Credit Card Prediction SC1015 – Presentation Transcript

Slides	Transcript piece
1	Hello! Our team consist of Davis, En lih and Tomoki and we would be presenting on our project on credit card approval prediction
2	Here is our table of contents
3	We would first be moving on to the introduction of our dataset
4	**(Intro to Dataset cover)**
5	On the background of our project, Firstly, the usage of credit card is extremely prevalent in today's day and age. However, given that credit card is built on the system of trust whereby the client can charge purchases to the card and make the necessary payments at the end of the month, it is important for banks to know who are the clients that impose a higher risk. Secondly, the dataset for the consideration of the client's credit card approval is exhaustive. This is especially so as financial firms will collect as many information as possible to determine if the client impose a high risk. After all, credit system is akin to a short-term loan
6	Now we would be elaborating on the prevalence of the issue. Notably it has been shown that the average consumer credit card debt was higher than it ever had been. Additionally, there are many people which would be affected worldwide, with 2.8 billion credit card users and 70% of people having atleast one credit card
7	Problem Definition: How does the various data submitted by Credit Card Applicants determine their credit scores (i.e. Clients' Risk) based on prediction of future defaults and credit card loans?
8	Our team's objective is to distinguish the good clients from the bad clients based on the time taken for them to repay their loans. A Good client either has no loans or can repay their loans within 29 days and hence impose a lower risk to the bank, whereas a Bad client is one that takes longer than 30 days to repay their loan, or loan default and hence impose a higher risk to the bank
9	Next we would be covering on the Data engineering
10	**(Data Engineering)**
11	From our datasets, it is noted that there are two excel files included;
12	The first file (Application_record.csv) contains the client's information such as their gender, if they own a car, if they own a realty, the number of children they have, etcetera

13	The second file (Credit_record.csv) contains the respective client's ID with their 'status' over the months. The 'Status' of the clients is tied to their loan repayment period, as seen in the list above. As mentioned earlier, since we define good clients as those that has no loans or have repaid their loans within 29 days, 'X', 'C' and '0' status are considered the good clients (which is the majority), whereas bad clients are those with loans repayment more than 30 days, id est '1', '2', '3', '4', '5'
14	Exploring the dataset under the client's details, we also did a data visualization of our variables, and found our dataset to be of a Mixed Type with a notably large portion of our datasets are Categorical variables whereas only a few are Numeric variables. Here is a bigger picture of our numeric data and categorical data
15	(zoomed in diagram of data visualization)
16	Next, our team would be moving on to the Data cleaning portion. Here is an abstracted version of the changes we have made.
17	We removed the duplicate IDs since they were inconsistent to ensure the integrity of the data. The customer's details also included details like 'days_birth' and 'days_employed'. This may not be very useful and hence for a better analysis, we have converted them into years, allowing us to get a more in depth and qualitative analysis in the later parts. Additionally, some data types in the data were indicated wrongly such as count of the family members being indicated as 'float values'. Hence we have converted them back to the correct data types. Finally, we also noted that there were *null* values in the Occupation type. We ensured that the individual was really unemployed through comparison of values with the 'days_employed' as there were some cases where 'occupation_type' was noted to be *null* while there were non-zero records of the 'days employed' column. Hence, we compared and noted that if the 'occupation_type' was nill and 'days_employed' was 0, the client was indeed unemployed hence the *null* values would be converted to 'unemployed'. The remaining *null* values will contain field where our client had employed days and hence were employed with occupation not known. In these cases, we have converted the null values into 'unknown' hence ameliorating the issue.
18	Going back to the earlier slide, as part of data cleaning, we have also created a new column 'GOOD_OR_BAD_CLIENT' to determine from the client's 'status'. This would be our *response variable* We have found that we have a total of 98.65% records of good clients and 1.35% of bad clients. However, we have noted that this is not an accurate description of the good and bad clients as they contain a record of the client's status over a month. Id est, a client can be considered bad for failing to pay their loans the past 4 months but managed to pay all their loans on time this month, marking it as 4 bad records and 1 good records for

	that particular client (tagged to his id). This would lead us to our Exploratory Data Analysis
19	To solve this issue, we have used the inbuilt function in pandas which allows us to aggregate the risk of the clients given the same ID. For the same example as mentioned previously, the client with 4 months of bad records and 1 months of good record would still be marked as a bad client since for most of the months with the bank, the client had either accumulated or defaulted on their loan repayments. Conversely, a client with 4 months of good record and 1 months of bad record would still be marked as good since the client despite late payments of his loan on one of the months, have been making their loan repayments on time
20	After using our aggregation function, we got a more accurate gauge of the percentage of good and bad clients as seen above
21	Finally, we concatenated the two excel files after rectifying the data and doing the Exploratory Data Analysis as well as data wrangling. The two excel files can be concatenated as we are able to use panda's .merge function by indicating the client's individual unique IDs post data engineering.
22	Next, we would be moving on to our core analysis consisting of our machine learning models
23	**(Core Analysis)**
24	Since our dataset is extremely imbalanced, prediction of our machine learning model will become biased towards the majority. In this case, it may always predict our client to be a 'good client'. SMOTE; or Synthetic Minority Oversampling Technique, will allow us to overcome this issue by generating synthetic samples for the minority classes, in this case our 'bad clients'. This forces the machine learning models to also consider 'bad clients' upon using SMOTE and hence predictions would be more accurate and correct
25	These are the machine models that we will cover.
26	Now we would move on to our first machine learning model; Decision Tree Classification. While Decision Tree Classification model is one of the most commonly used machine learning model due to its ease of use, in our case we had only obtained a model accuracy of 57.6% for our train dataset and 58.1% for our test dataset. Hence we decided to use Random Forest instead to improve on our model accuracy
27	For our Random Forest model, it is an ensemble learning method which operated by constructing a multitude of decision trees, and will return the mean prediction of the individual trees. This allows the model to be more accurate and factual. In our case, we have obtained a model accuracy of 79.1% for out train dataset and 71.8% for out test dataset.

28	XGBoost is something new that we learnt. It is an ensemble learning method which use numerous learning algorithms to obtain a better predictive performance as compared to the individual algorithms alone. While decision trees is one of the most used models, they often exhibit highly variable behavior which may result in errors. Hence, we use boosting to ensure that the trees are built sequentially such that each subsequent tree will reduce the errors of the previous tree. - In contrast to the Random Forest technique where trees are grown to their maximum extent, XGBoost make uses of trees with fewer splits. Such small trees are easily comprehended and readable.
29	As seen here, XGBoost is our best model for prediction if our client is a 'good client' or a 'bad client'; with a model accuracy of 87.7% for our train dataset and 83.2% for our test dataset.
30	We can improve our model further through Grid Search Cross Validation to find out the best possible 'n_estimators' and 'max_depth' which is 700 and 2 respectively as can be seen from the result above.
31	After grid search, the model accuracy for train dataset decreases slightly to 86.4% but the accuracy of test dataset increases to 85.0% which is slightly better
32	Finally, we would conclude our project and highlights the outcomes as such
33	**(Conclusions & Outcomes)**
34	Firstly, we learnt about Synthetic Minority Oversampling Technique; or SMOTE Given that our data was extremely imbalanced with most of our clients being classified as a 'good client', most models did not fit well initially. To solve this, we used the concept of SMOTE as covered earlier While Decision Tree was used, the accuracy was not ideal and hence we had used Random Forest instead, which we had also covered on earlier Decision Tree also exhibited highly variable behaviors, and hence we had used XGBoost which was another new machine learning model that we assimilated into our project and covered on earlier.
35	Out of all the machine learning models, we have found that the machine learning model that best suit our data set was XGBoost, which gave us the highest model accuracy of 85.0%. This meant that XGBoost would be a very useful model for the banks to predict if a client were a good client or a bad client; given the risks of client defaulting on their payments upon credit card which banks should consider prior to the approval of their credit cards. However, it is noted that while this model worked the best for our dataset, it may not always be the best model in all cases.
36	For our extra improvements, we want to find out explicitly what are the best predictors for our response. We used Recursive Feature Elimination to achieve this.

37	From the above, we can see that count of children, client's total income, age, years employed and family status played the biggest key importance. The rest of the variables are not a good estimate even though they are in the list as not all types were included in the list
38	However, something we can further improve on our Predictor variables is the use of optimal feature selection, which would allow us to better get an answer for the best set of variables for models, hence giving a better solution. With that, we would like to conclude our presentation. Thank you!
39	(Thank you!)
40	(Reference list)