Este documento é um passo a passo de como foram efetuadas as análises e como podem ser reproduzi-las.

Introduzindo a respeito, o repositório de Data-science está particionado em:

- Bases:
    - Armazena a base de dados original e suas separações
- Modelo:
    - Guarda o melhor modelo, gerado pelo notebook “testeDeModelos”
- Notebooks:
    - Armazenam todos os notebooks utilizados para análise

Primeiramente nós fizemos a plotagem de boxplots de todos os atributos no notebook “analiseAtributosBoxplot”, com isso conseguimos perceber que há muitos outliers em alguns atributos, além disso também foi possível perceber, que alguns atributos como o V17, V12, V14 e V10 possuem medianas muito diferente entre as classes o que nós dá a percepção que estes 4 e mais alguns atributos com comportamento parecido são bons para a classificação das classes.

Porém para não tomar ações precipitadas, vamos analisar mais dois tipos de plot, no caso iremos ver o Violin plot e um Scatter plot no notebook “extracao_de_atributos”, nele novamente confirmamos o que já sabíamos em relação aos atributos mais importantes, porém graças ao plot de violino, conseguimos ter um embasamento melhor para saber quais atributos devem realmente ser levados em consideração.

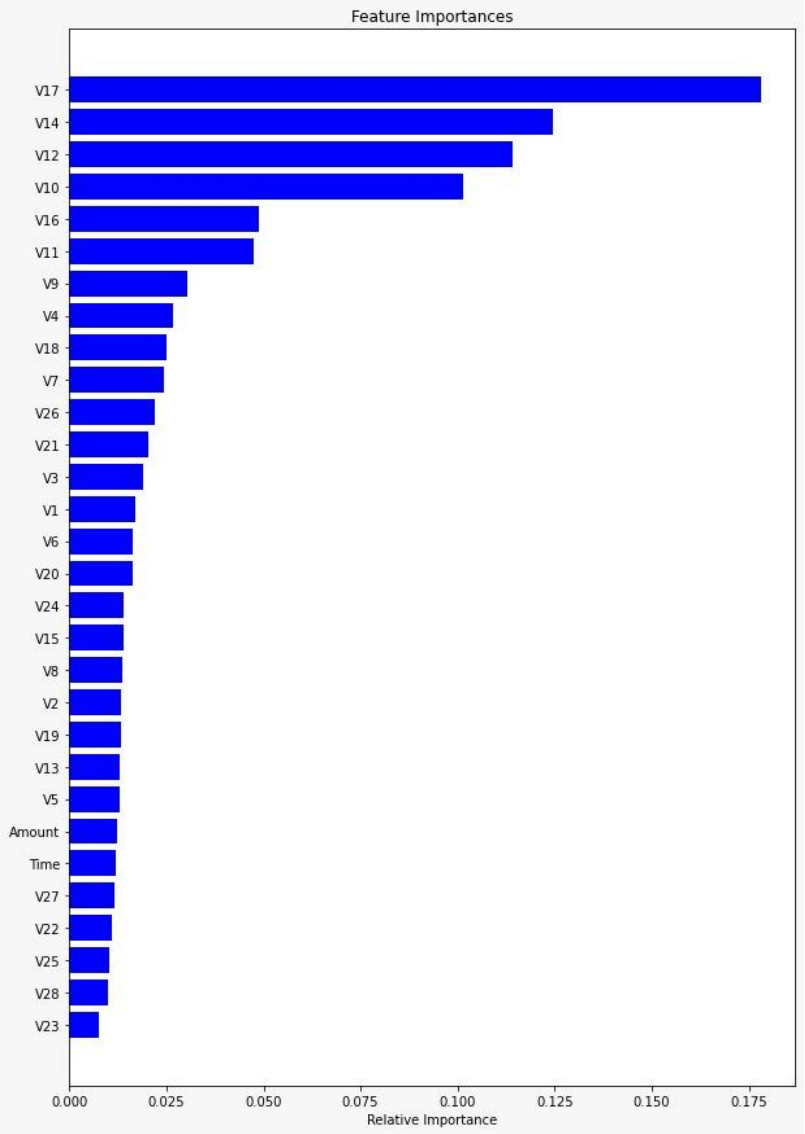
Com base nos dois notebooks já citados escolhemos os seguintes atributos para usar durante o treinamento do modelo, pois acreditamos que estes atributos facilitarão o treinamento do modelo: V3, V4, V5, V7, V9, V10, V11, V12, V14, V16, V17, V19, Amount

Agora com os melhores atributos adquiridos devemos conseguir arranjar a melhor separação entre: treino, teste e validação, para isso utilizamos da técnica de Nested cross validation, e esta separação está localizada no notebook: “separacaoBases”, a estratégia consiste em um cross validation externo para separação do teste e do treino geral, um cross validation interno que pega o treino geral e o divide em treino e validação. Com essa estratégia fomos capazes de separar as bases de forma que conseguimos maximizar o resultado obtido pelo treinamento dos modelos, as bases foram salvas na pasta “Bases”, antes da separação alguns dos modelos estavam dependendo de sorte para obter um bom resultado, visto que a separação era aleatória.

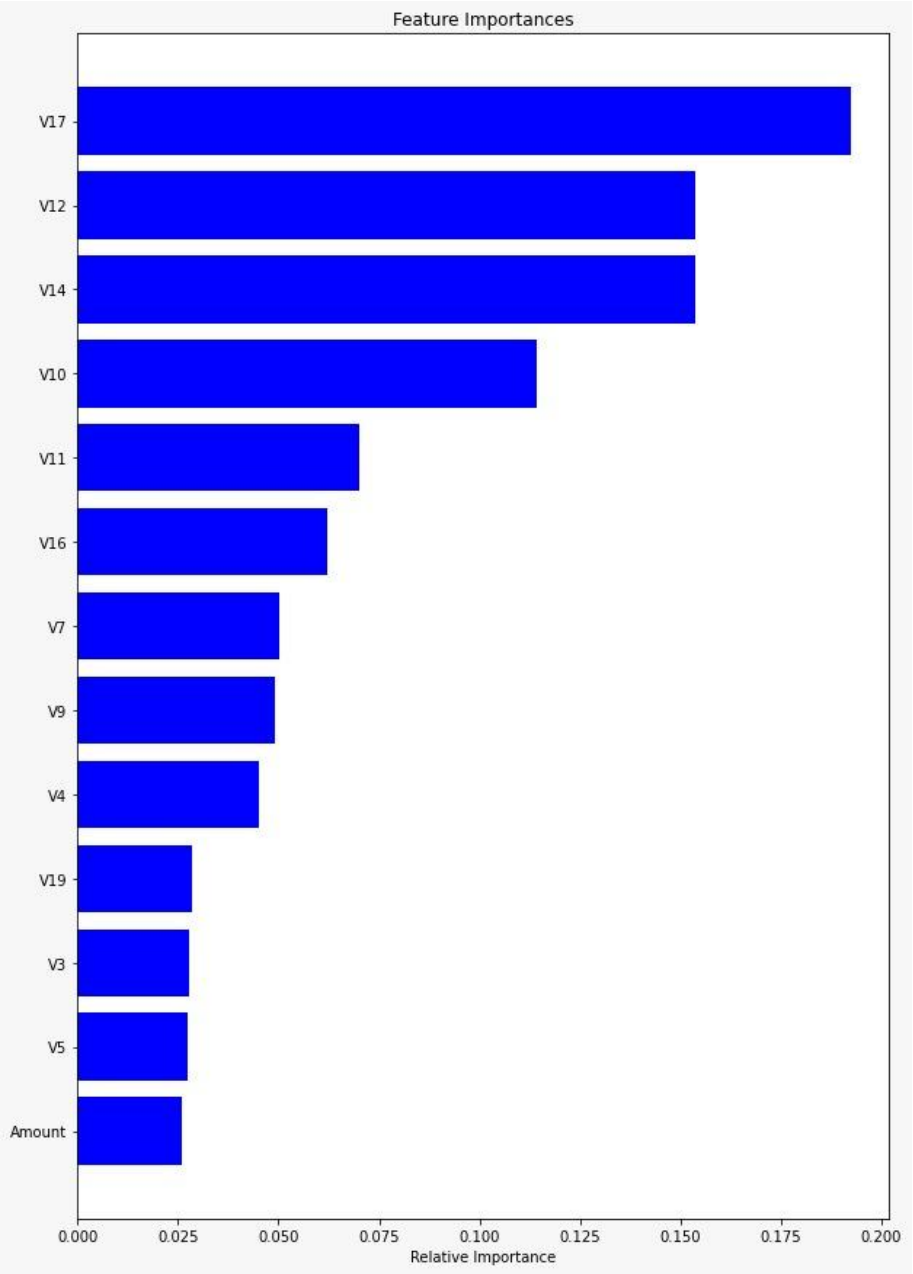
Agora com as bases separadas, e os melhores atributos escolhidos, fizemos uma bateria de testes em alguns modelos como: árvore de decisão, random forest, regressão logística, Knn vizinhos, entre outros, todos estes modelos estão no notebook: “testeDeModelos” assim como seus resultados, no fim chegamos a conclusão que o melhor classificador encontrado foi o random forest, conseguimos

treinar o modelo até uma precision/recall de 0.93, o que segundo alguns artigos que lemos a respeito da base é a mesma precision/recalls já encontrada por outros pesquisadores.

Como confirmação adicional dos atributos, após acharmos o melhor modelo e melhor separação rodamos o treinamento com todos os atributos e extraímos os melhores atributos para o treinamento da random forest, assim tendo a confirmação que os atributos selecionados são realmente importantes para o treinamento do modelo



Importância de todos os atributos.



Importância dos atributos selecionados.