**For sure. I did quite a few,but I think will pick my favorite, simply because it cleared a lot of data science basics for me.**

**about project a financial entity that want to issuie credit card and also want to increase credit card limit for existing customers.**

**the process they were doing before giving that project to our team , there were doing very hectic and challenges manual task to identifiy customer to give credit cards and because of that they were loose their business they were unable to identifiy which customer are paying credit card bills or which not and because of that they get loose their business.**

**therefore they decided to go with to automate the process of identifying potentional customer those not getting default the credit card bills.**

**So in that we get the project , yes after that ,client has stated his problem to our team and i was the part of team and after that we got in discussion and you know , i personally whenever i faced the new project i used to refer the research papers & different blogs to how we can solved the business problem and i given some inputs to the my team as well as to the clients .**

**so these are some essential features required to classifiy the problems that was stated by the client ,**

**ID of each client, Amount of given credit loan , Gender, EDUCATION, Marital status,Repayment status of five months(Sept-April),Amount of bill statement of five months(Sept-April) ,Amount of previous payment of five months(Sept-April) , etc.**

**so we have given some inputs to the client and we clear with client that how we solved your problem by which method (Machine Learning).**

**so after that client has given the data in csv format , so we have received the data file along with the schema file by our data engineer team ,in a schema file they have mentioned about a the number of columns, name of the features and its data types and we supposed to be cross checked the data with schema file.**

**for that we have created two folder one genuine folder and other is non-genuine folder if data is matching to the schema file then we transfer these data into genuine folder and if not matching then we transferred in non-genuine folder after that we communicate with client that some of your data is not matching with the schema file so i personally contacted with client about that...so i personally cross checked the data with schema file.**

**after that i checked the data and its data types ,so how many number of columns and the name of the features present in data so i have received around 1.2 millons of the data and around a 33 features.**

**numberical features -> age , credit limit ,bill statement and so on**

**categorical features -> Gender, EDUCATION, Marital status,Repayment status of five months**

**features with missing values-> credit limit, age**

**features withe outliers present -> bill transition of the months**

**features with one hot encoding-> Marital status, Repayment status of five months, location**

**features with labels encoding -> education, gender,**

**so after that we done the EDA , so in EDA we have checked the univariate analysis and multivariate analysis so in univariate analysis i have checked which are features are categorical which are numerical features like gender in which we have replace label encoding, education in multivariate analysis there are multiple nominal data so for that part i have done one hot encoding because of nominal data we can't compare there two categorical which is more weightage so therefore we decided to go for one hot encoding.**

**after that i have checked for the outliers for the numericsl fearturs , so i have confirmed the outliers by using boxplots , then again i confired with iqr method and with discussion with the domain expertise , we have decided to imputs those outliers with the mean, so as data analyst we have to analyse with the number of solution , until and unless we get good accuracy and good result, so we decided to go with mean mode median and we get the good accuracy with mean so we decided to go with mean.**

**So After done with the feature engineering we go for the feature selection part in that we used some features selection method ->**

**for categorical ->chi square method,mutual information gain**

**for numerical -> Pearson's correlation , spearman rank coefficient**

**so before going with the model building we supposed to cross check the assumptions and we done the same thing in that project. Assumptions like- linearity , no-multicolinearity with independent variables, after checking correlation i found that there are four features they are having higher correlation between the independent variable ,and we know that there should not be a multicolinearity bet independent variables.so we decided to drop that features with the concern with the domain experty .**

**as part of model building we have selected some of ml algorithms - Logistic, KNN, decison tree , random forest , adaboost , so i have selected this model with the concern with my teammates.**

**In model building**

**Logistics regression -> accuracy 78% , Rejected because not get enough accuracy**

**KNN -> get good accuracy 85% but rejected because of time complexity & also precision**

**Decision Tree-> overfitted model**

**Random forest -> get good accuracy 91% , not good precision value**

**Adaboost -> get accuracy about 94% ,also precision , f1 score , recall .**

**so i have selected percision , recall , f-1 score , aoc-roc curve evaluation matrix ,so i got the result but i decided to go with the precision only , i have conveyed the same thing to my teammates and clinet as well , the precision is important and why not the recall?**

**because of financial domain we have to minimise the false positive points that are like( a person is genuine and model is predicting is not genuine so the ultimate result can be loss to the business) those are falsely predicting that the are paying the credit card bills .so we have to minimise that points**

**take break....**

**so they ask question? your data is imbalanced or not ?**

**-Yes,,i have handled the imbalance data as well...techinique - so i have used oversampling method -smote , Adasen ,so as data analyst ..we know that there no method to get good accuracy with imbalanced dataset , so we need to try n number of solution to get good result .**

**how oversampling method works?**

**what are the challenges that you are face in the project?**

**-> yes i have face the chanllenges over there and i have solved the same thing**

**i have received data in 2nd from that come from different locations so that time i got less for 2nd batch , that get good result for first batch , so the accuracy ,precision get less so i stucked over there , we did R&D on that result and i read somewhere whenever data coming from different locations,it is having diff patterns and might that pattern is not caching by my model and that was the reason we got less accuracy ,so after analysing model we identified the patterns and we add location feature in our original dataset and again train the model . so finally we got good result that accepted by client , so this were the challenges that i overcome the challenges.**

**okay thought i am part of the team , but i myself solved the problem and that was the good learning for me and i appreciated by client as well.**

**on the basis that result client has refer my name for his 2nd project.**