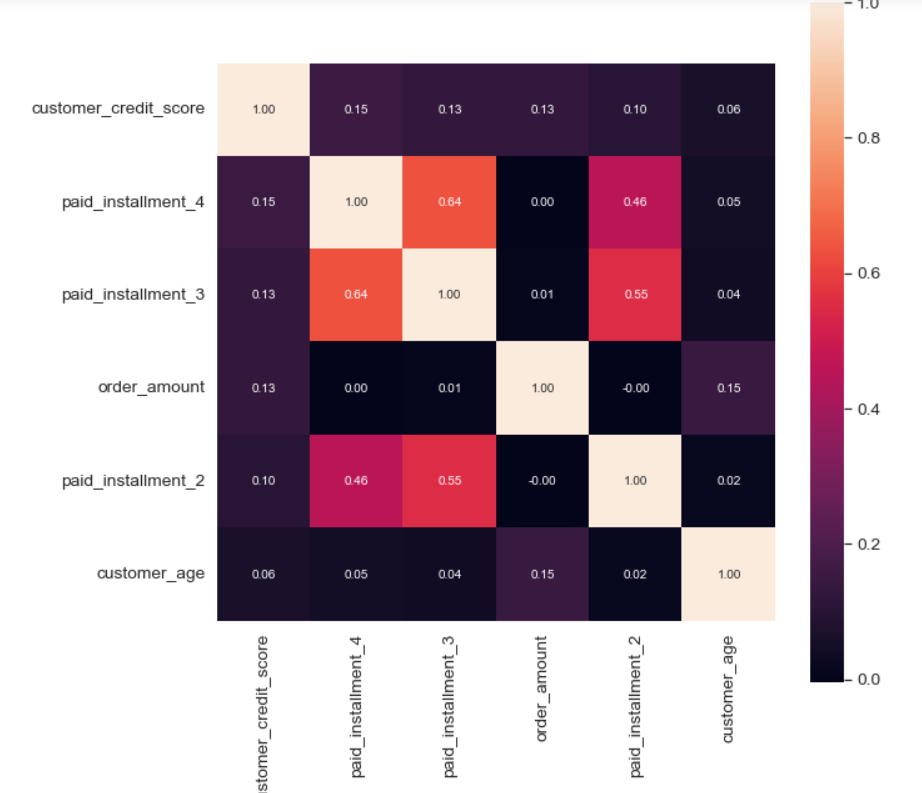
**Questions to explore:**

**Which features show strong correlation with a customer’s likelihood of paying back instalments?**

From the given data it is very difficult to get a good correlation matrix. Although below diagram shows a correlation matrix between some important features.



As we can see from the diagram, only few features have good correlation like paid\_installment\_3 has a good correlation with paid\_installment\_4. So, we can say that if the customer is paying the third instalment, then it is likely that he will pay the fourth instalment too.

**Which features should be discarded? Why?**

We can discard “approved\_for\_installments” as it has same value throughout and also it will not do any help to your model to differentiate between two different labels while on the other hand, it can even negatively affect your model by creating a bias in the data.

**What surprised you about the results/trends observed in the data?**

It is surprising to get accuracy of logistic regression, Random Forest & MLP Classifier as 1. Also, by plotting the confusion matrix we are getting no FP & FN. FP: Predicted values incorrectly predicted an actual positive. i.e., Negative values predicted as positive and FN: False Negative: Positive values predicted as negative.

**What additional data would you like to see that might help build a better instalment approval classifier?**

First, the data set is pretty small for a model to learn and give results. For getting a better result we need good volume of data set with features such as income of the customers, previous loan history of the customers, Balance etc.

**What would be your next steps enrich your training/build that can be used to make this classifier a real-time decisioning engine?**

To make it a real time, we can use the AutoML. But for Auto ML we need to have proper features in our data set and also the volume of data should be enough to train and learn the model properly.