Approach

The very step was to explore and visualize the data and perform basis check on it to get familiar with it. Target variable was highly imbalanced so I got an idea that for this model I have to try different sampling techniques and check which one is best for this model. Checked for missing value, luckily there were none in this dataset. Checked for correlation if any among columns. Created dummies for categorical variables. Created a new feature **totaldays** by subtracting origination date and first payment date. I have tried to use **PCA** and **SMOTE**(in combination) for this particular dataset with different ML algos but result was not that good. Finally I have decided to go without PCA as dimensionality was not a concern for this data set. I have used different models but **Adaboost classifier** has given me the best result. For parameter tuning ,I have used grid search.

Quality checks performed / Errors found:

- Target variable was highly imbalance.
- No missing values were present in data.

Data Visualization:

- No. of unique elements for each month

Column	No. of unique elements
m1	4
m2	5
m3	6
m4	7
m5	8
m6	9
m7	10
m8	10
m9	11
m10	12
m11	13
m12	13

Number of unique elements increases as month increases.

- number_of_borrowers and co-borrowers_credit_score are highly correlated to each other.

- Majority of source is “X” with a count of 63858 (55.02%).
- insurance_type has value “0” with a percentage of 99.67%.

Data Preprocessing and Feature Eengineering:

- Created a new feature **totaldays** by subtracting origination date and first payment date.
- Dropped origination date and first payment date from the dataset.
- Created dummies for categorical variables.
- After trying difference sampling techniques to deal with class imbalance went with SMOTE.

Model Choice Explanation:

- I tried different ML models like Random Forest, Logistic Regression, ,Decision Tree. All the models were not giving the desired results. Some models were underfitting like Random Forest and Logistic Regression whereas some models were overfitting like Decision Tree.
- I have also tried different ensembling classifiers such as XGBOOST,ADABOOST,Catboost.
- XGBOOST has given best F1 score but Interestingly ADABOOST has given me best result in public leader board of Analytics vidhya so I went that as my final solution
- I have used GridSearch to find out the best parameters.

Final AdaBosst Classifier model:-

```
AdaBoostClassifier(n_estimators=50,
                   learning_rate=0.45)
```

The most important features are loan_term and insurance_type.

Key Takeaways from the challenge:

- Always visualize the dataset to get the insights before applying any ML model to it.
- Try different models. Never stick to one model.
- Try to fine tune models using grid search
- PCA is not the solutions of every problem.
- For classification problems, Always check for class imbalance and try to treat them using different sampling techniques

Things a participant must focus on while solving such problems:

- Start with slowly and thoroughly reading the problem statement. Do a quick inspection of the data by using pandas builtin functions and note down my observations. By quick inspection I mean loading into a variable df, doing df.info(), df.nunique(), df.describe() and so on.
- Understand each of the predictors (attributes) and look for biases in the data. This is the key and it is not related to algorithms. It is mainly commonsense and business knowledge. For classification problems, examine and understand the percentage of positive labels in the target variable. For regression problems, examine the distribution of the target variable.
- Do standard scaling on continuous variables to bring them all to mean 0 and standard deviation 1. Next do a correlation plot. Drop the correlated columns (typically correlations > 0.7) to prevent model being biased.
- Try training with Logistic Regression model, if it is a classification problem, and Linear Regression for regression problem. Now start with hyperparameter tuning, if you know a bit of the theory. If not, make an attempt to read up and try. There is nothing to lose.

Shashank Paliwal
+91 8005097763
https://github.com/spaliwa1
spshashankpaliwal@gmail.com