**Limitations**

In the preprocessing part, we imputed missing values based on average and common values, which may introduce some errors into the data. Similarly, the manually-engineered features may introduce bias into the dataset. Furthermore, with 90 variables expanded from 20 original variables, the chance of overfitting also increased.

For the modeling part, logistic regression is built on two basic assumptions that the data observations are independently distributed, and that the predictive variables also should not be highly correlated. The first assumption depends on data collection schema, which is beyond our control in this case. As we see from the correlation heatmap in **Figure 3**, there are 2 pairs of highly correlated variables. In the modelling process, since we use regularization terms, the negative effect of this violation is mitigated.

First of all, random forest model is time-consuming to train and test, which may come into disadvantage in some loan evaluation scenarios when the portfolio has to be built in a timely fashion. Due to the best split mechanism of random forest, it tends to favor common occurrences. However, for default prediction, which is closer to outlier detection in nature, random forest tends to ignore those rare occurrences and outputs zero default probabilities. This leads to over-optimistic, risk-seeking portfolio decisions. Last but not least, random forest can also overfit to noisy data.

(The key is to understand the risk attitude inherent in the two models. )\* not a limitation, just a comment! maybe can put it somewhere, or not :D