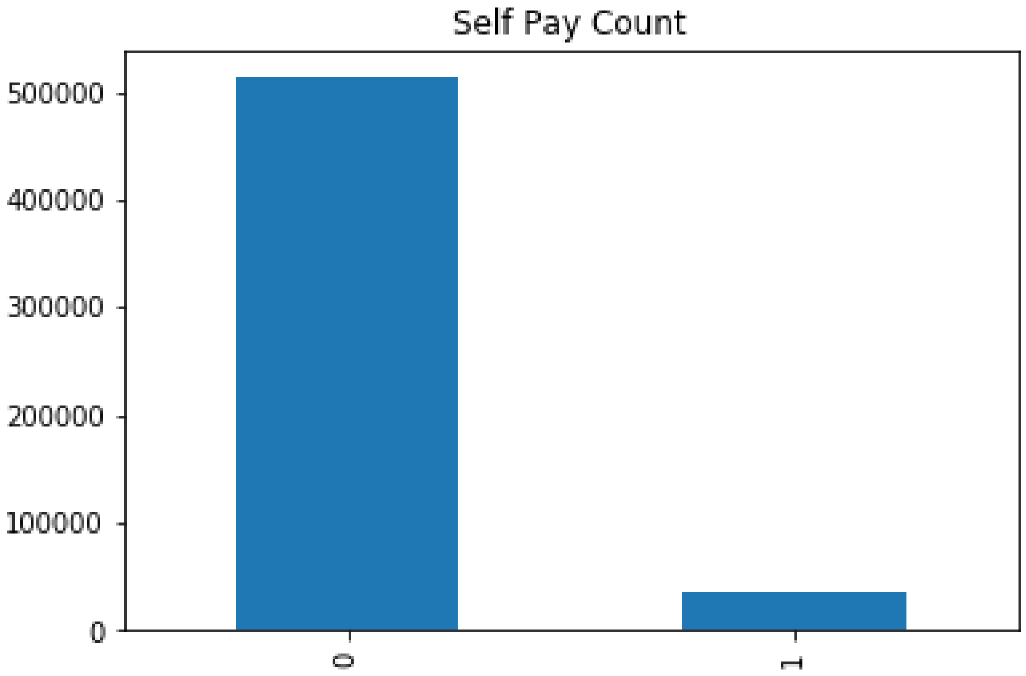
A brief Study by Akshat & Keshav

TVS Credits

Overview of the Data

Highly Imbalanced Dataset present .



Pay = 0 signifies the customer is not a self paying customer
Pay = 1 signifies the customer is a self paying customer

Focus has been paid on

i) Accurate detection of the self paying customers
ii) All the self paying customers should be identified

For this we have decided to use the F1 score(macro) as the metric for all our models

Dealing with the Data

Dropped

Certain features did not have sufficient amount of data. So they were dropped.

i) TIME PERSONAL LOAN
ii) TIME SINCE LIVE PERSONAL
iii) TIME SINCE CLOSED PERSONAL
iv) TIME SINCE LIVE BUSINESS

Certain features were not valuable enough to be analysed

i) CUST_ID
ii) MODEL CODE
iii) DEALER CODE
iv) DOB

Another feature : Type of product was dropped as another feature 'Asset Cost' was taken into consideration

MOLA : Maximum Live Amount
MOLAUL : Maximum amount of live amount , Unsecured loan

Imputation

In some places the columns of the features were missing some data points so we imputed them with the mean.
Since the data points missing were very small in number , imputing with the mean would suffice

i) FUTURE PRINCIPAL
ii) MOLAL
iii) MOLAUL
iv) TIME SINCE LAST LOAN
v) TIME SINCE FIRST CONSUMER LOAN

(vi) Residence was filled with the most common value : OWNED

UNIVARIATE ANALYSIS

Feature	P – value	Inference
App	Extremely Small	Levels in feature are useful
Qualification	0.6e-17	Levels in feature are useful
Area Code	0.0	Levels in feature are useful
MOLAUL	0.006	Not useful
MOLA	0.00371	Levels in feature are useful
Employment	0.01	Levels in feature are useful
Residence	Small	Levels in feature are useful

For Categorical Features chi square tests were used
For Continuous variables ANOVA was used
Any p value less than 0.05 was considered to be of statistical significance

UNIVARIATE ANALYSIS

SALARIED EMPLOYEES TEND TO HAVE BETTER CHANCES OF PAYING BACK

PAY	1	0	
EMPLOYMENT			
SELF	8186	174719	MORE PERCENTAGE OF 0S
SALARIED	10391	92925	LESS PERCENTAGE OF 0S

PEOPLE WITH HOUSES HAVE A LESS CHANCE OF PAYING BACK

PAY	1	0	
HOUSING			
OWNED	15203	236848	94% OF OWNED IS 0
RENT	3374	30796	90% OF RENT IS 0

AS MONTH OF BUSINESS PROGRESSES , DEFAULTERS INCREASE
Will be explained in later slides

Since it is a highly unbalanced data set even small percentage difference can lead to important results

BIVARIATE ANALYSIS

Pearson's Correlation was used for the continuous variables . We dropped the features that were very much correlated.

Another new feature called Score was created using Score = 0.45 * Reliability + 0.35 * Loan Amount + 0.20 * Loan Tenure

Features like Future Principal , MOLAUL , ASSET COST were dropped to prevent data from getting sparse.

Also a new feature was created called Reliability was created by using the below given formula Reliability = (Advance EMIs paid – (2 * Bounced)) *EMI

As Month of business progresses , the defaulters increase. This is shown by a negative correlation between the two features.(-0.62)

Duration is another feature created that shows time between customer acquired and the transaction date. Both the features were then dropped

Duration and reliability also have a negative correlation. (-0.42)

	Loan Tenure	EMI	Loan Amount	Down Payment	Month of Business	Future Principal	MOLA	Time Last Loan	Pay	Duration	Reliabilty
Loan Tenure	1.000000	-0.250239	0.292438	0.507289	0.087081	0.348761	0.150734	0.056339	0.004846	0.029963	0.195659
EMI	-0.250239	1.000000	0.836255	-0.035097	-0.041282	0.550505	0.529761	-0.041421	0.044750	-0.028265	-0.422954
Loan Amount	0.292438	0.836255	1.000000	0.301160	0.014713	0.743920	0.611553	-0.007005	0.045236	-0.002089	-0.298083
Down Payment	0.507289	-0.035097	0.301160	1.000000	-0.011279	0.338882	0.163559	0.001966	0.011790	-0.020722	0.307498
Month of Business	0.087081	-0.041282	0.014713	-0.011279	1.000000	-0.586252	0.012195	0.515469	-0.155761	0.673685	-0.622372
Future Principal	0.348761	0.550505	0.743920	0.338882	-0.586252	1.000000	0.430217	-0.303640	0.131546	-0.406921	0.275943
MOLA	0.150734	0.529761	0.611553	0.163559	0.012195	0.430217	1.000000	-0.052198	0.032760	0.000890	-0.208908
Time Last Loan	0.056339	-0.041421	-0.007005	0.001966	0.515469	-0.303640	-0.052198	1.000000	-0.064265	0.344794	-0.311841
Pay	0.004846	0.044750	0.045236	0.011790	-0.155761	0.131546	0.032760	-0.064265	1.000000	-0.125365	0.085243
Duration	0.029963	-0.028265	-0.002089	-0.020722	0.673685	-0.406921	0.000890	0.344794	-0.125365	1.000000	-0.425147
Reliabilty	0.195659	-0.422954	-0.298083	0.307498	-0.622372	0.275943	-0.208908	-0.311841	0.085243	-0.425147	1.000000

FEATURE GENERATION

The features created were

i) Duration : Time in days between the customer acquisition

ii) Score

iii) Reliability : A measure of payment history

A 0.81 correlation was found which echoes our hypothesis that a higher score leads to better reliability.

Reliability : The number of cheques bounced were given higher priority as most of the people had the same number of advance EMIs paid and cheques bounced. Also people who defaulted on higher EMI were putting the company at a greater risk .
Hence the EMI was also included.

Reliability Value

Pay	1	0
Mean	-3534	-6551
Median	-1334	-4568

Score

Pay	1	0
Mean	5326	3531
Median	5497	3764

The longer the Duration , the lower the reliability
This was shown from the previous Pearson Correlation .

The Month of Business is very closely related to the Reliability of the customer. When looked at closely , it can be concluded that the reliability is a very strong indicator of the payment history of the person.

UNBALANCED DATA

We used the technique of SMOTE to cure the problem of unbalanced dataset.
As an example a random forest classifier was first trained on the original dataset and then on the SMOTE data.

Random Forest Classifier trained on the original data set gave an F1 score (macro) of 0.54
The same model on the SMOTE data gave us an F1score of 0.56.

Also care has been taken not to use the oversampling technique on the test data. The models have been trained on the transformed data only to be used on the untransformed test data.

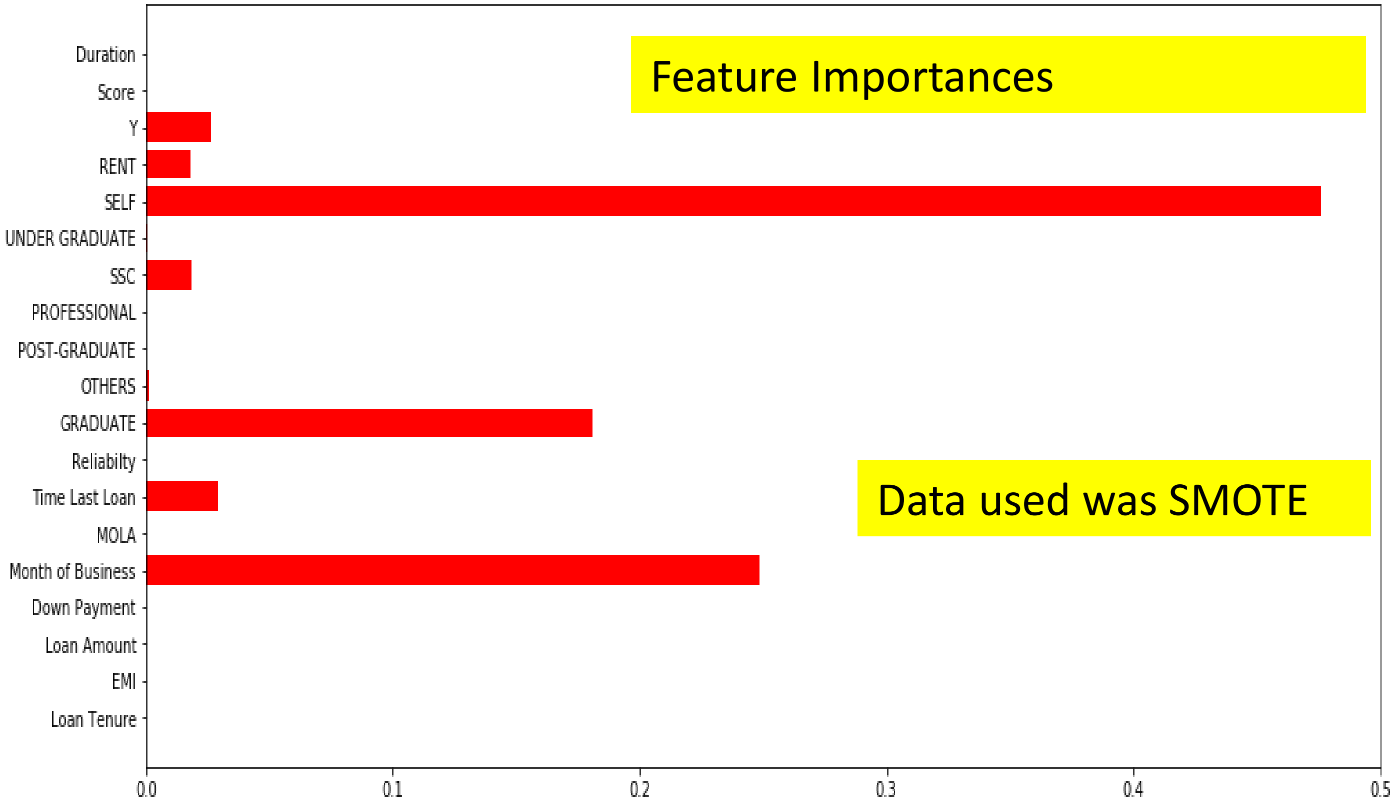
Dummies were created for all the categorical features for preparing the data for machine learning models.

MOLA had a few outliers which were corrected.

MODELS USED

A Random Forest was used to train and test the data.
Following are the results

	SMOTE	F1 SCORE MACRO
Random Forest	TRAINED ON ORIGINAL	0.49
Random Forest	TRAINED ON SMOTE	0.59



Look at the anomaly here.
The Month of Business is really important for the classification. But the feature Reliability is not.
This is because the Month of Business reflects of the properties of the Reliability and thus the Random Forest need not use the additional feature of Reliability.

INFERENCE FROM THE DATA

i) The employment status of the person matters a lot . Salaried people or fixed income people are categorized as self pay customers very often
ii) The month of the business is very important. As the months progress more and more number of people tend to fall out from the self paying zone.
iii) Qualification matters. Graduates tend to get stable jobs and hence their income is easily determined.
iv) Whether the housing is rented or not can play a big role in the classification
v) Your time since last loan also plays an important role in the classification problem

Loan Amount , Loan tenure are either in the Reliability feature or the Month of Business feature .
Which is why our model did not find it necessary to include those features to classify the customers.

Thank You!