Predição de Intenção de Compra de Casas

Henrique Hott

Janeiro 2025

1 Análise Exploratória

Uma análise inicial simples demonstrou a presença de valores negativos na coluna de tempo e de valores faltantes em diversas colunas, essas linhas foram simplesmente removidas do conjuntos de dados.

Ademais a partir da distribuição das variáveis foi possível perceber um desbalanceamento na variável alvo.

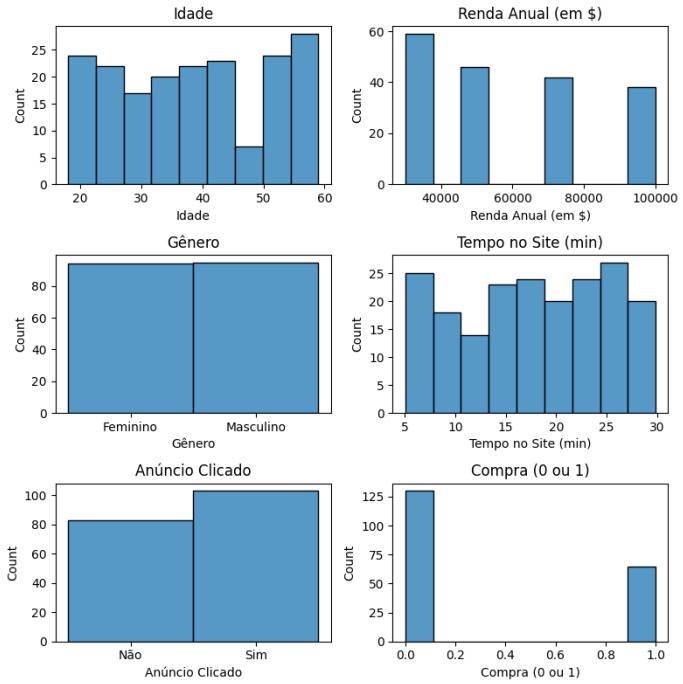

Figure 1: Distribuição das variáveis.

2 Tratamento, Combinação e Seleção de Variáveis

2.1 Tratamento

Todas a variáveis numéricas foram normalizadas entre 0 e 1, utilizando o maior valor presente, enquanto as categóricas foram transformadas em números inteiros.

2.2 Combinação

1. Renda Por Idade: Tenta distinguir usuários que tem mais poder de compra em relação aos seus pares na idade.

2. Renda Por Tempo no Site: Tenta distinguir e relacionar o comportamento de tempo gasto com poder de compra.

3. Tempo no Site por Idade: Tenta relacionar a idade e o interesse, considerado como tempo gasto, com a idade.

2.3 Seleção

A matriz de correlação nos da *insight's* valiosos sobre as relações entre as variáveis. Nela podemos ver duas das variáveis criadas tem mais correlação com o objetivo do que parte das que ja existiam no dataset.

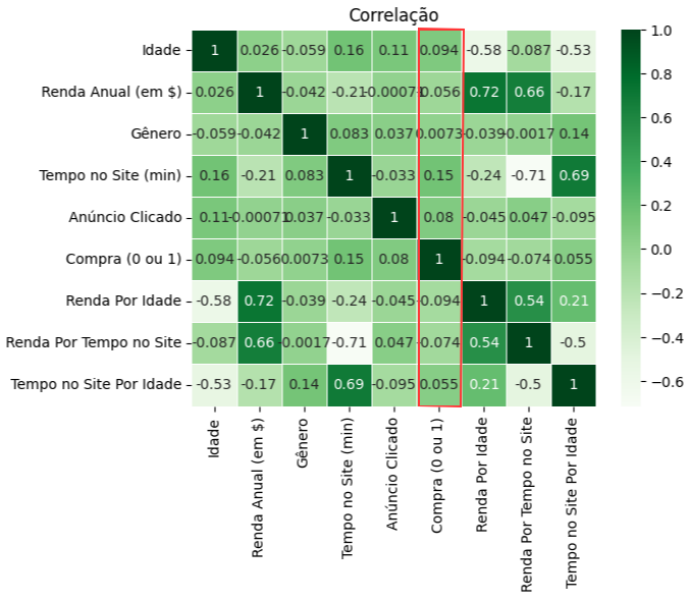

Figure 2: Distribuição das variáveis em relação à variável alvo.

Já a distribuição mostra demonstra quais variáveis pode trazer informações relevantes para o modelo por que seguem uma distribuição "bem definida".

Figure 3: Distribuição das variáveis em relação à variável alvo.

A partir do gráfico da distribuição das variáveis em relação a variável alvo e a partir da matriz de correlação foram escolhidas cinco - a quantidade foi definida para garantir a possibilidade de execução de um *grid search* em tempo viável - variáveis para serem utilizadas pelo modelo:

1. Idade
2. Tempo no Site (min)
3. Renda Por Idade
4. Renda Por Tempo no Site

5. Tempo no Site por Idade

3 Balanceamento do Conjunto de Dados

Foi observado que existe um desbalancemanto do conjunto de dados em relação a variável alvo, uma relação de 3:1. Considerando que a perca de uma venda tem valor significativo pois os valores no comercio de imóveis são mais altos, identificou-se a necessidade de contornar este problema.

Classe	**Quantidade**
Não Compra	119
Compra	31

Para isso foi utilizado e comparado duas estratégias:

1. Realizar o rebalanceamento dos dados a partir da tecnica de sobreamostrage aleatória.

2. Ajustar os pesos de maneira inversamente proporcional as frequências das classes dentro do classificador.

4 Resultados

4.1 Métricas

A partir da natureza do problema, que é o sucesso de uma possível venda de um imóvel, é importante definir um objetivo: Diminuir o número que vezes que um potencial comprador deixa de realizar a compra. Ou seja, é necessário escolher um modelo que tenha um bom *F1* balanceado entre *recall* e *precision*.

A fim de diminuir o efeito da aleatoriedade nos resultados foi realizado uma busca entre os hiperparametros utilizando a téninca de *GridSearch* em conjunto da utilização de *Stratified Cross Fold Validation* com cinco *folds* para escolher o melhor conjunto de parâmetros.

Classificador	**Subamostragem**	**Peso por Classe**
Logistic Regression	0.56	0.55
SVC	0.52	0.53
Decision Tree	0.66	0.55
Random Forest	0.63	0.62

4.2 Interpretação dos Modelos

4.2.1 SVC

Os parâmetro encontrados foram:

```
{
    "C": 10
    "kernel" : "sigmoid",
    "random_state" : 42
}
```

Listing 1: Parâmetro encontrado para o classificado SVC.

Para análisar o resultado do SVC utilizamos as dependências parciais de cada variável de entrada em relação a variável alvo.

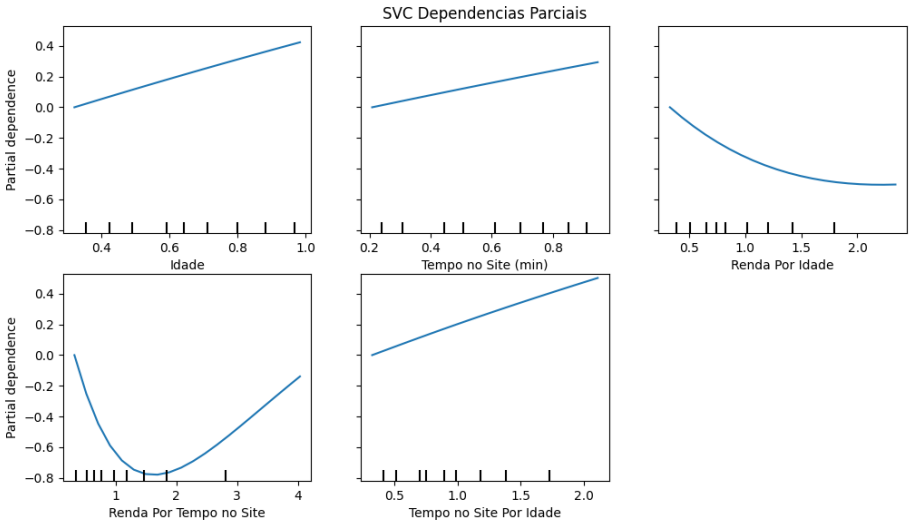

Figure 4: Dependências parciais

O principal fator a se considerar é a inclinação das curvas e o sinal da inclinação, quanto maior a magnitude da inclinação maior é a influência no modelo. Por exemplo, "Renda por Idade" é uma variável que, a medida que aumentam, a probabilidade de haver uma compra diminui.

4.2.2 Decision Tree

Os parâmetro encontrados foram:

```
{
    "criterion": "entropy"
    "max-depth" : "5",
    "random_state" : 42
}
```

Listing 2: Parâmetro encontrado para o classificador Decision Tree.

No ranqueamento das features da arvore é possível perceber que a Idade tem zero influencia na decisão tomada pela mesma.

	Feature	Importance
1	Tempo no Site (min)	0.520399
2	Renda Por Idade	0.222333
4	Tempo no Site Por Idade	0.155372
3	Renda Por Tempo no Site	0.101897
0	Idade	0.000000

Figure 5: Ranquamento das features da árvore de decisão.

5 Conclusão

A análise revelou que a variável "Tempo no Site (min)" desempenha um papel crucial no aumento das probabilidades de conversão em compras. Isso sugere que investir em estratégias que promovam o engajamento dos usuários, como a otimização da experiência de navegação, ferramentas de busca mais eficazes e estímulos para avaliação de produtos, pode ser determinante para aumentar as taxas de conversão.

Além disso, ambas as abordagens demonstraram que as variáveis criadas a partir da combinação de outras (como "Idade por Tempo no Site" e "Renda por Idade") tiveram um impacto positivo nos modelos. Mesmo em casos onde variáveis isoladas, como a "Idade", apresentaram baixa relevância, a interação entre variáveis contribuiu significativamente para informar os modelos, destacando o valor da engenharia de atributos.

Por fim, futuras iterações podem explorar outras combinações de variáveis ou diferentes estratégias de criação de novos atributos para aumentar a eficácia dos modelos. Também seria interessante testar um modelo baseado em métodos de votação (ensemble), como o Voting Classifier, que combina as previsões de múltiplos algoritmos para potencializar o desempenho geral.