Find Default (Prediction of Credit Card fraud)

This is a fraud detection case where we predict whether the transaction is fraud or not fraud.

1. **We first saw the data checked the null values.**
2. **Then we checked the correlation of each dependent and independent variable. We checked but we found most of variables have low correlation with the dependent variable. But we cannot drop all the variables. So, we tried another approach.**
3. **We found out the feature importance for Logistic Regression and found the 10 most important features.**
4. **We did Exploratory Data analysis and found out that :**
   We found that there is no relationship between the amount of transaction and whether transaction is fraud or not.
   We also found that the time interval is less the chances the transaction fraud is less.
   Also, we found that as V12 increases the chances of fraud also increase.
5. **Then we found out if there are any missing values in the dataset.There are no missing values in the dataset.**
6. **We removed the outliers of the data excluding class and time.**
7. **We Standardized the data.**
8. **Then we need to remove the duplicates.**
9. **Since we have imbalanced dataset present here in column Class fraud and non fraud**
10. **SMOTE is a method used to address class imbalance by generating synthetic samples of the minority class. In imbalanced datasets, where one class (the minority class) is significantly underrepresented compared to another (the majority class), traditional machine learning algorithms may perform poorly because they tend to bias towards the majority class.**
11. **FEATURE ENGINEERING: Here we found out the features with high skewness and then transformed those features.**
12. **APPLYING VARIOUS MODELS:**
    Here we found out that is this a fraud detection problem here recall should less as the transactions which are not predicted as fraud are fraud should be less.
    Also, accuracy is high and the model is not overfitting.
    Comparing with other models like Decision Tree or Random Forest Logistic Regression is better. So, I used Logistic Regression.
13. **Applied K fold cross validation for model validation and applying cross validation with 5 folds. As the mean accuracy for k fold cross validation of Logistic regression is good we can surely use it in making the predictions.**
14. **Then we used Hyperparameter Tuning found the best parameters and re-trained the model and re-train the model.**
    Use the best model for training the model
15. **MODEL DEPLOYMENT: We deployed the model using Flask and loaded the file.**