victor que fizeram a minha jornada do mestrado mais leve e contribuíram para o meu aprendizado e entregas.

Aos colegas de empresa Fabiano, Márcio, Hernany e Thiago pela disponibilidade e ajuda na construção do trabalho.

"Talvez não tenha conseguido fazer o melhor,
mas lutei para que o melhor fosse feito. Não sou o que
deveria ser, mas Graças a Deus, não sou o que era antes".
(Marthin Luther King)

Resumo

Nesse trabalho, realizamos um estudo a respeito da inadimplência em contratos habitacionais concedidos ao segmento de pessoa física por uma Instituição Financeira enquadrada no segmento S1, utilizando uma amostra de 31647 contratos pactuados entre Janeiro de 2016 e Janeiro de 2021, que tinham como garantia real imóveis entre 500 mil e 1 milhão de reais. A partir das informações disponibilizadas, selecionamos um conjunto de características relacionadas ao tomador e ao contrato. Incorporamos à análise variáveis macroeconômicas comumente utilizadas pela literatura, que demonstraram ter sido bastante relevantes na construção do modelo.Considerado o grande desbalanceamento identificado na base, utilizamos algumas técnicas de reamostragem e aplicamos 5 diferente classificadores na base balanceada e desbalanceada com o intuito de comparar o desempenho das combinações entre diversas técnicas para previsão de contratos que venham a ficar inadimplentes em 30, 60 e 90 dias. Os resultados obtidos indicam maior eficiência da técnica de reamostragem SMOTEENN e dos classificadores Random Forest, Regressão Logística e KNN.

Palavras-chave: Aprendizado de máquina; Risco de Crédito; Previsão; Classificação

Abstract

In this work, we did a study about the default in housing finance granted in the individual segment by a Financial Institution classified in the S1 segment, using a sample of 31647 contracts agreed between January 2016 and January 2021, that they had as a real guarantee properties between 500 thousand and 1 million reais. From the information provided, we select a set of characteristics related to the borrower and the contract. We integrate macroeconomic variables commonly incorporated in the literature into the analysis, which demonstrated to be quite relevant in the construction of the model. Considering the large imbalance identified in the base, we used some resampling techniques and applied 5 different classifiers to the balanced base and unbalanced base to compare the performance of combinations among several to predict contracts that will become defaulter in 30, 60 and 90 days. The results obtained indicate greater efficiency of the SMOTEEN resampling technique and the Random Forest, Logistic Regression and KNN Classifiers.

Keywords: Machine learning, Credit risk, Forecast, Classification

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

IBGE	Instituto Brasileiro de Geografia e Estatística
IPEA	Instituto de Pesquisa Econômica Aplicada
BACEN	Banco Central do Brasil
IPCA	Índice Nacional de Preços ao Consumidor Amplo
SELIC	Sistema Especial de Liquidação e Custódia
IBOVESPA	Índice Bovespa
PIB	Produto interno bruto

Sumário

1 Introdução

A intermediação financeira consiste em facilitar a transferência de recursos entre os agentes superavitários, aqueles cujas rendas superam as suas despesas, e os agentes deficitários, cuja renda não cobre as despesas. Nesse sentido, os intermediários financeiros desempenham papel importante dentro da economia real. Em períodos normais, o setor financeiro demonstra sua importância na capacidade de suavizar as flutuações econômicas e multiplicar os investimentos. No entanto, em períodos de crise, a fragilidade do setor financeiro pode contribuir para a instabilidade da economia.

Conforme observamos na literatura vigente, os bancos são considerados os intermediários financeiros mais relevantes, entre outras instituições, especialmente nos países em desenvolvimento. Segundo Ozgur et al. (2021), os bancos realizam tarefas relacionadas ao tamanho do risco, maturidade e transformação, equilibrando as grandes necessidades de financiamento de longo prazo dos mutuários com a coleta de poupanças de pequeno e curto prazo. Para Seven e Yetkiner (2016), o principal papel dos mercados e instituições financeiras em todas as economias é melhorar a eficiência da alocação de capital e estimular a poupança, impulsionando o crescimento, levando a uma maior formação de capital, a mobilização de poupança, a gestão de riscos e a facilitação de transações.

Ozgur et al. (2021) destaca que os bancos possuem um papel relevante em quase todo o mundo, uma vez que a concessão de empréstimos bancários pode afetar os lados financeiro e real da economia, por várias razões. Para as instituições bancárias, os empréstimos bancários são ativos essenciais e as principais fontes de rendimento dos bancos (MALEDE, 2014).

Desde a crise financeira 2007-2008, a supervisão da gestão de risco tem sido foco de atenção por parte dos reguladores que esperam ver nos bancos estruturas de medição de risco transparentes e auditáveis, pois a falência de um banco pode afetar outras instituições financeiras por meio do efeito de contágio. Esses choques também podem se espalhar para outros setores da economia, causando queda no preço dos ativos e diminuição da atividade econômica.

Nesse sentido, considerando a relevância dos empréstimos bancários para a atividade econômica e estabilidade financeira dos países, conforme observado na literatura da área, diversos são os estudos que tentaram elucidar os fatores que influenciam o comportamento dos empréstimos bancários nas economias emergentes e avançadas.

Nos últimos anos, técnicas de modelagem quantitativa tem sido usadas para otimizar a entrada de dados, reduzir custos e aumentar a lucratividade geral. O aprendizado de máquina “supervisionado” gira em torno do problema da previsão: produzir previsões de y

a partir de x, descobrindo padrões generalizáveis(MULLAINATHAN; SPIESS, 2017).

De acordo com FSB (2017), a utilização de Inteligência Artificial e Aprendizado de Máquina em serviços financeiros pode trazer benefícios importantes para a estabilidade financeira, contribuindo na eficiência da prestação do serviço, no processamento de informações sobre risco de crédito e menor custo de interação com o cliente, bem como na supervisão de riscos regulatórios e sistêmicos.

O risco de crédito, que envolve empréstimos não pagos, pode ser considerado o mais presente na atividade bancária. Podemos defini-lo como a possibilidade de perda para um banco devido à incapacidade dos devedores de empréstimos cumprirem pontualmente ou completamente as obrigações que assumiram como parte de seus contratos com a instituição. A avaliação eficaz do risco de crédito tornou-se um fator crucial para a obtenção de vantagens competitivas no mercado de crédito, que podem ajudar as instituições financeiras a conceder crédito a clientes com capacidade de honrar seus compromissos e rejeitar clientes que não poderão, reduzindo assim as perdas(ZHOU; LAI; YU, 2010).

Nesse estudo, propomos um comparativo de vários classificadores de aprendizado de máquina da literatura discutida, usando diferentes estratégias de amostragem, considerando o desequilíbrio característicos de dados desta natureza, com o objetivo de lidar com a previsão da inadimplência em empréstimos.

Nos últimos anos, muitos pesquisadores tem empregado abordagens de aprendizado de máquina nas pesquisas de pontuação de crédito.Malekipirbazari e Aksakalli (2015) realizam um estudo comparando diferentes métodos de aprendizado de máquina, incluindo Random Forest(RFs), Regressão Logística (LR) e k-vizinhos (k -NN).Por sua vez, Yeh e Lien (2009) investigam os prós e contras de seis técnicas de mineração de dados: K-vizinhos (KNN), Regressão Logística, Análise Discriminante, Classificador Bayesiano Naive, Redes Neurais Artificiais e Árvores de Decisão. Os autores concluíram que a acurácia entre as seis técnicas mostra que há poucas diferenças nas taxas de erro entre os métodos, no entanto, as redes neurais artificiais realizam a classificação com mais precisão do que os outro cinco métodos.

Alguns estudos procuraram demonstrar que técnicas inteligentes como Redes Neurais Artificiais (RNA), Árvore de Decisão (DT), Raciocínio Baseado em Caso (CBR), Máquina de Vetor de Suporte (SVM) podem ser utilizadas como métodos alternativos para previsão de falência corporativa (OLSON; DELEN; MENG, 2012; TSAI; WU, 2008).Ozgur et al. (2021) focaram no impacto de 19 variáveis específicas do banco, macroeconômicas e globais nos empréstimos bancários, comparando o desempenho do modelo de regressão com métodos de aprendizado de máquina, como Regression Tree, Boosting, Bootstrap Aggregating (bagging), Random Forest, Extremely Randomized Trees (extra-trees), and Extreme Gradient Boosting (xgboost).

A principal contribuição do presente trabalho está relacionada com a aplicação de metodologias atuais de Machine Learning para aprimorar o poder preditivo da instituição financeira detentora dessa carteira, permitindo a identificação de potenciais tomadores que venham a ficar inadimplentes, tendo como base as características do mutuário, da operação de crédito e fatores macroeconômicos.

A presente dissertação está organizada da seguinte forma. A seção 2 detalha o nosso conjunto de dados e analisa cada conjunto de variáveis que usamos. Na seção 3, apresentamos os modelos, as especificações utilizadas para a previsão e o procedimento adotado para escolha dos parâmetros. Em seguida, mostramos nossos resultados gerais e examinamos os melhores modelos na Seção 4. Por fim, na Seção 5, apresentamos uma breve conclusão sobre nossos achados.

2 Dados

Nesse trabalho, utilizamos dados de operações de crédito habitacional concedidas por uma relevante instituição financeira brasileira do segmento S1[1], entre janeiro de 2016 e janeiro de 2021, a fim de manter um horizonte preditivo mínimo de 12 meses.

Os dados correspondem somente a linha de crédito destinada à pessoa física para compra de imóveis novos ou usados, para pagamento de prestações mensais. Com o objetivo de manter a amostra de contratos dentro de um mesmo segmento, selecionamos apenas aqueles que tinham como garantia da operação imóveis com valores entre 500 mil e 1 milhão de Reais.

O conjunto de dados utilizado demonstrou-se bastante desequilibrado, pois o número de inadimplentes é consideravelmente inferior ao montante daqueles que permaneceram com os empréstimo sem atraso. Conforme observado por Namvar et al. (2018), situações que envolvem desequilíbrio de classe surgem quando há um número muito maior ou menor de objetos em uma classe do que em outra e, nesses casos, prever efetivamente o risco de crédito a partir de um conjunto de dados desequilibrado é difícil, pois afeta a capacidade do modelo de discriminar entre bons tomadores e potenciais inadimplentes, e os algoritmos de mineração de dados ignoram as classes minoritárias, focando na majoritária.

2.1 Seleção das Amostras

Foram extraídos 45.784 contratos da base de dados, dentre eles, aqueles que poderiam ser considerados adimplentes e os definidos como inadimplentes. No entanto, identificou-se a necessidade de realizar tratamento prévio nos dados antes da aplicação dos modelos, resultando em um total 31647 contratos .

De forma geral, o pré-processamento compreende atividades de preparação, organização e estruturação dos dados e trata-se de uma etapa altamente relevante para o desenvolvimento do trabalho que deve ser executada antes da aplicação dos modelos, a fim de aumentar a qualidade dos dados e proporcionar maior confiabilidade nas informações.

Nesse sentido, realizamos a limpeza dos dados, removendo outliers e eliminando registros nulos da base semelhante ao executado por Namvar et al. (2018). Todas as linhas que possuíam qualquer registro em branco em qualquer uma das variáveis foram excluídas da amostra. Optamos por não realizar nenhuma técnica de preenchimento dos valores nulos

[1] O S1 é composto pelos bancos múltiplos, bancos comerciais, bancos de investimento, bancos de câmbio e caixas econômicas que tenham porte igual ou superior a 10% (dez por cento) do Produto Interno Bruto (PIB) ou exerçam atividade internacional relevante, independentemente do porte da instituição

que utilizasse média, mediana ou frequência, práticas comumente utilizadas em trabalhos semelhantes, por considerar que se tratava de uma quantidade muito pequena de situações que, uma vez excluídas, não prejudicariam os resultados do trabalho.

Nos trabalhos de Lee et al. (2004), Yeh e Lien (2009), Steenackers e Goovaerts (1989) encontramos algumas das variáveis explicativas que poderiam constar nesse estudo. Ao todo, serão utilizadas 28 variáveis independentes, detalhadas na sessão a seguir.

Algumas variáveis possuíam grande variabilidade de valores ou categorias. Para essa situação, utilizamos uma técnica de agrupamento dos dados semelhantes em clusters. Removemos os valores considerados outliers e transformamos os dados categóricos em variáveis numéricas, por meio das técnicas de Label Encoder e One hot encoder.

Em seguida, realizamos transformação de dados originais em formatos mais apropriados e adequados para o processo de mineração, envolvendo a atividade de padronização utilizando a biblioteca StandardScaler do Python, técnica que coloca os dados em uma mesma escala.

Após a consolidação dos dados e seleção das variáveis, chegamos ao total de 28 variáveis que serão apresentadas a seguir:

2.1.1 Variáveis Independentes

As variáveis independentes podem ser divididas em 3 categorias: contrato, tomador e macroeconômicas.

2.1.1.1 Variáveis dos Contratos

As características dos contratos foram fornecidas pela instituição financeira a partir de uma extração realizada no sistema que compila as informações referentes às operações de crédito habitacional.

As variáveis selecionadas estão detalhadas na Tabela 1.

2.1.1.2 Variáveis dos Tomador

Outras variáveis utilizadas no estudo estão relacionadas às características dos mutuários, conforme Tabela 2.

2.1.1.3 Variáveis Macroeconômicas

De acordo com Figlewski, Frydman e Liang (2012), a incorporação de fatores macroeconômicos juntamente com outras variáveis relacionadas nos modelos de risco de crédito leva a um aumento altamente significativo do poder explicativo. Esses autores agrupam os fatores em três grandes classes: Relacionados às condições macroeconômicas

Tabela 1 – Variáveis Contrato

Variável Contrato	Descrição
Indexador	Índice selecionado no momento da contratação para atualização monetária da operação de crédito
Taxa de Juros	Taxa de juros pactuada na contratação da operação de crédito
Valor da Prestação	Obrigação mensal do tomador com a operação de crédito
Valor do Contrato	Valor que foi financiado na operação de crédito (R$)
Valor da Garantia	Valor do imóvel objeto da contratação (R$)
Tipo do Contrato	Informa se é um contrato normal ou fruto de renegociação
Prazo do Contrato	Prazo discriminado em meses utilizado no financiamento
Prazo Remanescente	Prazo discriminado em meses que faltam para vencimento do contrato
Quantidade de Operações Contratadas pelo tomador	Quantidade de operações de crédito contratadas pelo tomador
Créditos do Tomador a Liberar	Créditos contratados com recursos a liberar
Crédito do Tomador a Vencer	Créditos contratados a vencer
Créditos Vencidos	Créditos contratados vencidos
Créditos Baixados Prejuízo até 48 meses	Operações em inadimplemento por prazo igual ou superior a 60 meses, na data-base ou operações com vencimentos baixados como prejuízo há mais de 48 meses
Parcela Relativa	Variável criada a partir da divisão do valor da parcela mensal pela renda bruta
Vlr_Entrada	Variável criada a partir da diferença entre o valor da garantia e o valor do contrato firmado.

Fonte: Elaborado pelo Autor

Tabela 2 – Variáveis dos Tomador

Variável Tomador	Descrição
Gênero	Sexo do tomador de crédito
Idade	Idade do Tomador de crédito
Renda Bruta	Renda anual recebida pelo tomador
Estado Civil	Estado civil do tomador
UF	Unidade da Federação onde reside o tomador de crédito
Ocupação	Atividade, serviço ou trabalho principal do tomador
Escolaridade	Nível de escolaridade atingido pelo Tomador

Fonte: Elaborado pelo Autor

gerais (taxa de desocupação, inflação, etc); Relacionadas à direção em que a economia está seguindo (PIB, mudança no sentimento do consumidor, etc) e Fatores das condições

do mercado financeiro (Taxa de Juros, retorno do mercado de ações, etc).

Nesse sentido, selecionamos 5 variáveis dispostas na Tabela 3. A série histórica do PIB e Taxa de desocupação foram coletadas no portal IPEADATA. Para avaliarmos a variável inflação, utilizamos o IPCA que foi obtida na página do IBGE. Por fim, temos a Taxa básica de juros, Selic, e o Índice BOVESPA, principal indicador de desempenho das ações negociadas na Bolsa de Valores, extraídos junto ao BACEN, por meio do site da instituição, e do portal da B3, respectivamente.

Tabela 3 – Variáveis Macroeconômicas

Variável Macroeconômica	Descrição
Índice BOVESPA[1]	O Ibovespa é o principal indicador de desempenho das ações negociadas na B3 e reúne as empresas mais importantes do mercado de capitais brasileiro
IPCA[2]	Índice Nacional de Preços ao Consumidor Amplo
Selic[3]	Taxa básica de juros da economia
PIB(Produto Interno Bruto)[4]	Soma de todos os bens e serviços finais produzidos pelo país
Taxa de Desocupação[4]	Percentual de pessoas desocupadas, na semana de referência, em relação às pessoas na força de trabalho no mesmo período

Fonte: B3[1], $IBGE$[2], $BACEN$[3], $IPEADATA$[4]

Uma visão geral das variáveis macroeconômicas e suas estatísticas descritivas é apresentada na Tabela 4.

Tabela 4 – Estatísticas Descritivas das Variáveis Macroeconômicas

	IPCA	SELIC	IBOVESPA	DESOCUPAÇÃO	PIB
mean	0.330412	3.550614	100108.625431	13.921038	-1.912214
std	0.254973	2.744737	10029.933941	1.365497	3.345829
min	-0.380000	1.190000	62711.470000	11.100000	-10.730000
25%	0.240000	1.900000	99369.150000	12.800000	-3.710000
50%	0.240000	1.900000	99369.150000	14.800000	-3.710000
75%	0.330000	4.400000	100967.200000	14.800000	1.290000
max	1.350000	12.900000	126801.660000	14.900000	12.300000

Fonte: Elaborado pelo Autor

2.1.2 Variáveis Dependentes

A variável dependente deste estudo foi calculada a partir de um campo na base disponibilizada referente à quantidade de dias de atraso. Para esse estudo, consideramos 3 cenários em que o tomador seria considerado "mau pagador":

Cenário	Critério
Cenário 1	>= 30 dias de atraso
Cenário 2	>= 60 dias de atraso
Cenário 3	>= 90 dias de atraso

,

Tabela 5 – Cenários de Default

Codificamos a variável dependente igual a 1 se o solicitante ultrapassou a quantidade de dias estipuladas na tabela acima, assumindo como inadimplente. Caso contrário, a variável assume valor zero se, em nenhum dos meses, o tomador deixou ultrapassar 30, 60 ou 90 dias de atraso.

3 Metodologia

3.1 Classes Desbalanceadas

Situações de desequilíbrio de classe são muito comuns em trabalhos que envolvem problemas de classificação. Esse desequilíbrio ocorre quando temos uma amostra que apresenta o número muito de superior de uma das classes comparativamente à outra objeto de estudo. Nesses casos, os classificadores tem uma tendência à priorizar a classe majoritária, enquanto a minoritária é ignorada. Pesquisadores e estudiosos da área parecem concordar com a hipótese de que o desequilíbrio entre as classes é o maior obstáculo na indução de classificadores em domínios desequilibrados.(BATISTA; PRATI; MONARD, 2004)

Para solução desse problema, foram projetados diversos algoritmos, dentre eles os de reamostragem que gera um conjunto de dados de treinamento de equilíbrio antes de construir o modelo de classificação. Os três tipos de reamostragem são sobreamostragem, subamostragem e um híbrido dos dois. (NAMVAR et al., 2018)

Pesquisas anteriores demonstraram que as técnicas de reamostragem podem melhorar o desempenho de classificação e vários tipos de técnicas foram comparados na literatura para tentar determinar a maneira mais eficaz de superar um grande desequilíbrio de classe. Batista, Prati e Monard (2004) recomendam a utilização dos métodos SMOTE+Tomek ou SMOTE+ENN para conjuntos de dados com um pequeno número de instâncias positivas e para aquelas amostras que possuem uma quantidade maior de exemplos positivos, o método de sobreamostragem aleatória. Japkowicz (2000) demonstrou que tanto a sobreamostragem quanto a subamostragem são métodos muito eficazes de lidar com o problema, embora a abordagem de downsizing funcione melhor do que a abordagem de sobreamostragem em grandes amostras.

No presente trabalho, realizamos testes com as técnicas de SMOTE, SMOTEENN, Tomek Links e Nearmiss. A que apresentou maior contribuição para o melhor desempenho foi SMOTEENN, resultado semelhante ao obtido por Batista, Prati e Monard (2004) com as técnicas de SMOTEENN e SMOTE+Tomek que foram muito boas para conjuntos de dados com um pequeno número de exemplos positivos do seu estudo.

3.2 Modelos

O principal objetivo da nossa pesquisa é realizar um comparativo entre vários classificadores de aprendizado de máquina da literatura discutida, usando diferentes

estratégias de amostragem com o objetivo de lidar com a previsão da inadimplência em operações de crédito a partir do mapeamento de um conjunto de características de empréstimos, tomadores e variáveis macroeconômicas.

Para esse estudo, selecionamos algumas técnicas muito utilizados pela literatura como Random Forest, K-Vizinhos mais próximos (K-NN) e Regressão Logística (NAMVAR et al., 2018; MALEKIPIRBAZARI; AKSAKALLI, 2015; SUN et al., 2018). Incluímos, ainda, o método de Gradient Boosting Chen e Guestrin (2016) e XGBoost (WANG et al., 2022).

3.2.1 Regressão Logística

Inicialmente, selecionamos o modelo de regressão logística, que é utilizado para prever a probabilidade de uma determinada classe, neste caso, probabilidade de inadimplência. O modelo mostra a relação entre os recursos e, posteriormente, calcula a probabilidade de um resultado.

Essa abordagem é bastante utilizada na literatura principalmente com questões que envolvem risco de crédito por ser, inclusive, mais aceita pelos órgãos reguladores, dada a sua menor complexidade e maior explicabilidade. Segundo Bracke et al. (2019), os modelos de regressão são considerados relativamente interpretáveis, especialmente quando seus coeficientes de regressão tem um significado econômico claro.

Para Dumitrescu et al. (2022), a regressão logística continua sendo a referência na indústria de risco de crédito, principalmente porque a falta de interpretabilidade dos métodos de conjunto é incompatível com as exigências dos reguladores financeiros.

3.2.2 Random Forest

O modelo de árvore de decisão é bastante utilizado em estudos que envolvem problema de classificação e tem como premissa a formação de trilhas que criam uma divisão dos dados, com bases nas características, em pequenos grupos para que, ao final, quando introduzido um novo dado, ele possa nos dizer em qual grupo esse dados se encaixa melhor.

O Algoritmo Random Forest é baseado na utilização de diversas árvores de decisão que funcionam em conjunto. Todas as árvores individuais fazem previsões independentes da classe de saída correta e a mais comum é selecionada como valor de saída.

Segundo Malekipirbazari e Aksakalli (2015), as árvores de decisão por si só nem sempre são competitivas com outras técnicas de classificação, então, para melhorar a precisão das árvores, às vezes é necessário empregar métodos de conjunto, como boosting e bagging.

Alguns estudos tem apresentado resultados de estudo indicando que as técnicas de Random Forest e Gradient Boosting têm um desempenho muito bom em um contexto de pontuação de crédito e são capazes de lidar comparativamente bem com desequilíbrios de classe pronunciados nesses conjuntos de dados (BROWN; MUES, 2012).

3.2.3 Gradient Boosting

Gradient Boosting é uma técnica Boosting, que faz parte do grupo de classificadores Ensemble e utiliza a combinação de resultados de preditores fracos, com o intuito de produzir um modelo mais eficiente. Ele melhora a precisão de uma função preditiva por meio da minimização incremental do termo de erro. Depois que o aprendiz de base inicial (mais comumente uma árvore) é cultivado, cada árvore da série é ajustada aos chamados "pseudo-resíduos" da previsão das árvores anteriores com o objetivo de reduzir o erro (FRIEDMAN, 2001; FRIEDMAN, 2002).

3.2.4 XGBoost

O XGBoost é composto de várias árvores de decisão, cada uma com um poder preditivo baixo, e a tecnologia Boosting aumenta a capacidade de ter melhores resultados. No problema de classificação, embora a precisão de previsão global de cada árvores não seja alta, ele pode demonstrar poder de previsão relevante em algum aspecto dos dados.

O algoritmo Xgboost usa o método bagging para reduzir o viés, o método boosting para diminuir a variância e regressões fts para aumentar a eficiência e a precisão nas funções objetivas (PETROPOULOS et al., 2019).

O modelo de avaliação de risco de crédito pessoal baseado no XGBoost possui forte capacidade de discriminação de inadimplência e robustez (WANG et al., 2022). Segundo Ma et al. (2018), XGBoost é um dos métodos mais avançados de aprendizado de máquina desenvolvidos nos últimos anos.

3.2.5 K-vizinhos mais próximos (K-NN)

O algoritmo de K-vizinhos mais próximos (K-NN) é relativamente simples, mas ainda amplamente utilizado em problemas de classificação. A técnica classifica uma instância de dados considerando apenas as k instâncias de dados mais semelhantes no conjunto de treinamento. O resultado é atribuído conforme a classe da maioria dos k-vizinhos mais próximos.

A principal vantagem dessa abordagem é que não é necessário estabelecer um modelo preditivo antes da classificação. As desvantagens são que K-NN não produz uma fórmula de probabilidade de classificação simples e sua precisão preditiva é altamente afetada pela medida de distância e pela cardinalidade k da vizinhança (YEH; LIEN, 2009).

3.3 Métricas de Avaliação

A métrica mais tradicional e popular em um problema de classificação binária é a acurácia.Contudo, ao avaliar conjuntos de dados desbalanceados, a acurácia tende a enfatizar a classe majoritária, dificultando o bom desempenho do classificador na classe minoritária (NAMVAR et al., 2018). Além disso, ela não considera que falsos positivos são mais importantes do que falsos negativos(ABELLáN; CASTELLANO, 2017).

Desta forma, selecionamos outras métricas de avaliação comumente encontradas na literatura Namvar et al. (2018), Moscato, Picariello e Sperlí (2021): Accuracy (ACC), Recall , Precision , F1, área sob a curva (AUC), G-mean e Matriz de Confusão.

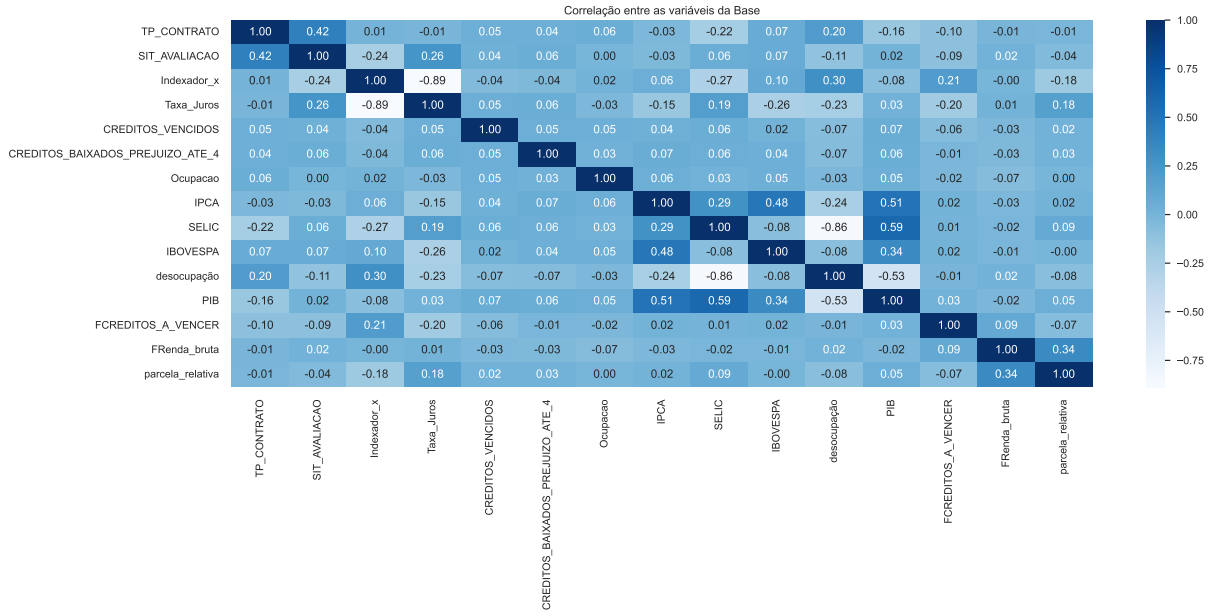
4 Resultados

O objetivo da nossa avaliação é comparar o desempenho de várias técnicas de classificação com diferentes métricas, após utilização de diversas estratégias de reamostragem, considerando que os dados obtidos são desbalanceados.

Escolhemos 5 classificadores (Regressão Logística, Random Forest, Gradient Boosting, XGBoosting e K-vizinhos mais próximos) e utilizamos 4 diferentes estratégias de reamostragem (Nearmiss, SMOTE, Tomek Links, SMOTE-NN).

Inicialmente, trabalhamos com o total de 28 variáveis. No entanto, após a aplicação de uma técnica de Feature Selection para redução da dimensionalidade, selecionando as 15 mais relevantes para uso na construção do modelo. A Figura 1 mostra as correlações para os 15 principais atributos.

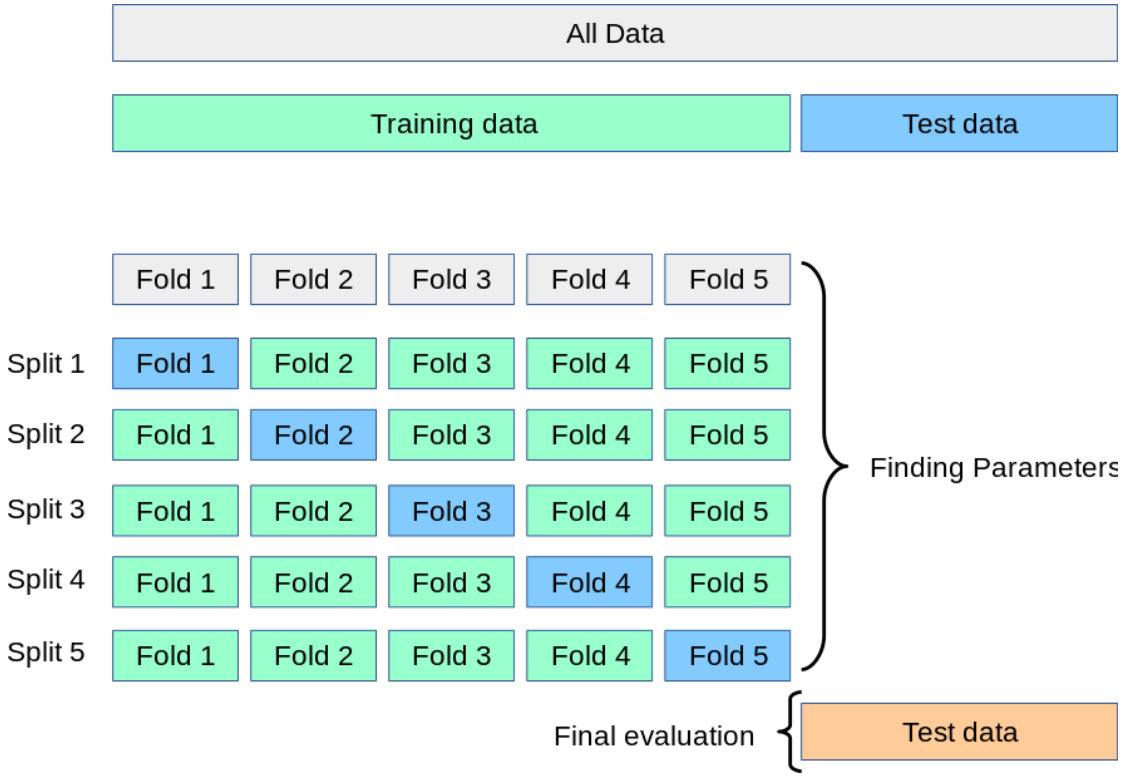
Figura 1 – Correlação entre as variáveis da Base



O total de amostras foi dividido em um conjunto de treino e teste, representando 70% e 30%, respectivamente. Aplicamos a técnica de balanceamento apenas no conjunto de treinamento, por meio de reamostragem, validando o conjunto de teste de forma desbalanceada.

Realizamos validação cruzada para a escolha dos hiperparâmetros com GridSearch (cv=5), no qual os dados de treino foram divididos em 5 partes *(folds)* e o modelo é treinado várias vezes utilizando 4 delas, sempre deixando 1 para teste, conforme ilustrado na Figura 2.

Figura 2 – Scikit-learn: Cross Validation



Fonte: scikit-learn.org

Os resultados são apresentados a seguir:

4.1 Avaliação dos modelos com variável dependente 30 dias.

Em um primeiro momento, utilizamos os contratos que apresentaram atrasos maiores que 30 dias como a variável dependente. Após o pré-processamento dos dados, iniciamos a análise com uma amostra de 29795 casos adimplentes e 1852 inadimplentes.

A Tabela 6 lista os resultados de classificação dos algorítimos e técnicas de reamostragem examinadas. As melhores pontuações estão em negrito.

No geral, Random Forest parece ser o melhor classificador em termos de Acurácia, Precision, G-Mean e AUC, 97,15%, 77,31%, 95,70% e 95,7%, respectivamente, pois obteve o mais alto desempenho. Enquanto XGBoosting pontuou melhor em F1 e Recall, obtendo 75,31% e 99,28%. Os números do Gradient Boosting são muito semelhantes aos apresentados por Random Forest e XGBoosting. O classificador que menos se destacou foi KNN, uma vez que não obteve pontuação superior em nenhum dos testes realizados.

Observamos que as técnicas de reamostragem que mais contribuíram para aumento do desempenho foram TomekLinks, imprimindo melhores resultados nas métricas de

Acurácia e F1, enquanto SMOTEENN trouxe melhor resultado para AUC e G-Mean. NearMiss, por sua vez, apresentou o melhor Recall. Em termos gerais, a eficácia dos diferentes métodos de subamostragem depende da medida usada para medir o desempenho.

A combinação que apresentou maiores níveis de Precision (Amostra Desbalanceada + Random Forest) teve maior índice de acerto na classe majoritária, embora tenha tido também o pior desempenho na previsão da classe minoritária.

Tabela 6 – Métricas de Avaliação (30 Dias)

		Random Forest	Gradient Boosting	XGBoosting	Regressão Logística	KNN
Accuracy	**NEARMISS**	83.97%	82.81%	91.76%	94.44%	89.34%
Accuracy	**SMOTEENN**	94.94%	95.03%	94.98%	94.59%	94.08%
Accuracy	**TomekLinks**	**97.15%**	97.07%	97.07%	95.0%	96.18%
Accuracy	**SMOTE**	96.13%	96.29%	95.48%	96.3%	94.94%
Accuracy	**Sem Balanceamento**	97.14%	97.07%	97.07%	94.92%	96.18%
AUC	**NEARMISS**	0.909	0.903	0.953	0.951	0.937
AUC	**SMOTEENN**	**0.957**	0.954	0.953	0.953	0.94
AUC	**TomekLinks**	0.858	0.857	0.873	0.602	0.803
AUC	**SMOTE**	0.91	0.87	0.894	0.846	0.888
AUC	**Sem Balanceamento**	0.855	0.857	0.873	0.594	0.803
F1 Score	**NEARMISS**	41.91%	40.22%	58.53%	66.88%	51.99%
F1 Score	**SMOTEENN**	69.11%	69.31%	69.05%	67.51%	65.0%
F1 Score	**TomekLinks**	74.98%	74.45%	**75.31%**	32.62%	65.66%
F1 Score	**SMOTE**	72.09%	70.72%	68.15%	69.34%	65.47%
F1 Score	**Sem Balanceamento**	74.72	74.45	75.31	30.74	65.66
G-mean	**NearMiss**	90,55%	89,88%	95,20%	95,10%	89,34%
G-mean	**SMOTE**	90,84%	86,34%	89,15%	83,58%	88,52%
G-mean	**SMOTEENN**	**95,70%**	95,41%	95,30%	95,26%	93,98%
G-mean	**TomekLinks**	84,87%	84,73%	86,61%	45,39	78,31%
G-Mean	**Sem Balanceamento**	84,46%	84,73%	86,61%	43,78%	78,31%
Precision	**NEARMISS**	26.6%	25.25%	41.5%	51.35%	35.31%
Precision	**SMOTEENN**	53.81%	54.28%	54.01%	52.05%	49.71%
Precision	**TomekLinks**	77.04%	76.13%	74.39%	77.18%	69.26%
Precision	**SMOTE**	62.45%	65.79%	58.03%	67.4%	54.56%
Precision	**Sem Balanceamento**	**77.31%**	76.13%	74.39%	76.43%	69.26%
Recall	**NEARMISS**	98.74%	98.74%	**99.28%**	95.86%	98.56%
Recall	**SMOTEENN**	96.58%	95.86%	95.68%	96.04%	93.88%
Recall	**TomekLinks**	73.02%	72.84%	76.26%	20.68%	62.41%
Recall	**SMOTE**	85.25%	76.44%	82.55%	71.4%	81.83%
Recall	**Sem Balanceamento**	72.3%	72.84%	76.26%	19.24%	62.41%

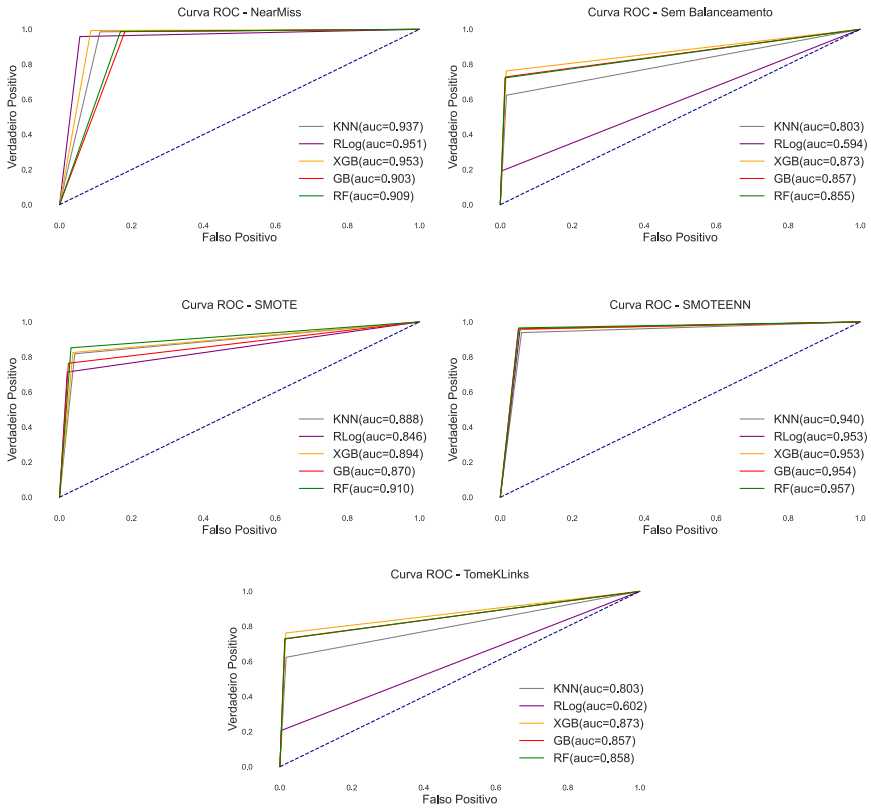
Fonte: Elaborado pelo Autor

De acordo com Namvar et al. (2018), G-Mean e AUC são as medidas mais eficazes para avaliar os resultados de classificação quando trabalhamos com classes desequilibradas.

Neste sentido, os melhores resultados foram alcançados utilizando Random Forest, após aplicação da técnica de reamostragem SMOTEENN, que apresentou 95,70% para ambas as métricas consideradas mais eficazes. A Figura 3 demonstra o benchmarking das curvas ROC, ratificando o desempenho das técnicas mais bem sucedidas.

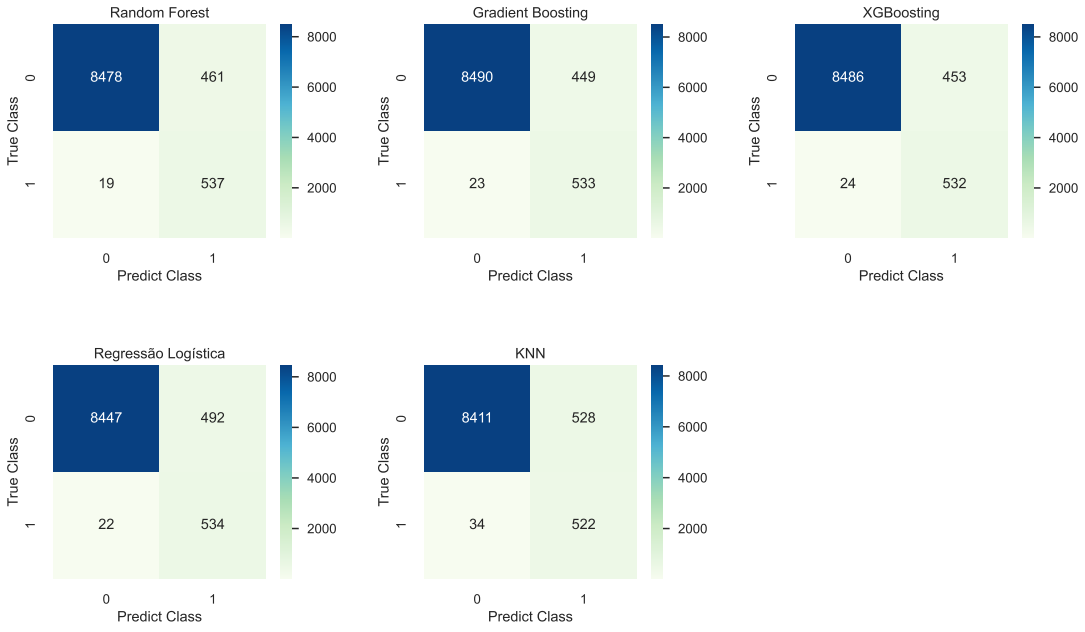
Na Figura 4 temos as matrizes de confusão dos classificadores, utilizando a técnica de reamostragem SMOTEENN que apresentou resultados mais eficazes. As demais matrizes, referentes às outras estratégias encontram-se no Apêndice A.

Figura 3 – Curva ROC (30 Dias)



Fonte: Elaborado pelo Autor

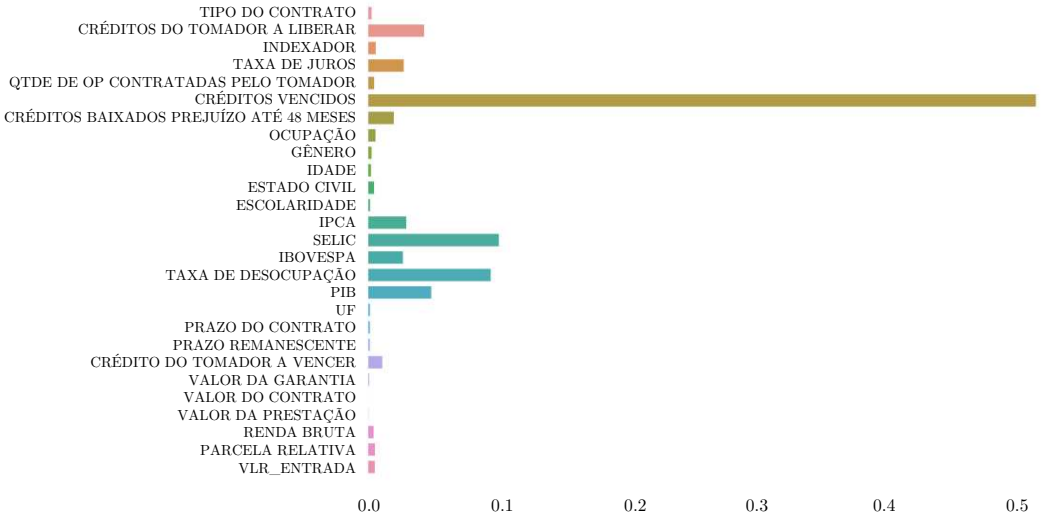
Figura 4 – Matriz de Confusão (SMOTEENN - 30 Dias)



Fonte: Elaborado pelo Autor

Considerando o modelo com resultados mais eficazes, procuramos demonstrar a importância de cada variável para o classificador por meio da Figura 5.

Figura 5 – Importância de cada variável



Fonte: Elaborado pelo Autor

4.2 Avaliação dos modelos com variável dependente 60 dias.

Para avaliar a performance dos modelos que utilizam como variável dependente *default* de 60 dias, utilizamos os mesmos classificadores e estratégias de balanceamento da variável anterior.

Neste caso, conforme observamos na Tabela 7, temos a Regressão Logística apresentando melhor desempenho em 4 das 6 métricas utilizadas, 3 delas com a técnica SMOTEENN, com 94,2% (AUC), 94,16% (G-Mean), 67,92% (Precision) e 94,98% (Recall). XGBoosting obteve melhor desempenho na Accuracy e f1, com 97,82% e 65,48%, respectivamente.

De forma geral, verificamos que todos os modelos tiveram uma redução de desempenho se comparados com os resultados obtidos na variável dependente de 30 dias. Este fato pode ser explicado porque houve uma piora no desbalanceamento da base, passando para 30651 adimplentes e 996 inadimplentes, o que traz mais dificuldade de aprendizado para o modelo.

Tabela 7 – Métricas de Avaliação (60 Dias)

		Random Forest	Gradient Boosting	XGBoosting	Regressão Logística	KNN
Accuracy	**NearMiss**	91.34%	80.95%	91.83%	96.09%	96.74%
Accuracy	**SMOTE**	97.49%	97.44%	97.55%	95.93%	95.53%
Accuracy	**TomekLinks**	97.7%	97.64%	**97.82%**	97.05%	96.98%
Accuracy	**SMOTEEN**	96.39%	96.64%	96.64%	93.4%	94.6%
Accuracy	**Sem Balanceamento**	97.7%	97.64%	97.78%	97.05%	96.98%
AUC	**NearMiss**	0.842	0.8	0.836	0.808	0.808
AUC	**SMOTE**	0.858	0.814	0.861	0.901	0.867
AUC	**TomekLinks**	0.762	0.786	0.791	0.559	0.706
AUC	**SMOTEEN**	0.917	0.918	0.924	**0.942**	0.901
AUC	**Sem Balanceamento**	0.762	0.786	0.79	0.559	0.706
F1 Score	**NearMiss**	35.78	20.7	36.61	50.99	55.2
F1 Score	**SMOTE**	64.79	61.24	**65.48**	56.53	52.14
F1 Score	**TomekLinks**	59.33	60.84	63.1	20.45	46.95
F1 Score	**SMOTEEN**	60.17	61.89	62.25	47.53	49.85
F1 Score	**Sem Balanceamento**	59.33%	60.84%	62.52%	20.45%	46.95%
G-mean	**NearMiss**	83,86%	79,96%	83,18%	79,17%	79,04%
G-mean	**SMOTE**	84,84%	79,53%	85,24%	89,92%	86,17%
G-mean	**SMOTEENN**	91,52%	91,64%	92,33%	**94,16%**	89,96%
G-mean	**TomekLinks**	72,61%	75,87%	76,58%	34,66%	64,76%
G-Mean	**Sem Balanceamento**	72,61%	75,87%	76,35%	34,66%	64,76%
Precision	**NearMiss**	23.34%	11.91%	24.22%	42.14%	48.6%
Precision	**SMOTE**	58.09%	58.54%	58.78%	42.61%	39.35%
Precision	**TomekLinks**	67.09%	63.74%	67.56%	**67.92%**	52.48%
Precision	**SMOTEEN**	46.09%	48.14%	48.17%	31.7%	35.22%
Precision	**Sem Balanceamento**	67.09%	63.74%	66.67%	67.92%	52.48%
Recall	**NearMiss**	76.59%	78.93%	74.92%	64.55%	63.88%
Recall	**SMOTE**	73.24%	64.21%	73.91%	83.95%	77.26%
Recall	**TomekLinks**	53.18%	58.19%	59.2%	12.04%	42.47%
Recall	**SMOTEEN**	86.62%	86.62%	87.96%	**94.98%**	85.28%
Recall	**Sem Balanceamento**	53.18%	58.19%	58.86%	12.04%	42.47%

Fonte: Elaborado pelo Autor

Novamente, a técnica SMOTEENN demonstrou maior eficácia na melhora dos resultados quando comparamos com a base desbalanceada e as demais técnicas de SMOTE, NearMiss e TomekLinks. A Figura 6 ilustra a superioridade do método por meio da Curva ROC.

Na Figura 7 temos as matrizes de confusão dos classificadores, utilizando a técnica de reamostragem SMOTEENN,que apresentou resultados mais eficazes. As demais matrizes, referentes às outras estratégias encontram-se no Apêndice B.

Figura 6 – Curva ROC (60 Dias)

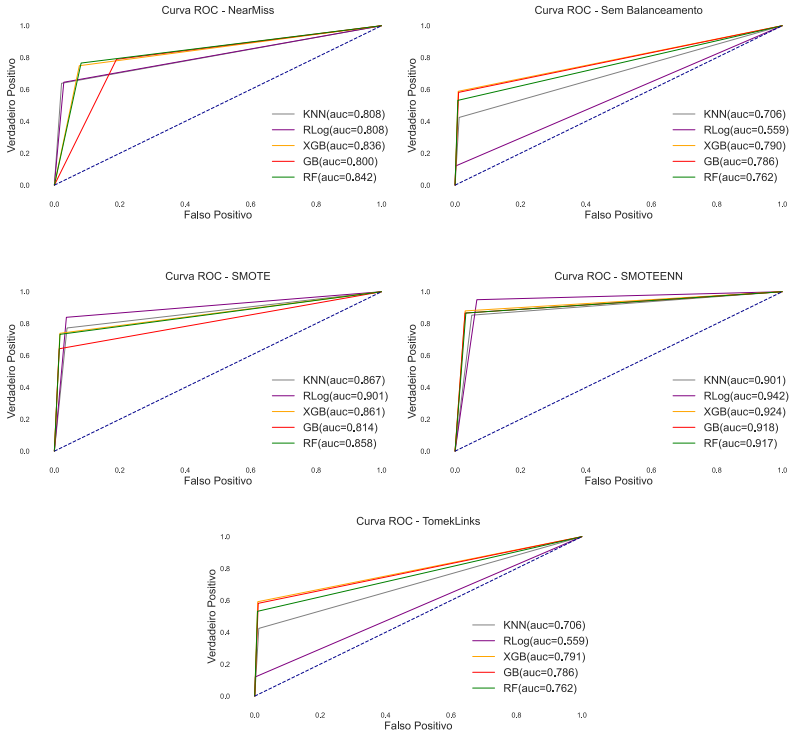
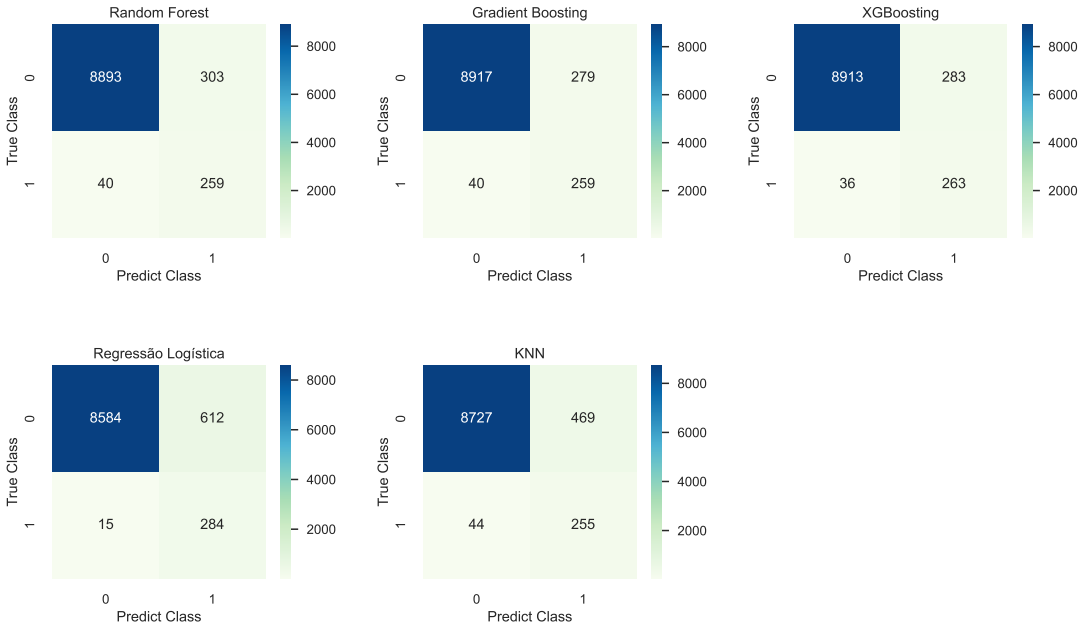


Figura 7 – Matriz de Confusão (SMOTEENN - 60 Dias)



Fonte: Elaborado pelo Autor

4.3 Avaliação dos modelos com variável dependente 90 dias.

Por fim, a Tabela 8 demonstra que, quando tratamos do modelo que tem como variável dependente *default* 90 dias, critério utilizando pelo BACEN, temos Random Forest com melhores resultados na Accuracy e Precision, 98,58% e 72,92%, respectivamente. Regressão Lógística tem os melhores resultados de 93,10%(AUC) e 91,08%(Recall), enquanto XGBoosting é superior no f1 com 52,98%. KNN apresenta um desempenho muito bom em uma das principais métricas utilizadas nesse estudo, pontuando com 95,92% de G-Mean.

Neste caso, também percebemos redução nos resultados após aumento do desbalanceamento, na qual tivemos 31124 adimplentes e 523 inadimplentes.Não tivemos a prevalência em termos de quantidade de resultados superiores em nenhuma das estratégias de balanceamento. No entanto, se considerarmos as métricas definidas por Namvar et al. (2018) como mais eficazes, temos um destaque para SMOTE no G-Mean e SMOTEEN na AUC.

Com relação à curva ROC, assim como as variáveis anteriores (30 e 60 dias), temos SMOTEENN com desempenho superior às demais, conforme Figura 8.

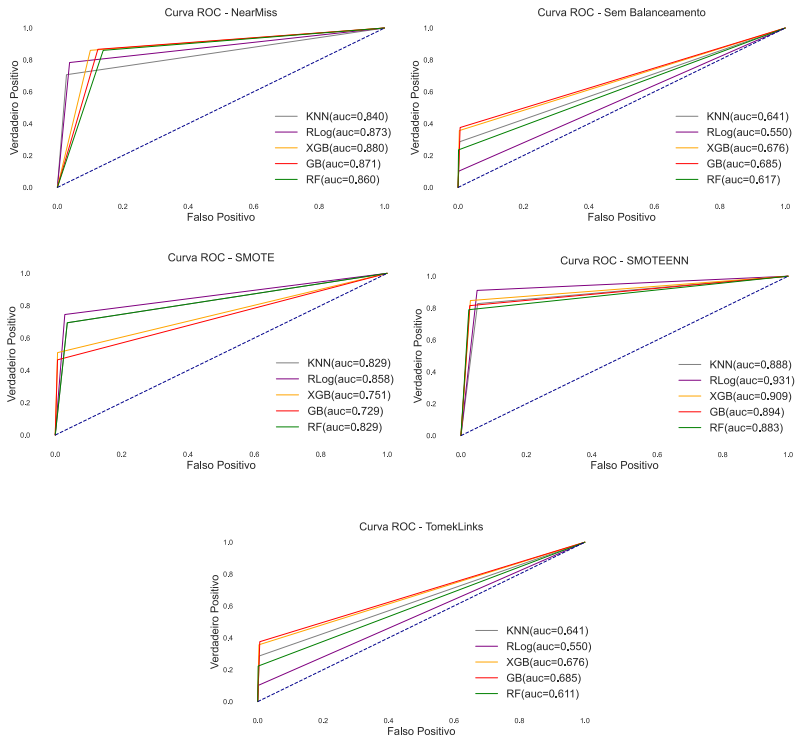
Tabela 8 – Métricas de Avaliação (90 Dias)

		Random Forest	Gradient Boosting	XGBoosting	Regressão Logística	KNN
Accuracy	**NearMiss**	85.98%	87.6%	89.86%	95.97%	96.78%
Accuracy	**TomekLinks**	**98.58%**	98.35%	98.56%	98.35%	98.34%
Accuracy	**SMOTEEN**	97.28%	97.04%	96.9%	95.0%	94.69%
Accuracy	**SMOTE**	**98.58%**	98.44%	98.5%	96.75%	95.92%
Accuracy	**Sem Balanceamento**	98.6%	98.35%	98.56%	98.35%	98.34%
AUC	**NearMiss**	0.86	0.871	0.88	0.873	0.84
AUC	**TomekLinks**	0.611	0.685	0.676	0.55	0.641
AUC	**SMOTEEN**	0.883	0.894	0.909	**0.931**	0.888
AUC	**SMOTE**	0.72	0.729	0.751	0.858	0.829
AUC	**Sem Balanceamento**	0.617	0.685	0.676	0.55	0.641
F1 Score	**NearMiss**	16.87	18.78	21.9	39.11	42.05
F1 Score	**TomekLinks**	34.14	42.91	44.98	16.93	36.29
F1 Score	**SMOTEEN**	49.01	47.67	47.5	37.59	34.04
F1 Score	**Sem Balanceamento**	35.75	42.91	44.98	16.93	36.29
F1 Score	**SMOTE**	50.91	49.66	**52.98**	43.09	36.03
G-mean	**NearMiss**	85,98%	87,12%	87,93	86,84%	82,90%
G-mean	**SMOTE**	66,60%	67,95%	71,13%	85,07%	**95,92%**
G-mean	**SMOTEENN**	87,79%	89,06%	90,69%	93,05%	88,64%
G-mean	**TomekLinks**	47,18%	61,10%	59,60%	31,89%	53,40%
G-Mean	**Sem Balanceamento**	48,51%	61,10%	59,60%	31,89%	53,40%
Precision	**NearMiss**	9.35%	10.53%	12.55%	26.06%	29.92%
Precision	**TomekLinks**	**72.92%**	50.0%	60.87%	50.0%	49.45%
Precision	**SMOTEEN**	35.53%	33.68%	33.0%	23.68%	21.42%
Precision	**SMOTE**	59.32%	53.28%	55.17%	30.31%	24.33%
Precision	**Sem Balanceamento**	74.0%	50.0%	60.87%	50.0%	49.45%
Recall	**NearMiss**	85.99%	86.62%	85.99%	78.34%	70.7%
Recall	**TomekLinks**	22.29%	37.58%	35.67%	10.19%	28.66%
Recall	**SMOTEEN**	78.98%	81.53%	84.71%	**91.08%**	82.8%
Recall	**SMOTE**	44.59%	46.5%	50.96%	74.52%	69.43%
Recall	**Sem Balanceamento**	23.57%	37.58%	35.67%	10.19%	28.66%

Fonte: Elaborado pelo Autor

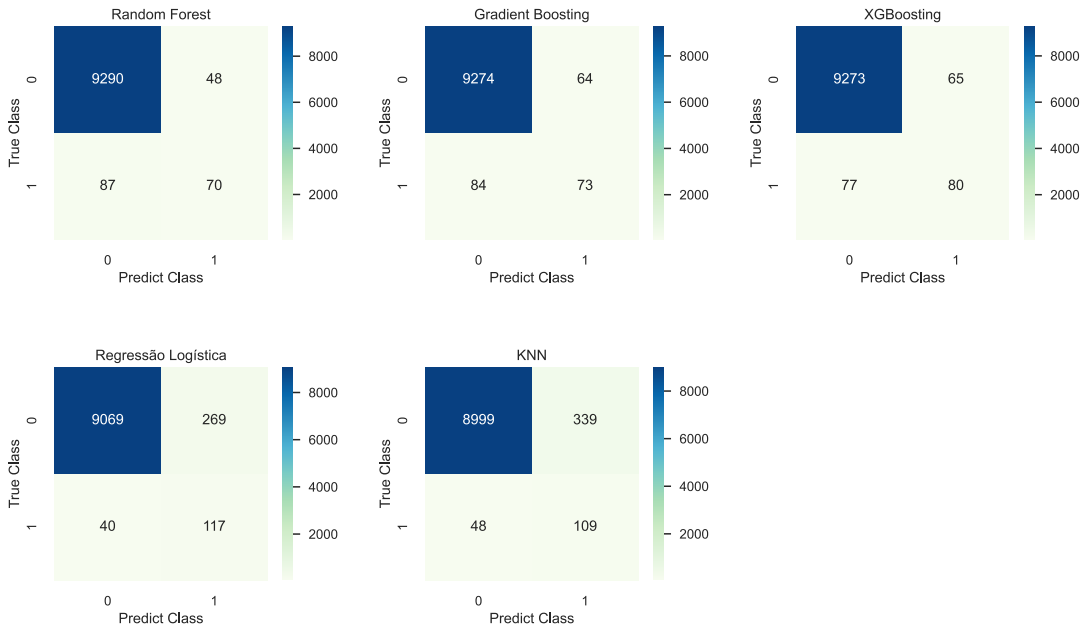
As Figuras 9 e 10 são apresentadas as matrizes de confusão para as técnicas que apresentaram melhor desempenho. As demais encontram-se no Apêndice C deste trabalho.

Figura 8 – Curva ROC (90 Dias)



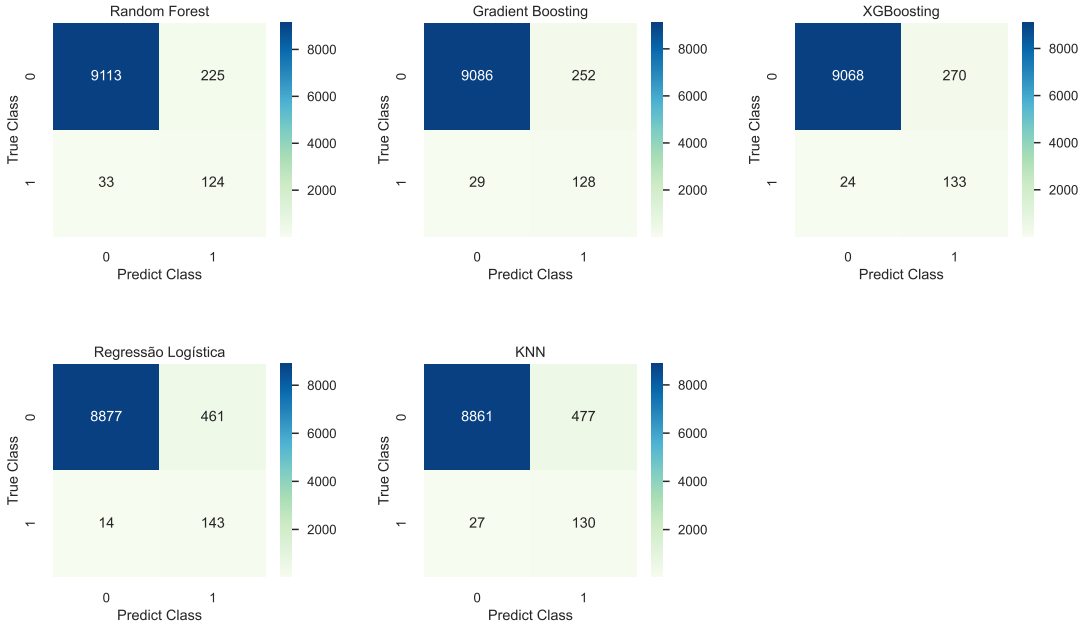
Fonte: Elaborado pelo Autor

Figura 9 – Matriz de Confusão (SMOTE - 90 Dias)



Fonte: Elaborado pelo Autor

Figura 10 – Matriz de Confusão (SMOTEENN - 90 Dias)



Fonte: Elaborado pelo Autor

5 Conclusões

Este estudo examinou algumas das principais técnicas de classificação e comparou o desempenho entre elas. Utilizamos alguns métodos mais tradicionais como a Regressão Logística e KNN, e outros que vem apresentando bons resultados em estudos mais recentes (Random Forest, Gradient Boosting e XGBoosting).

Avaliamos três cenários diferentes, utilizando variáveis dependentes de 30, 60 e 90 dias. Aplicamos técnicas de balanceamento SMOTE, SMOTEEN, TomekLinks e NearMisss, comparando com a aplicação do modelo sem balanceamento.

Após aplicação das técnicas na base selecionada, obtivemos os melhores resultados quando empregamos SMOTEENN nos três cenários. Os classificadores que mais se destacaram, foram Random Forest, Regressão Logística e KNN, quando utilizadas as variáveis dependentes de 30, 60 e 90 dias, respectivamente.

Os resultados apresentados demonstram a utilidade de técnicas de Machine Learning para avaliação do risco de crédito de clientes que já possuem empréstimo em andamento com a Instituição Financeira, possibilitando a identificação de potenciais operações de crédito que venham a inadimplir durante o seu curso.

Na avaliação da performance dos modelos, verificamos que, conforme a quantidade de dias da variável dependente aumentava e as classes iam se tornando menos balanceadas, a capacidade preditiva dos classificadores piorava.

Importante destacar que a introdução de variáveis macroeconômicas no modelo demonstrou-se relevante, considerando que, após a aplicação de técnica para seleção de variáveis, todas permaneceram.

Ressalta-se que todas as aplicações foram realizadas com a base de dados disponibilizada pela Instituição Financeira sem que pudéssemos avaliar a precisão das características do contrato e tomador, uma vez que a empresa disponibilizou as informações sem qualquer dado de identificação, demonstrando extrema preocupação em não ferir a Lei Geral de Proteção de Dados Pessoais. Nesse sentido, partimos do pressuposto que as informações eram fidedignas e correspondiam, de fato, ás operações de crédito disponibilizadas.

Sugere-se para estudos futuros, ampliar a quantidade de amostras utilizadas e inserir uma avaliação a respeito dos custos atrelados às classificações erradas de forma a direcionar a escolha da métrica de avaliação para aquela que traga mais vantagens à Instituição Financeira.

Referências

ABELLáN, J.; CASTELLANO, J. G. A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, v. 73, p. 1–10, 2017. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417416306947>. Citado na página 34.

BATISTA, G.; PRATI, R. C.; MONARD, M. C. *Um estudo do comportamento de vários métodos para balancear dados de treinamento de aprendizado de máquina*. 2004. 20–29 p. Disponível em: <https://doi.org/10.1145/1007730.1007735>. Citado na página 31.

BRACKE, P. et al. Machine learning explainability in finance: an application to default risk analysis. Bank of England Working Paper, 2019. Citado na página 32.

BROWN, I.; MUES, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, v. 39, n. 3, p. 3446–3453, 2012. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095741741101342X>. Citado na página 33.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939785>. Citado na página 32.

DUMITRESCU, E. I. et al. *Machine Learning for Credit Scoring: Improving Logistic Regression with Non-Linear Decision-Tree Effects*. [S.l.], 2022. Disponível em: <https://ideas.repec.org/p/hal/journl/hal-03331114.html>. Citado na página 32.

FIGLEWSKI, S.; FRYDMAN, H.; LIANG, W. Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *International Review of Economics Finance*, v. 21, n. 1, p. 87–105, 2012. ISSN 1059-0560. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1059056011000670>. Citado na página 26.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001. Citado na página 33.

FRIEDMAN, J. H. Stochastic gradient boosting. *Computational statistics & data analysis*, Elsevier, v. 38, n. 4, p. 367–378, 2002. Citado na página 33.

FSB. Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. *Relatório técnico, Conselho de Estabilidade Financeira*, 2017. Disponível em: <https://www.fsb.org/wp-content/uploads/P011117.pdf>. Citado na página 22.

JAPKOWICZ, N. Learning from imbalanced data sets: A comparison of various strategies. In: . [S.l.]: AAAI Press, 2000. p. 10–15. Citado na página 31.

LEE, Y. et al. A data mining approach to constructing probability of default scoring model. In: *Proceedings of 10th conference on information management and implementation*. [S.l.: s.n.], 2004. p. 1799–1813. Citado na página 26.

MA, X. et al. Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, Elsevier, v. 31, p. 24–39, 2018. Citado na página 33.

MALEDE, M. Determinants of commercial banks lending: Evidence from ethiopian commercial banks. *European Journal of Business and Management*, v. 6, n. 20, 2014. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.687.5785&rep=rep1&type=pdf>. Citado na página 21.

MALEKIPIRBAZARI, M.; AKSAKALLI, V. Risk assessment in social lending via random forests. *Expert Systems with Applications*, v. 42, n. 10, p. 4621–4631, 2015. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417415000937>. Citado 2 vezes nas páginas 22 e 32.

MOSCATO, V.; PICARIELLO, A.; SPERLí, G. A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, v. 165, p. 113986, 2021. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417420307636>. Citado na página 34.

MULLAINATHAN, S.; SPIESS, J. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, v. 31, n. 2, p. 87–106, May 2017. Disponível em: <https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87>. Citado na página 22.

NAMVAR, A. et al. Credit risk prediction in an imbalanced social lending environment. *International Journal of Computational Intelligence Systems*, v. 11, p. 925, 05 2018. Citado 6 vezes nas páginas 25, 31, 32, 34, 37 e 42.

OLSON, D. L.; DELEN, D.; MENG, Y. Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, v. 52, n. 2, p. 464–473, 2012. ISSN 0167-9236. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167923611001709>. Citado na página 22.

OZGUR et al. Machine learning approach to drivers of bank lending: evidence from an emerging economy. *Financial Innovation*, v. 7, n. 1, p. 1–29, 2021. Citado 2 vezes nas páginas 21 e 22.

PETROPOULOS, A. et al. A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. *IFC Bulletins chapters. In: Bank for International Settlements (ed) Are post-crisis statistical initiatives completed?*, 2019. Citado na página 33.

SEVEN, U.; YETKINER, H. Financial intermediation and economic growth: Does income matter? *Economic Systems*, v. 40, n. 1, p. 39–58, 2016. Citado na página 21.

STEENACKERS, A.; GOOVAERTS, M. A credit scoring model for personal loans. *Insurance: Mathematics and Economics*, v. 8, n. 1, p. 31–34, 1989. ISSN 0167-6687. Disponível em: <https://www.sciencedirect.com/science/article/pii/0167668789900449>. Citado na página 26.

SUN, J. et al. Imbalanced enterprise credit evaluation with dte-sbd: Decision tree ensemble based on smote and bagging with differentiated sampling rates. *Information Sciences*, v. 425, p. 76–91, 2018. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025517310083>. Citado na página 32.

TSAI, C.-F.; WU, J.-W. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, v. 34, n. 4, p. 2639–2649, 2008. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417407001558>. Citado na página 22.

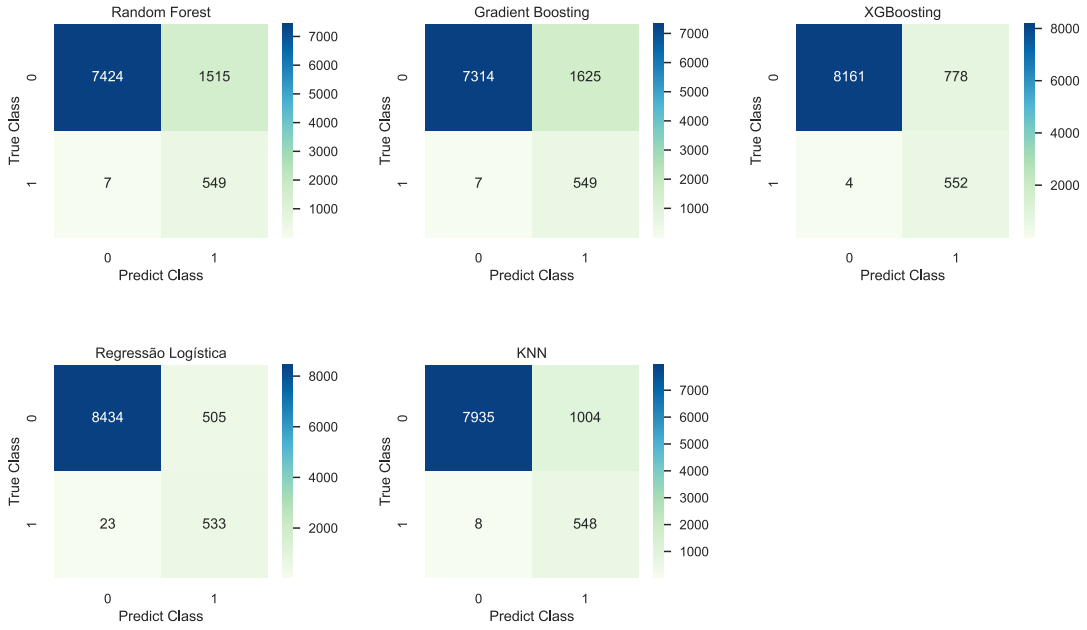
WANG, K. et al. Research on personal credit risk evaluation based on xgboost. *Procedia Computer Science*, v. 199, p. 1128–1135, 2022. ISSN 1877-0509. The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 2021): Developing Global Digital Economy after COVID-19. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050922001442>. Citado 2 vezes nas páginas 32 e 33.

YEH, I.-C.; LIEN, C. hui. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, v. 36, n. 2, Part 1, p. 2473–2480, 2009. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417407006719>. Citado 3 vezes nas páginas 22, 26 e 33.

ZHOU, L.; LAI, K. K.; YU, L. Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications*, v. 37, n. 1, p. 127–133, 2010. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417409004394>. Citado na página 22.

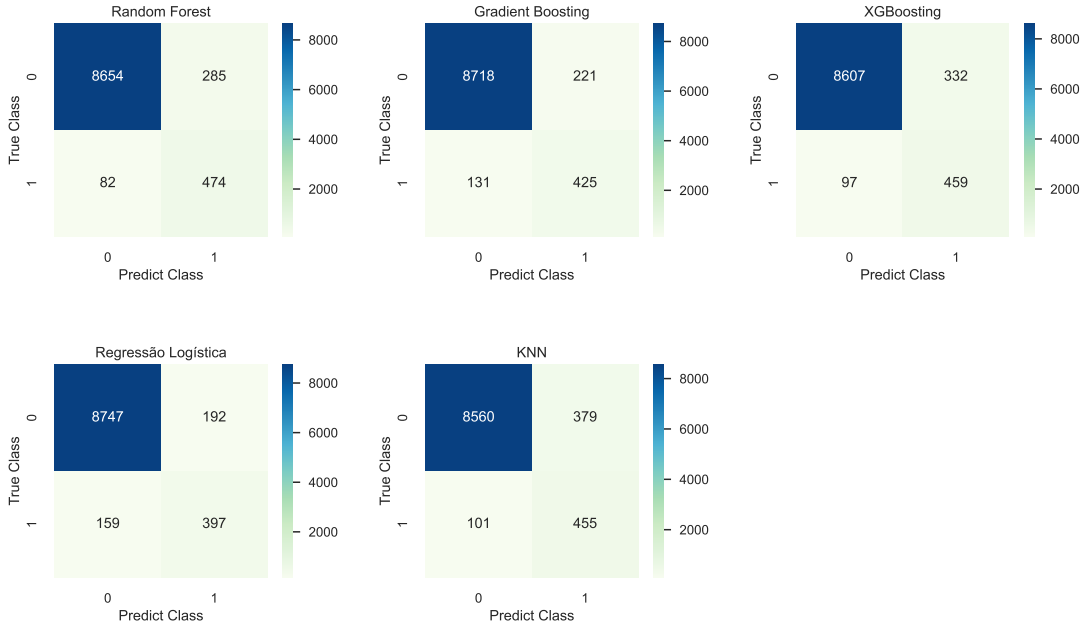
APÊNDICE A – Matris de Confusão - Variável 30 dias

Figura 11 – Matriz de Confusão (NearMiss - 30 Dias)



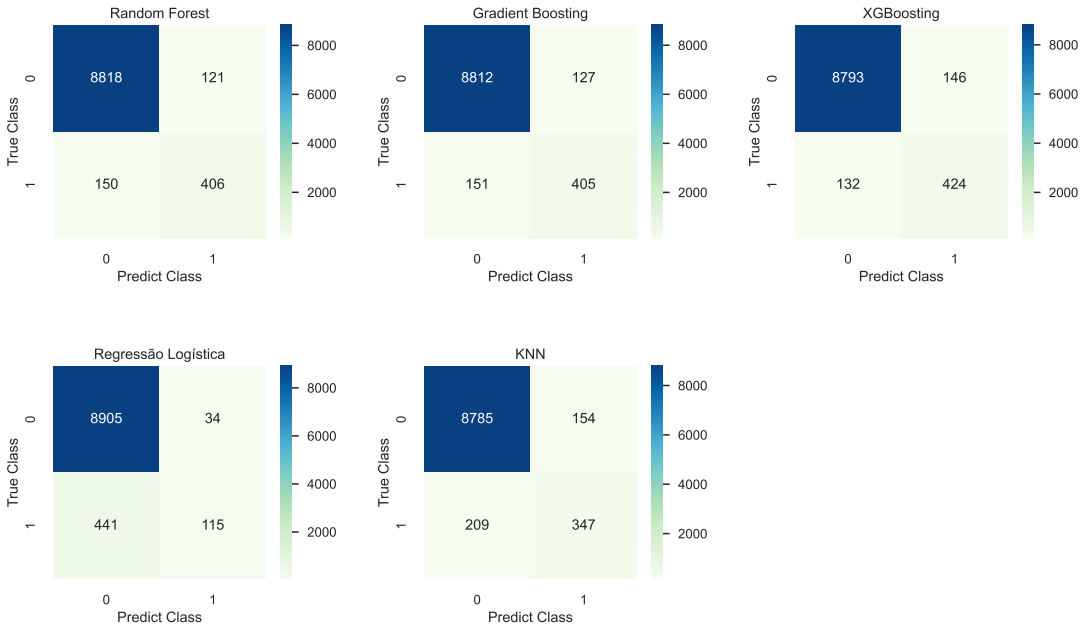
Fonte: Elaborado pelo Autor

Figura 12 – Matriz de Confusão (SMOTE - 30 Dias)



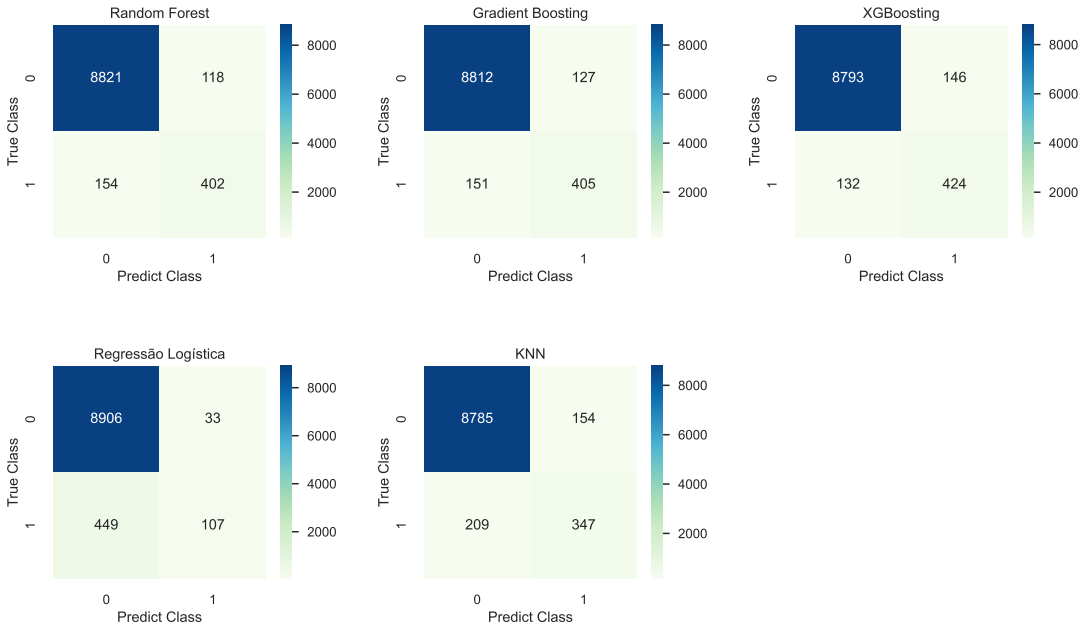
Fonte: Elaborado pelo Autor

Figura 13 – Matriz de Confusão (TomekLinks - 30 Dias)



Fonte: Elaborado pelo Autor

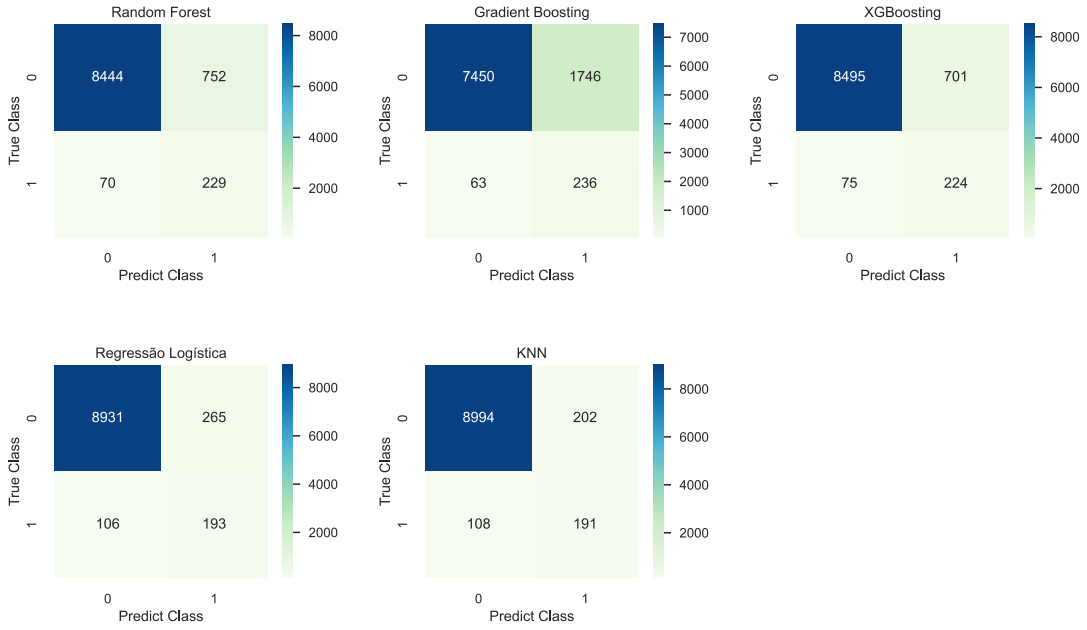
Figura 14 – Matriz de Confusão (Base Desbalanceada - 30 Dias)



Fonte: Elaborado pelo Autor

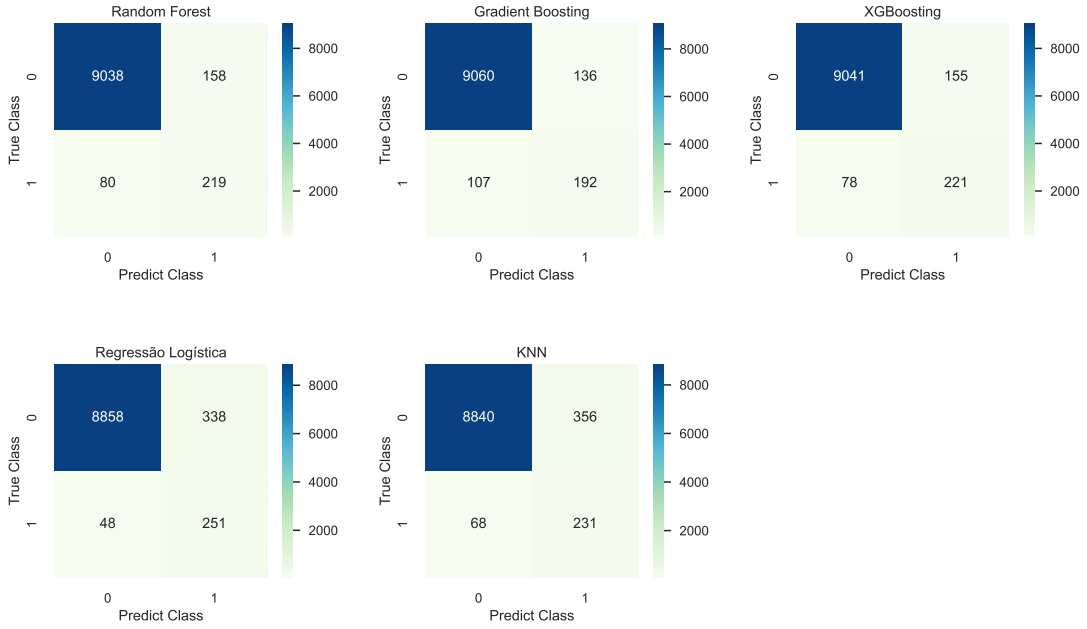
APÊNDICE B – Matris de Confusão - Variável 60 dias

Figura 15 – Matriz de Confusão (NearMiss - 60 Dias)



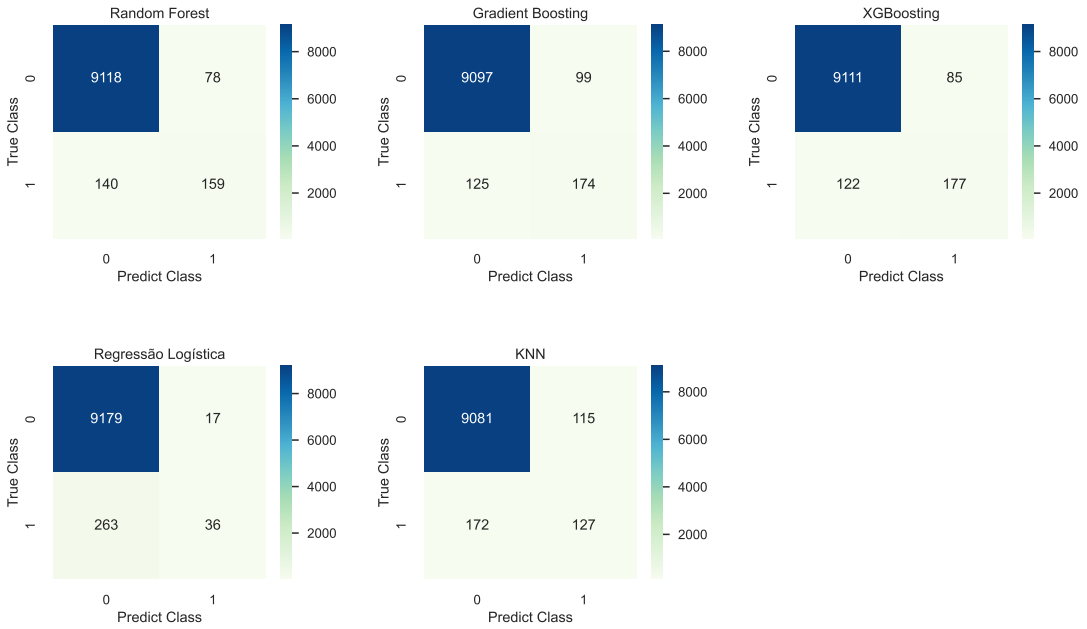
Fonte: Elaborado pelo Autor

Figura 16 – Matriz de Confusão (SMOTE - 60 Dias)



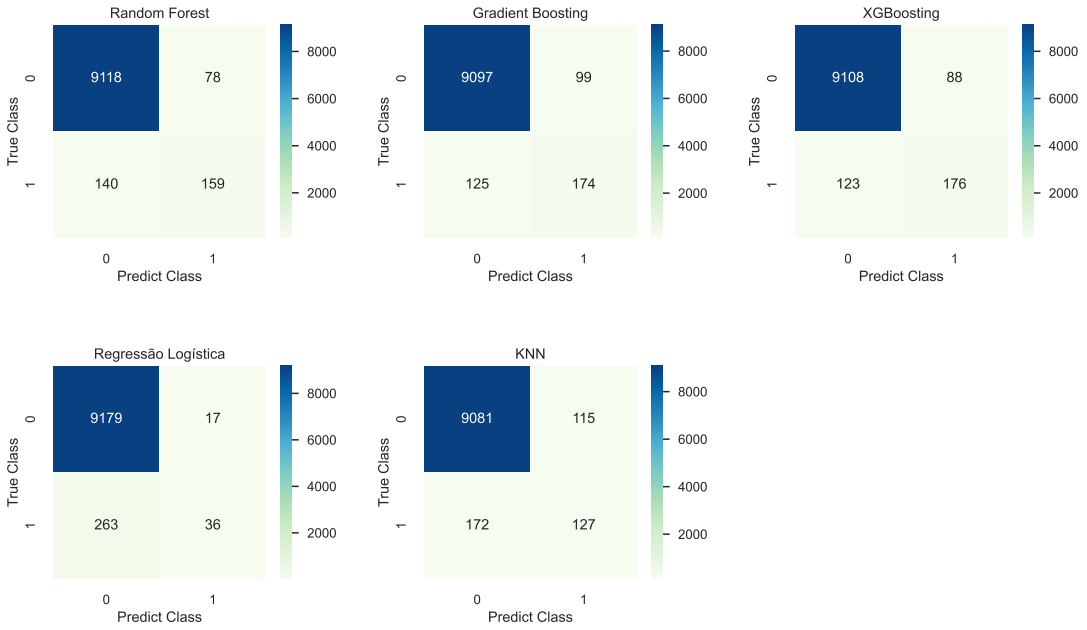
Fonte: Elaborado pelo Autor

Figura 17 – Matriz de Confusão (TomekLinks - 60 Dias)



Fonte: Elaborado pelo Autor

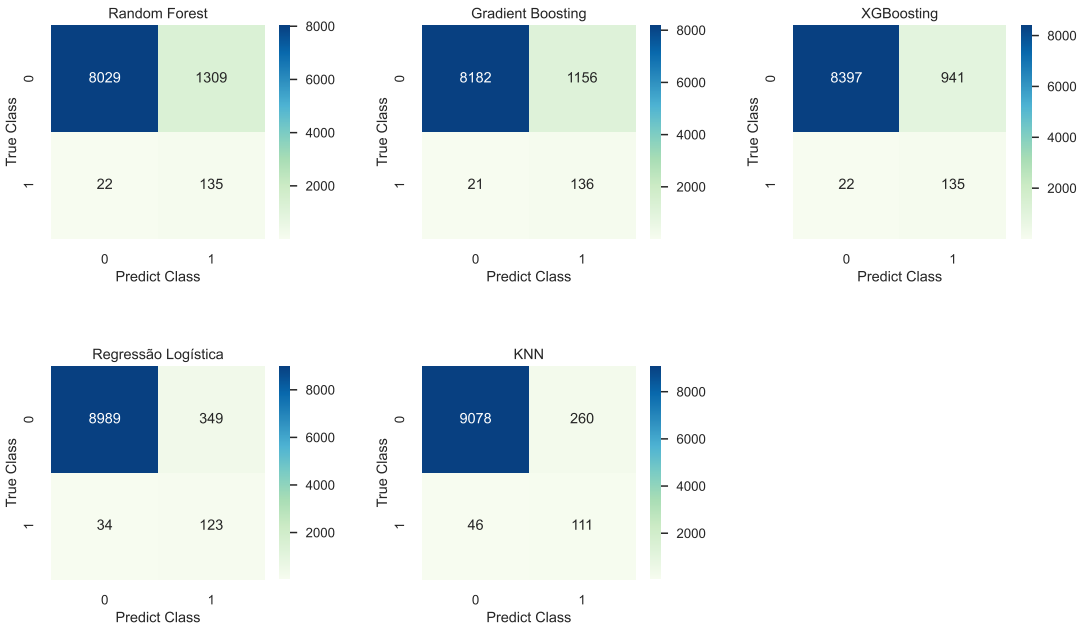
Figura 18 – Matriz de Confusão (Base Desbalanceada - 60 Dias)



Fonte: Elaborado pelo Autor

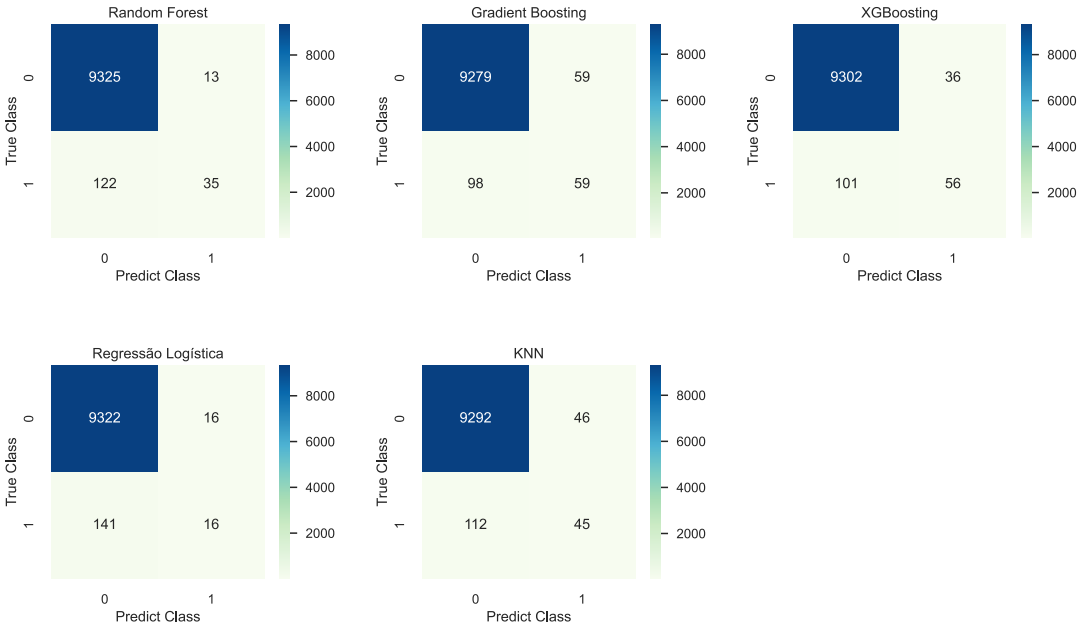
APÊNDICE C – Matris de Confusão - Variável 90 dias

Figura 19 – Matriz de Confusão (NearMiss - 90 Dias)



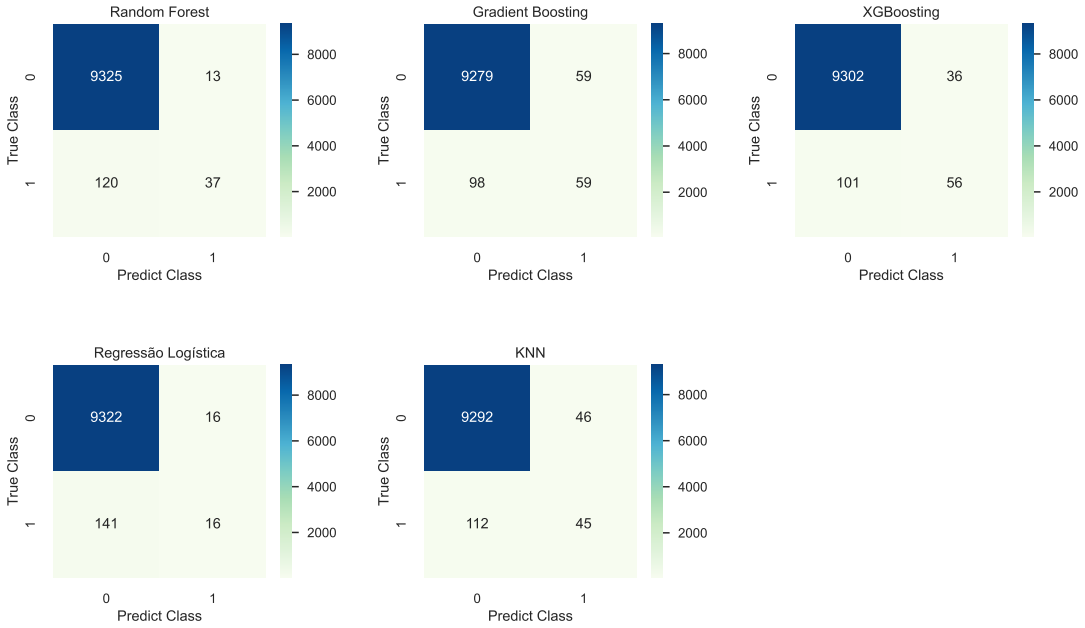
Fonte: Elaborado pelo Autor

Figura 20 – Matriz de Confusão (TomekLinks - 90 Dias)



Fonte: Elaborado pelo Autor

Figura 21 – Matriz de Confusão (Base Desbalanceada - 90 Dias)



Fonte: Elaborado pelo Autor