VIDEO TRANSCRIPT

INTRODUCTION:

1)Hello this is group 9 with the HCDR project , with myself, Hanish, Nidhi and Srinivas
2)The contents are as follows.
3)The four p's in this project are as follows:
a) for our completed phase zero this was all about setting the scope of the project, understanding the project requirements and deadlines. The biggest items were identifying the machine learning algorithms and identifying the loss functions and metrics that we wanted to apply for the project.
b)In the present we did EDA and built baseline logistic regression pipeline and rebalanced the data and we calculated accuracy and AUC
c)In the planned phase 2 we plan to use decision making trees , random forests and SVM
d)The problem we are facing is we need more prior knowledge about the data which will help us in feature engineering and hyperparameter tuning.

EDA:

We performed exploratory data analysis on the target's dataset features.

- We did descriptive analysis on the dataset such as data type of each feature, dataset size summary statistics such as the number of observations, mean, standard deviation, maximum, minimum, and quartiles for all features.

- We generated charts on descriptive statistics of the target dataset.

We compared few attributes like gender differences, income difference, whether an applicant owns a car or a home and after detailed analysis using bar plots, these are the conclusions we have drawn-
There are more women than men in both defaulters and non-defaulters. The difference between women and men was much larger in the non-defaulter group. Men are most likely to default than Female based on percentage of defaulter_count. Men had more income than women in both target and non-target. Men in non-target have higher income than men in target. There are more people who own houses than people who don't own houses in non-target. There are more people who own houses than people who don't own a house in the target.

We also included correlation plots, heatmaps to measure the strength of the relationship between two variables, and as well as visual representations of a number of key features that would help us better understand the characteristics of the average defaulter.

MACHINE LEARNING PIPELINES:

Phase 1 Machine learning pipelines:

Here we have implemented the logistic regression model upon the home credit dataset

First, we partition the data into train and test data. For correct findings, we separated 20% of the test data with a random seed set to 42.
Next, we developed a logistic regression baseline pipeline. Based on numerical properties and a standard scaler, we create a numerical pipeline. We use the median to fill in the blanks. With this numeric pipeline, we perform a logistic regression.

Here we are using the baseline model on the selected non defaulters data and the non defaulters data is about 10 times greater than defaulters data. We are trying to balance that

And the further details are provided in Jupyter notebook.

CONCLUSION:

The results of our model are as below -
- The baseline model ran with a training and testing accuracy of 91.5 and with an AUC of 0.499
- However, when we tried to check the accuracy of the baseline model with 50k and 75k data tge accuracy dropped to a great extent, but the AUC increased a little.

The end results are as below ->
- We know that AUC is a great indicator of high quality prediction
- And the higher the AUC, the better the model.
- One observation was that altough the baseline model had great accuracy, it had very less AUC.
- Which doesnt gaurantee a great prediction.
- However, with rebalancing, it was evident that the AUC increased even with lesser test accuracy.
- Hence, we can say rebalancing vaguely did improve the accuracy.
- Another important observation is that, the baseline model does not have any feature engineering, and hyper parameter tuning, which may further improve the reliability of the model.