Walk-through of the Project

i. Cleansing, Preprocessing and EDA

- Look at missing values
- Distribution of interes rate
- Categorical Variables -Explore categorical variables and interest rate
- Numerical Variables -Explore numerical variables and interest rate

ii. Feature engineering

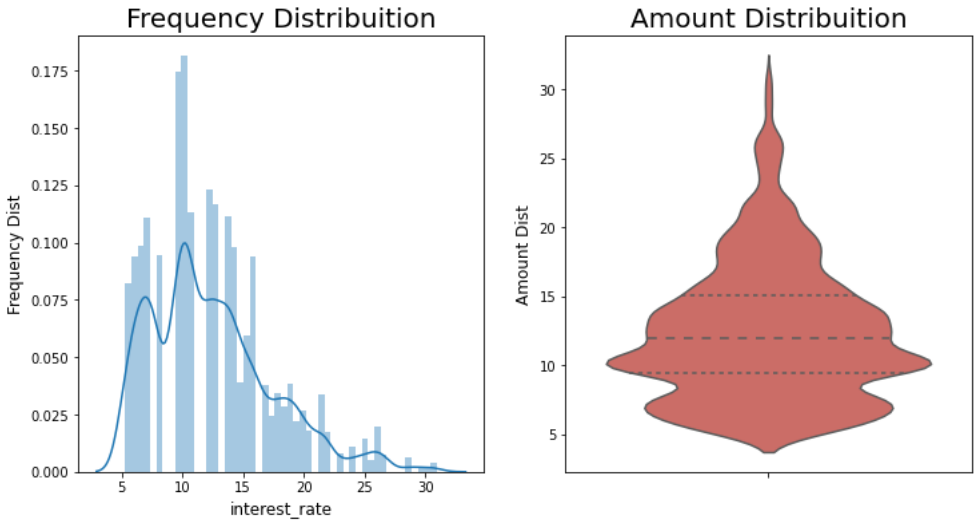
- Adding more variables
- Scaling & Getting dummy
- Feature selection (Lasso CV)

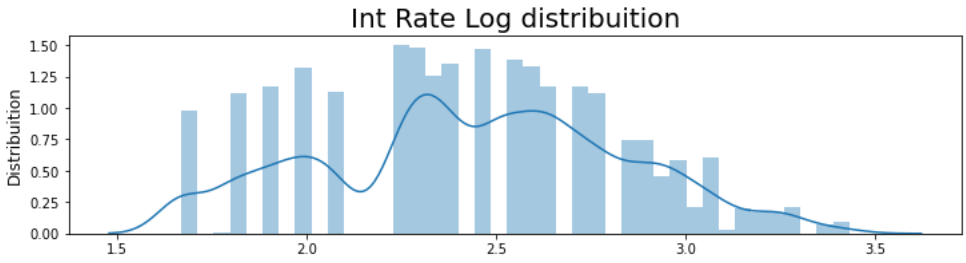
iii. Model

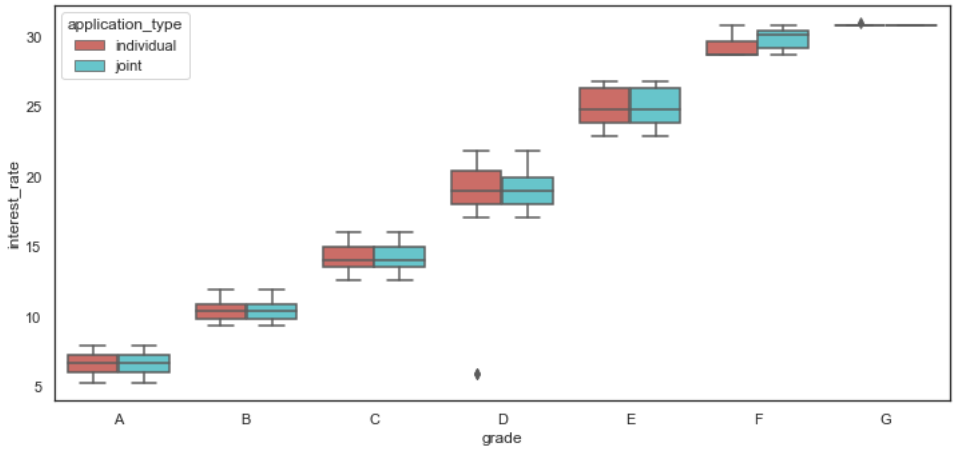
- Random Forest
- XGBoost

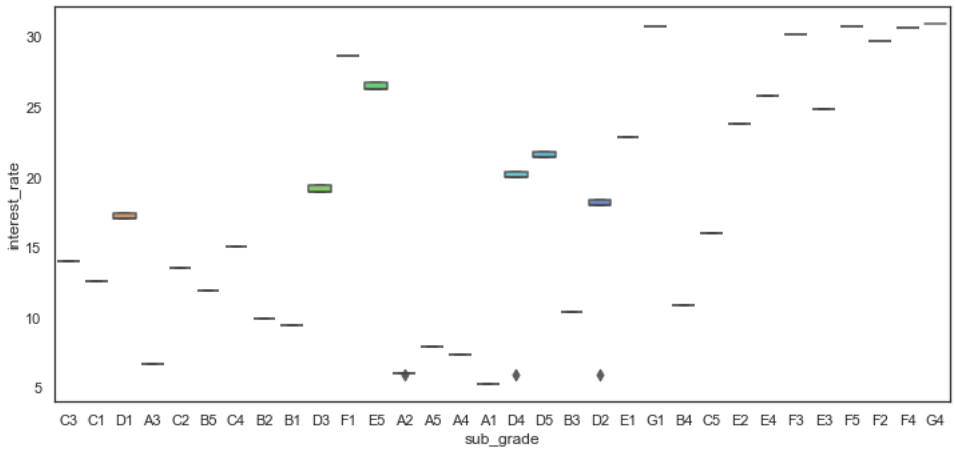
Conclusion

- EDA
  - 10000 sample size with 55columns.
  - Many variables containing outliers and missing values
  - Interest rate distribution are right-skewed. If we use linear regression, we should log-transform the interest rate



- 

    - Grades an subgrades are highly correlated to interest rate
- 

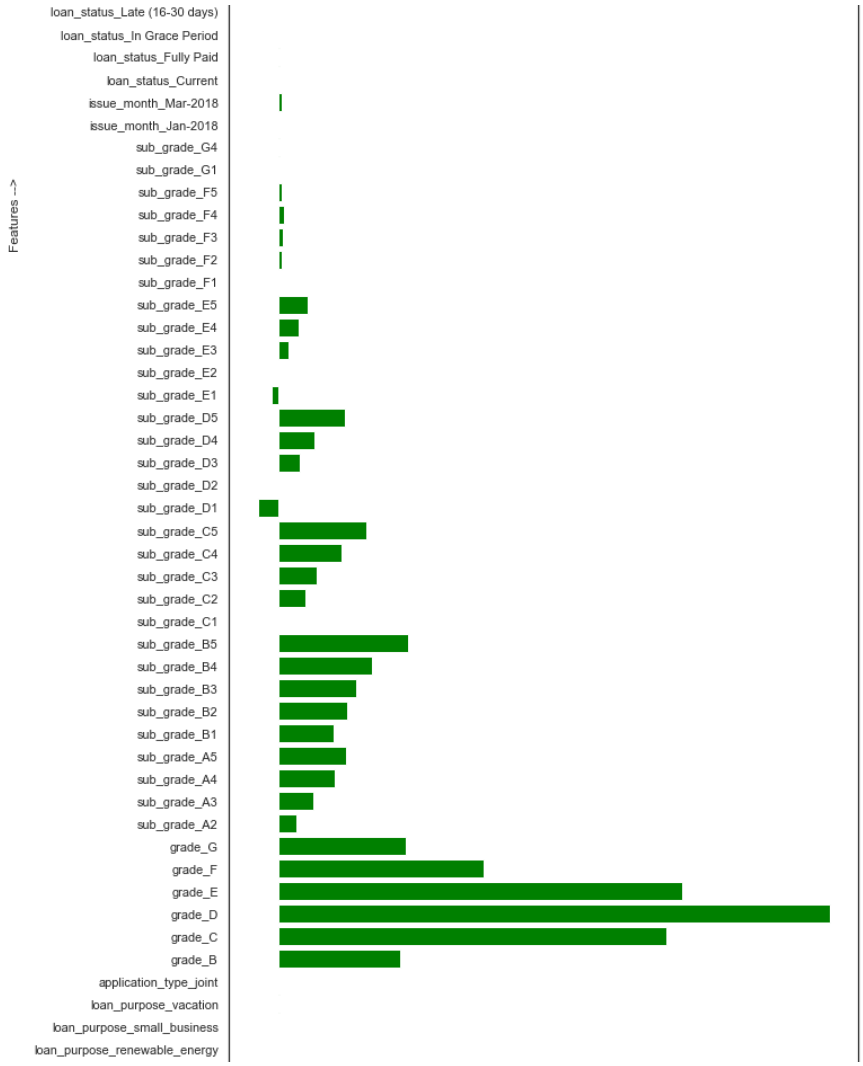


- Model Selection
    -

Metrics	Random Forest	XGboost
Mean squared error	0.36	0.62
Mean Absolute Percentage Error (MAPE)	3.55	4.64
Accuracy	96.45	95.36

- Random forest would be a better choice

- Feature Selection
  - The feature I choose are about Grade and Subgrade -grade: Grade associated with the loan. -subgrade: Detailed grade associated with the loan.



    - However, we don't know what grade does are given. Only when we find out what influence grades, we can deep dive into different variables that affecting interest rate.

- Next step:
    - Add more models (Neural Networks and Linear regression)
    - Explore more about how does grades and sub-grades influences the interest rate. Correlation does not mean causal inferences
    - Explore more on the parameters, optimizing the performance of the model