

DPDzeroData analyst assignment - 2024

This assignment is designed to give you a flavour for the kind of data analysis a typical data analyst does at DPDzero. Data analysts here leverage **Jupyter notebooks** heavily to do their tasks and you are expected to use the **same tool** to do the given task.

Please keep in mind that this is an open ended assignment

Dataset

Derive values from the raw data

Calculate the risk labels for all the borrowers.

label all customers based on where they are in their tenure


Segment borrowers based on ticket size

Give channel spend recommendations

Submission guidelines

Dataset

DPDzero is a data driven organisation where most decisions are taken based on data analyst recommendations. Here is a typical data set (it is anonymised) that a data analyst may look at.

 [Data_Analyst_Assignment_Dataset.csv](#) 1131.9KB

The dataset contains Loan data for various borrowers in a loan portfolio

The columns are as follows

Column Name	Description
Amount Pending	This is the EMI amount pending.
State	The borrower's state.
Tenure	Total tenure of the borrower. This is the total tenure of the loan.
Interest Rate	Interest rate of the loan.
City	The city of the borrower.
Bounce String	This is a string that explain's customer's bounce behaviour since the disbursal of the loan - customer did not end up making the payment <ul style="list-style-type: none"> • S or H- No bounce in that month • B or L - Bounce in that month • FEMI - first EMI - no known behaviour • Last character denotes the last month - first character denotes the first month on book - customer was on book for 4 months and he has bounced the in the last month
Disbursed Amount	The total disbursed amount of the loan.
Loan Number	The unique identifier for the loan.

We recommend that you explore the data before you go to the next section of assignment.

Derive values from the raw data

When a data analyst gets data from the lender at DPDzero, a lot of information should be derived and data set needs to be enhanced. As part of this assignment, derive the following values

Calculate the risk labels for all the borrowers.

Unknown risk	New customers
Low risk	Customers who have not bounced in the last 6 months
Medium Risk	These are customers who have bounced max twice in the last 6 months - The bounce should not have occurred in the last month
High risk	every other customer

label all customers based on where they are in their tenure

Early tenure	Customers who are in the book for 3 months
Late tenure	Customers who are 3 months away from closing the loan
Mid tenure	Everyone else



Hint: Read the delinquency string description above

Segment borrowers based on ticket size

Distribute the data into 3 cohorts based on ticket size. This is to be done such that sum of amount pending in each cohort should be approximately equal. Apply the following labels on each borrower based on this logic:

1. Low ticket size 2. Medium ticket size 3. High ticket size

Note that at the end of this exercise you would have a lot of folks with low ticket size and a few people in high ticket sizes - sum of amount pending for all these cohorts should be approximately equal

Give channel spend recommendations

At DPDzero, we employ various channels to communicate with the borrowers so that we can get the repayment done - Different channels have different costs & various degrees of effectiveness.

You are allowed to spend 3 kinds of resources to reduce the overall bounce

1. Whatsapp bot: This is the cheapest medium - it will cost 5 rupees per borrower
2. Voice bot: This is the mid-cost - it will cost 10 rupees per borrower
3. Human calling: This is the costliest option - it will cost 50 rupees per borrower

Whatsapp bot will work well in any of the following scenarios

1. Customers with great repayment behavior
2. Customers with first EMIs
3. Customers who have low EMIs

Voice bot will work well if all the following conditions are met

1. Customer who know Hindi or English
 - a. Metropolitan areas have high probability of english speakers
 - b. People with low interest rates are also typically english speakers
 - c. There are many states in India where the borrowers typically know Hindi
2. Customers who have had low bounce behaviour
3. Customers with low or medium sized EMIs

Human calling will work on all scenarios but is the costliest option and you need to use this channel only where absolutely necessary

Your job is to segment the borrowers into these 3 channels of spend category and minimise the overall spend while maximise on time repayment.

Submission guidelines

Create a PDF report of your Jupyter notebook and submit it here

<https://forms.gle/eKxkzAVUcsZygtrr9> (Alternate link)

1. summary of borrowers (with graphs) based on risk
2. summary of borrowers (with graphs) based on ticket sizes
3. Summary of borrowers (with graphs) based on tenure completion
4. Spend recommendation for borrowers (with graphs) - you need to articulate on how you have minimised spend while keeping in mind high repayment rate
5. Any other interesting insights you have derived from the above data.

We recommend that you spend your time in doing some really good work - read the questions carefully a couple of times before you start working on the solution - not all the information is given to you and the assignment is open ended.

Due to high volume of submissions, we may not be able to communicate to you if your submission has been rejected - but you will hear from us within 1 week if the submission is accepted.

Please do not email us about your submission - we will check the google form submissions. There is no need to email us about this.

