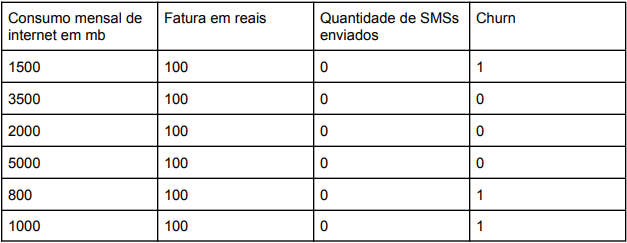
**FEATURE SELECTION COM FEATURE-ENGINE**

Contextualizando, vamos supor que uma empresa de telefonia queira prever o churn de seus clientes para oferecer promoções para aqueles que possivelmente darão. Para isso, a empresa conseguiu fazer um dataset com 200 features, mas para fazer uma análise exploratória com essa quantidade de colunas não é nada simples, pegar as correlações mais fortes apenas também pode ser impreciso, como prosseguir?

Antes, precisa-se ser explicado o conceito de variância nos dados:



Quanto maior uma coluna varia, maior é a sua variância, pelo dataset acima podemos notar que existem 2 features com variância igual a 0. As features com pouca variância são, via de regra, pouco valorosas para a predição, uma vez que ela não consegue diferenciar bem os possíveis targets, pois ela tem o mesmo valor independente de qual seja esse target.

Existem diversas formas de agir diante desse cenário onde temos um excesso de features:

1. Escolher as features uma a uma, analisando quais têm uma variância aceitável.
2. Remover as features com maior correlação, uma vez que, se A e B variam da mesma forma, então a relação delas com o target são muito parecidas, ou seja, ter A e B como features é redundante.
3. Analisar quais são as features mais relacionadas com o target(features importances), se A tem uma correlação com o target mais forte do que C, provavelmente A vai ser mais importante para essa predição.

Primeiramente é importante deixar claro que não existe uma forma de resolver todos os problemas, tem que ir testando, testar usar cada uma, usar mais de uma(remover as features constantes e depois as de forte correlações, por exemplo).