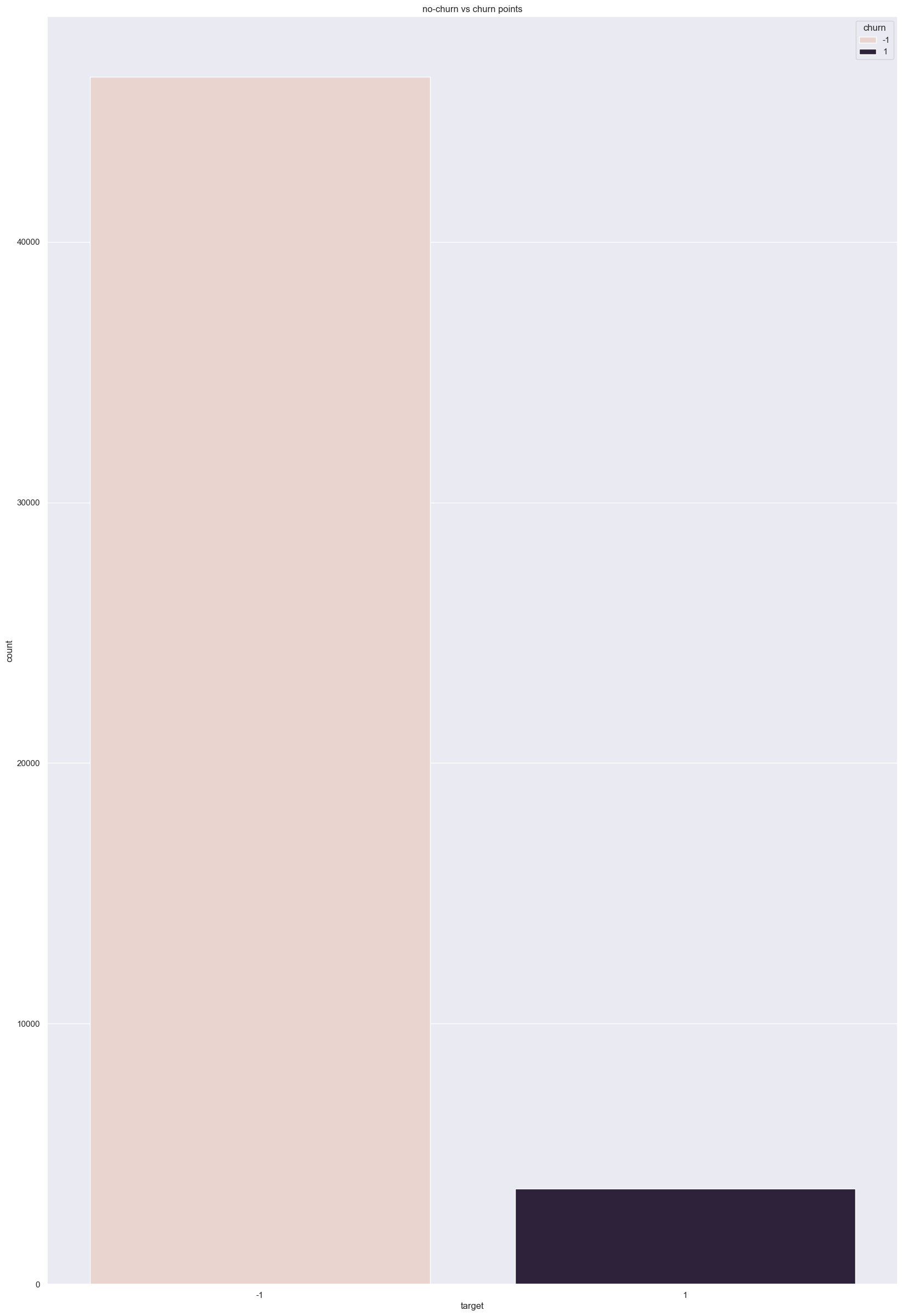
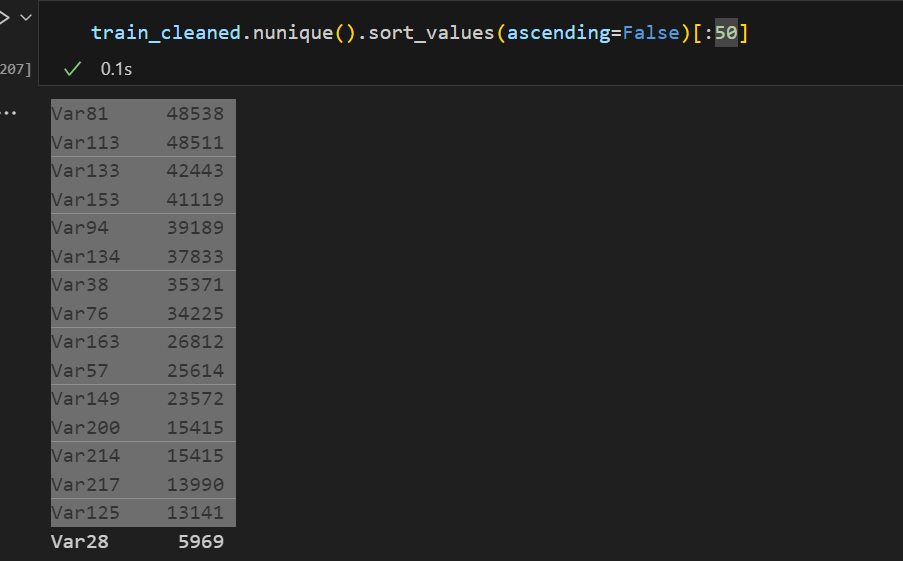
EDA & Modelling

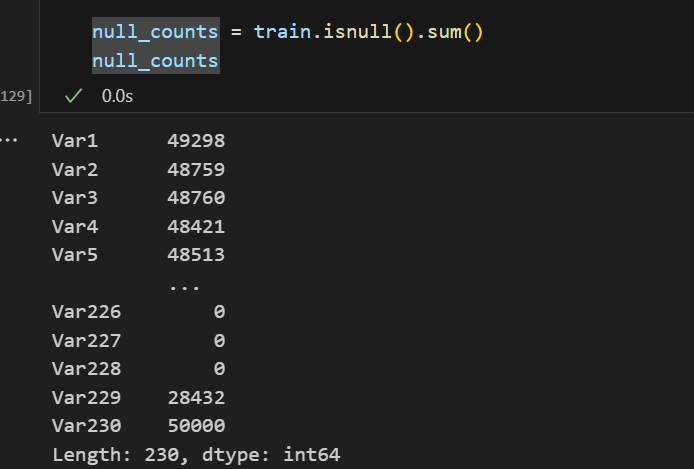
## Summary

* 1. 50,000 rows and 230 Columns and each row represents unique customer and we have 50K labels probably referring to a classification problem such as in case of churn and retained customers.
     1. 38 categorical variables with no metadata or interpretable string without a mapping table
     2. Distribution of churn vs no churn, as can be seen highly imbalanced datasets more towards reality also.



* + 1. Clearly some variable have more information and unique ones could be more important to churn for example var 81 , var 113 and var 133



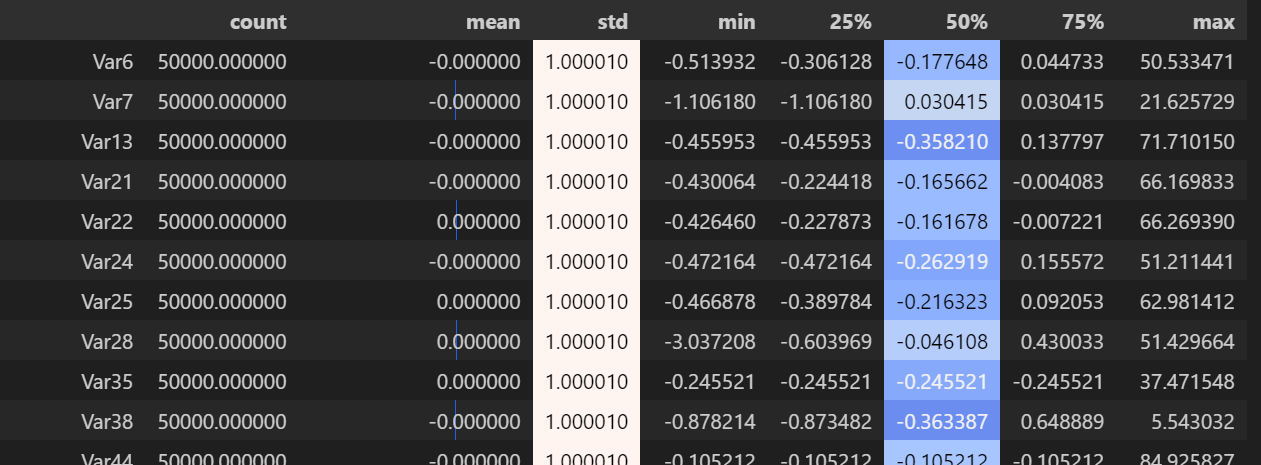
* 1. Missing data – only 76 variables have less than 20% null data in them
     1. 

## Imputation of missing data

* 1. Missing data was filled with linear interpolation method in fill na methods
  2. Other methods like knn, imputing mean /median etc. could also been used

## Distributions & box plots

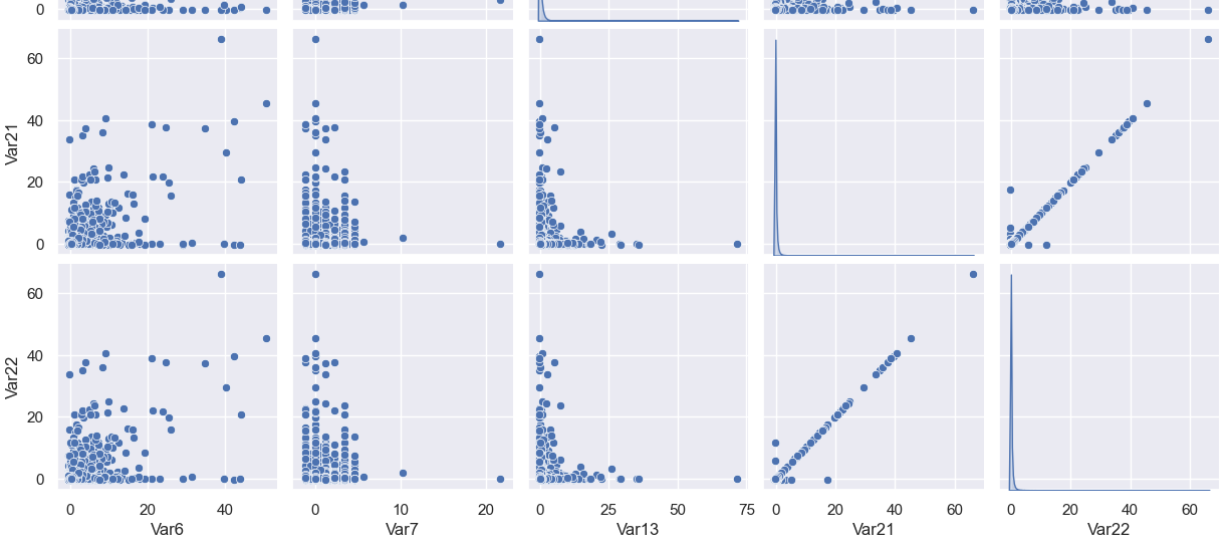
* 1. General summary shows the variable have skewness and median are bigger or smaller than mean after scaling too.



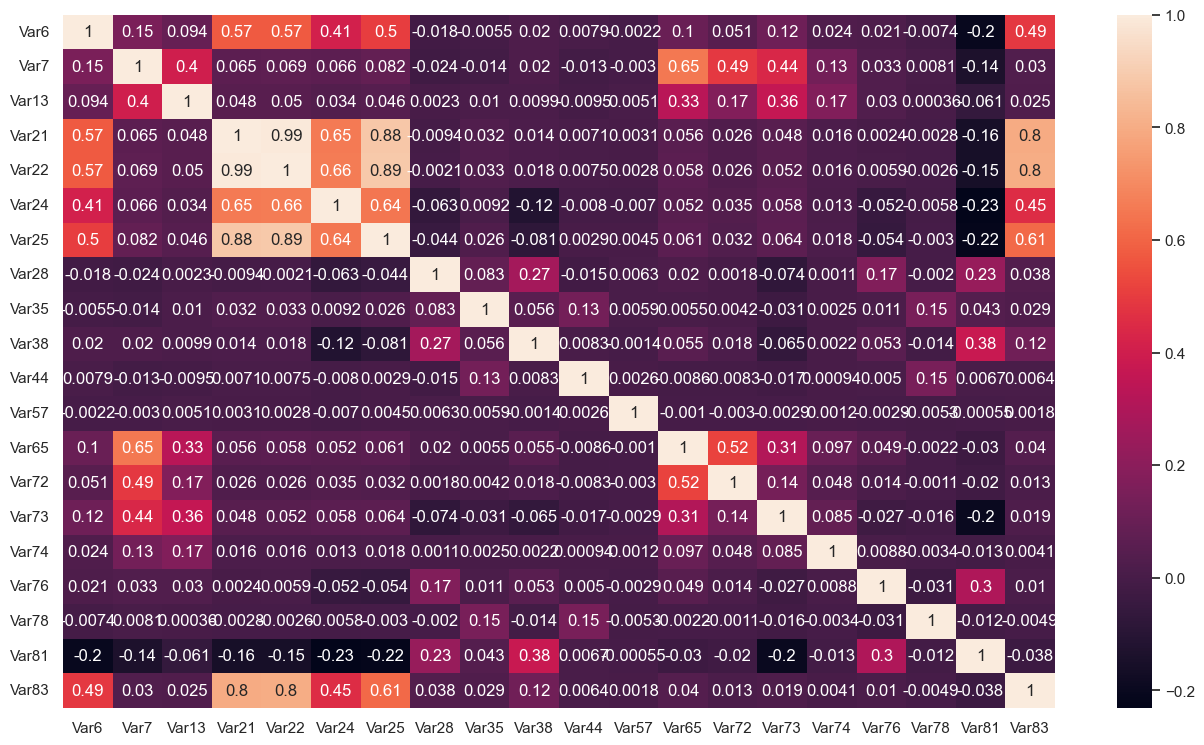
* 1. Var 6 ,var 7 Var13 and var21 id rightly skewed / positively skewed



* 1. Var 21 and Var22 are linearly related as seen in the scatter -plot



* 1. More variables are correlated, Var 25, var 21 are also correlated more than >0.5 for correlation coefficient.



## Encoding categorical variable

* 1. Target encoding was used as it seems to be working well incase of high cardinality of variables
  2. Other methods like ranking and frequency encoding can also be done

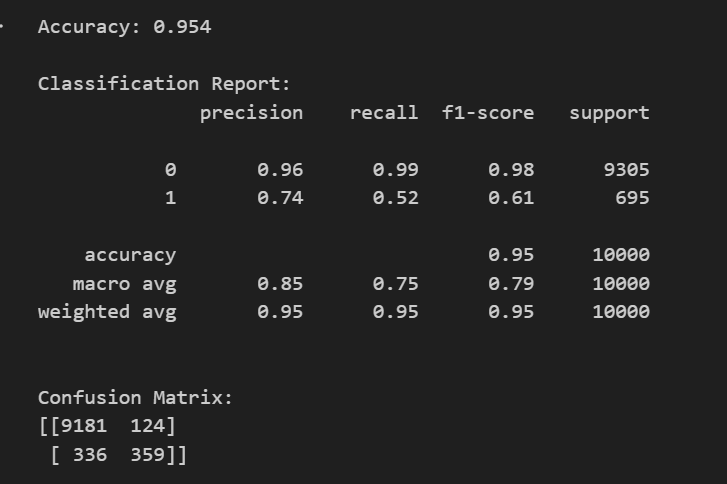
## Feature selection

* 1. Feature importance in logistics regression could be ranking by coefficients of variable. In Random Forests it is how much variables contribute to information gain across all tress where the variables got used. Xgboost has its own method to handle high cardinality.
  2. PCA is useful to get high informative lower dimensional variables from higher dimensional variables.

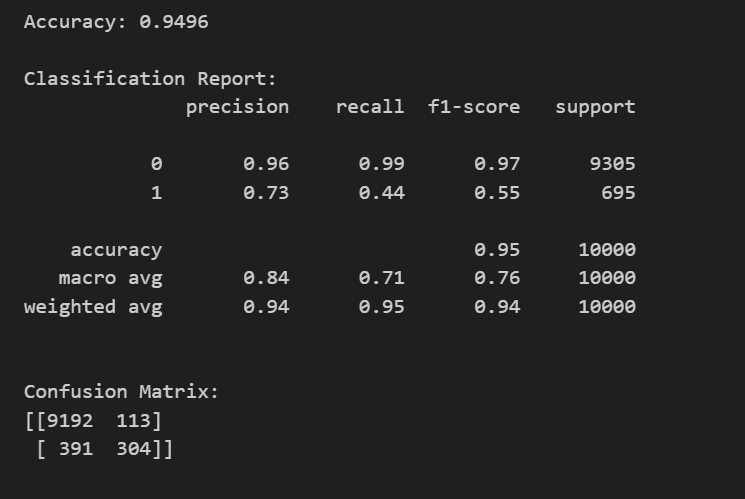
## Model testing

The model metrics should be used would be not the accuracy but the precision or recall based on what is more costly False negative or False positive. A combination of both to look at would be F1 score. Also AUC is helpful too in if we want to know the model is good at distinguishing between classes or not.

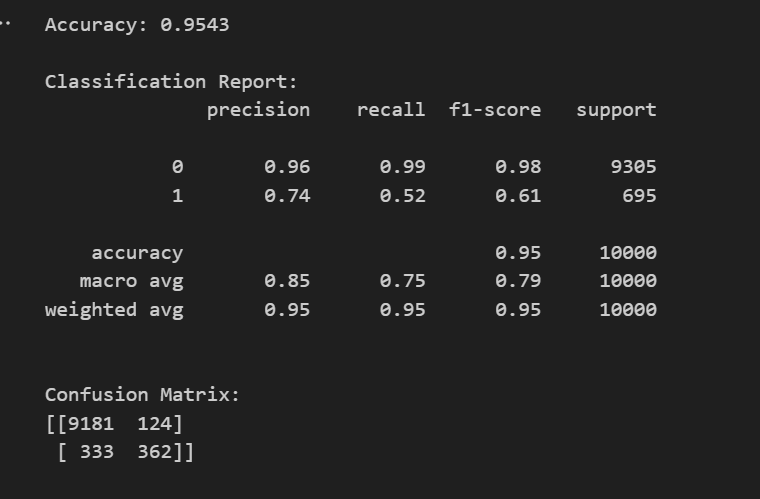
* 1. Logistic Regression with all the variables results show very low recall for churn class. For example, if we are targeting the false negative would mean somebody is classified as churned when he is not. So, we look at recall to take care of False negative in this case

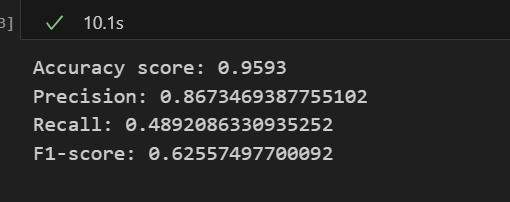


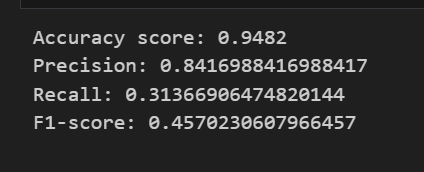
* 1. LR with PCA



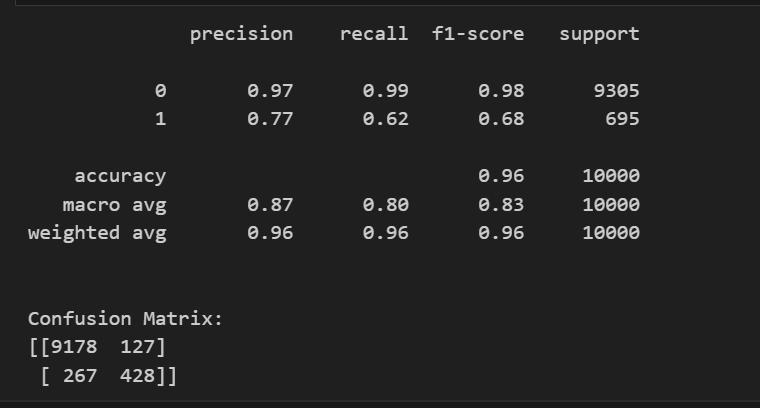
LR with feature selection which uses coefficients to rank the variables and results are similar



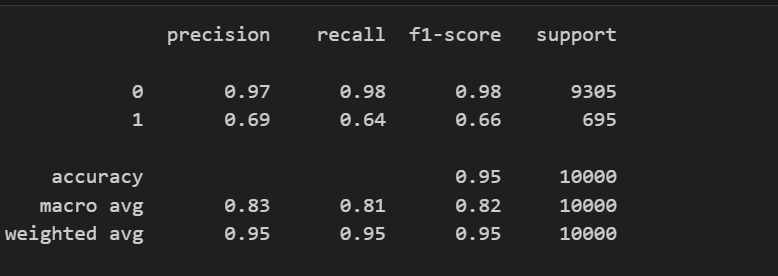
* 1. Random Forests with all variables
  2. 
  3. RF with PCA



* 1. Xgboost results are better too



1. SMOTE with RF cannot beat xgboost too



* Xgboost is better performing if we compare recall and F1 scores too out of all the experiments done
* Performance metric is chosen based on focus on the minority class over here and AUC tell how much the model is able to distinguish between the classes. for logistic regression is 0.5 like a naïve model, for random forest it is 0.96 and xgboost giving auc around 0.97, so being the best among all.
* Next steps could also be treating the skewness in the data ,testing some feature combination, if possible, also selecting some specific features based on recursive elimination methods
* Try more methods and test the hyperparameter tuning using grid search etc. and apply cross validation too to avoid overfitting before finalizing the solutions. Also testing methods of under sampling