Predicting Client's default risk

This analysis/forecast is going to be made using Random Forest algorithim. The motivations behind choosing such solution in this particular case are as follows:

- It produces good results without much hyper-paremeter tuning
- It is simple and diverse, therefore a good choice when the analysis needs to be done quickly (as in this case)
- It's been used in similar cases (Schonlau and Zou, 2020)
- Random forests are more accurate when the dataset is large (as in this case)
  To build the algorithm 14 features were used. Here's a quick overview with explanation as to why this variable was chosen:

Feature Name	Description	Why it was used
OCCUPATION_TYPE	What kind of occupation does the client have	Occupation type plays a very big role in current income but also income perspectives and perspectives of a Client if he/she is let go (if the Client had a good job earlier it is more probable that he/she gets a good job next time)
DAYS_BIRTH	Client's age in days at the time of application	Age of the Client is important, since older Clients tend to be more reliable in terms of their financial situation than younger Clients
FLAG_OWN_CAR	Flag if the client owns a car	Car ownership proves that the Client is able to finance a big buy and still be in a good financial condition, car has also a big share in people's total wealth and as such can add to Client's payment ability
FLAG_OWN_REALTY	Flag if client owns a house or flat	A flat is an anther big aspect of the Client's total wealth and potentially is a asset that can be liquidated (although not with ease). Also, flat or house ownership proves the Client is able to make cyclical payments
CNT_CHILDREN	Number of children the client has	The more children a Client has, the bigger financial responsibility for her/him, this can potentially be a very big risk factor
AMT_INCOME_TOTAL	Income of the client	One of the most important features, it is a primary variable to assess the Client by. The higher the income the bigger the probability the Client will make payments on time
NAME_INCOME_TYPE	Clients income type	Type of income is extremely important because different income sources have different level of dependability and some are less risk-creating than others

Feature Name	Description	Why it was used
NAME_EDUCATION_TYPE	Level of highest education the client achieved	The higher educated the Client the better his material situation will be - now or potentially (statistically speaking
DAYS_EMPLOYED	How many days before the application the person started current employment	If the person just started new job it can idicate he/she often changes occupation and his/hers income is not constant. And thus, if the Client has been working in the same job for the long period of time, it proves he/she is stable and trustworthy
REG_REGION_NOT_LIVE_REGION	Flag if client's permanent address does not match contact address	If Client permanent adress and contact adress are different it can idicate he/she is still somehow dependant (for example a student on the sources from the permanent adress.
ORGANIZATION_TYPE	Type of organization where client works	Some types of organizations lay off their workers often. Workers from some types of organizations tend to be in better material conditions (corporate workers)
DAYS_DELAYED_PAYMENT	Number of days the Client is late with payment	The longer the delay higher the risk of repeating this situations and even defaulting. This is one of the most important variables.
AMOUNT_UNDERPAID	Amount by which the Client underpaid the installment	The higher this amout the bigger risk of repeating this situation and potentially defaulting. Also a very important variable
EXT_SOURCE_2	Normalized score from external data source	Since this variable was chosen for the dataset it must hold a high value. External analysis is also very important.

The other features were not used on the account of maintaining the simplicity of the model and their lesser importance in comparison with the chosen ones.