Jakka Shashank
IMSc. Applied Mathematics
IIT Roorkee
7895189165
jakkashashank1@gmail.com

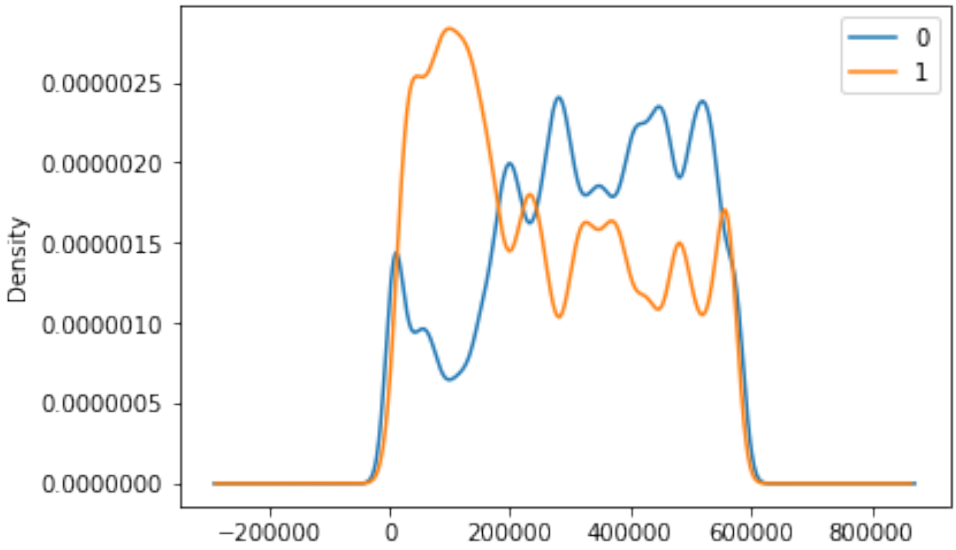
Problem 2 Report

Supervised Modeling with Emphasis on LAUC

Key Findings:

•Average Value of the feature Keys Bias:

1. The classes are balanced.
2. So we should expect almost same average value of keys for both 0 and 1 classes.
3. But a simple groupby shows otherwise.(with 0 having much higher value of keys on average) .
4. This suggest there is some mechanism for assigning keys. (like increment in time implies increment in key values)
5. Thus Giving Key feature a good predictive value.



High possibility of 1 in some range of key values.

Suggesting a time frame pattern.

- **One hotted merged:**

1. Categorical Variables **V11 - V54** actually are one hotted encodings of 2 variables.
2. Label encoding sometimes works better than one hot for trees algorithms.
3. So they have been **merged** to namely Cat_1, Cat_2.
4. This improved the auc score and also **reduced the training time** many times. (given that we already have over 400k rows)

- **Target Encoding of Cat_2:**

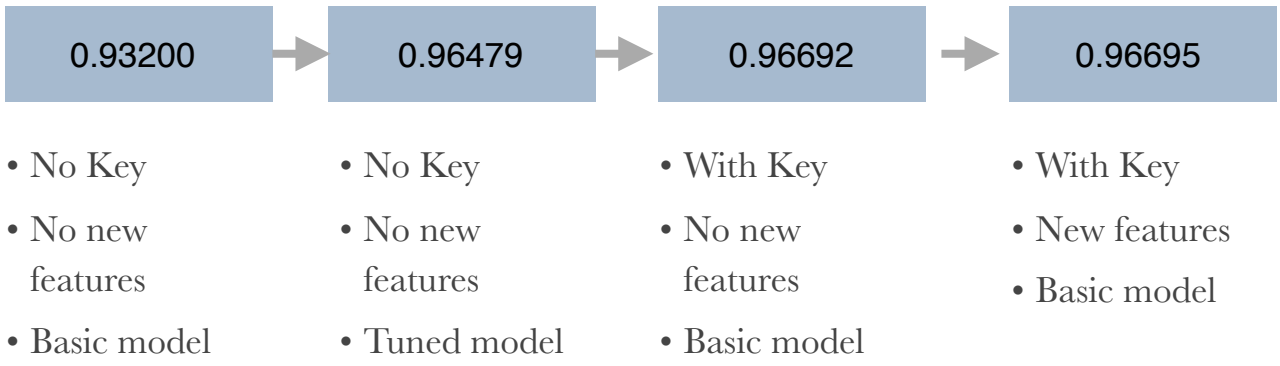
1. Now Cat_2 has 40 unique labels in it.
2. This allows us to encode the variable in turn by **probabilities.**
   - That is encode **label 1** in Cat_2 **with P(class 1 | label 1).**
   - As cardinality of Cat_2 is high, probabilities are more continuos.
3. Now we need to be very careful so as to not leak the information of Y into X.
4. So to do that 10 splits have been made.
5. The probabilities for one split is filled based on another **out of sample** 9 splits. (like CV)

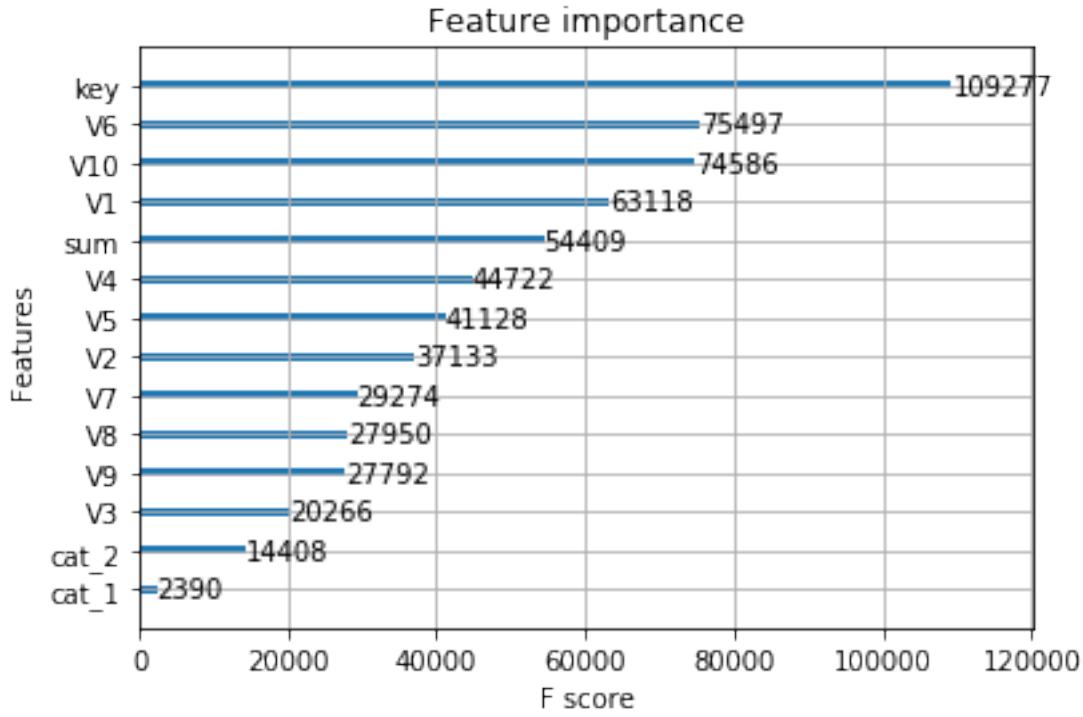
- **Other feature Engineering:**

1. As the features are anonymised, there is not much room for feature engineering.
2. But some aggregating features like sum, mean, variance, standard deviation of continuous variables can have predictive value given if they have enough variance in them.
3. Upon local validation, only sum of continuous variables had some predictive value.

- **Final Model:**

1. A XGBoost model with new feature is trained with params:
   - Depth : 20 (as we have more data and classes are imbalanced)
   - No Boost : 600 (found by early_stopping 100)
   - Learning rate : 0.1
   - Subsample : 0.8 (as we have so many rows already)




Feature importance
key
V6
V10
V1
sum
V4
V5
V2
V7
V8
V9
V3
cat_2
cat_1
109277
75497
74586
63118
54409
44722
41128
37133
29274
27950
27792
20266
14408
2390
Features
0
20000
40000
60000
80000
100000
120000
F score