Our Business Case

Profiling Customer through Credit risk assessment

Brought to you by: Group 5, Dixy Chicken :)

Agenda for this 7 min

1. Overview of our project
    a. Business objective, use case
    b. Project plan
2. Data exploration, preparation, cleaning
3. Data Modelling
4. Model Evaluation and Improvements

Overview

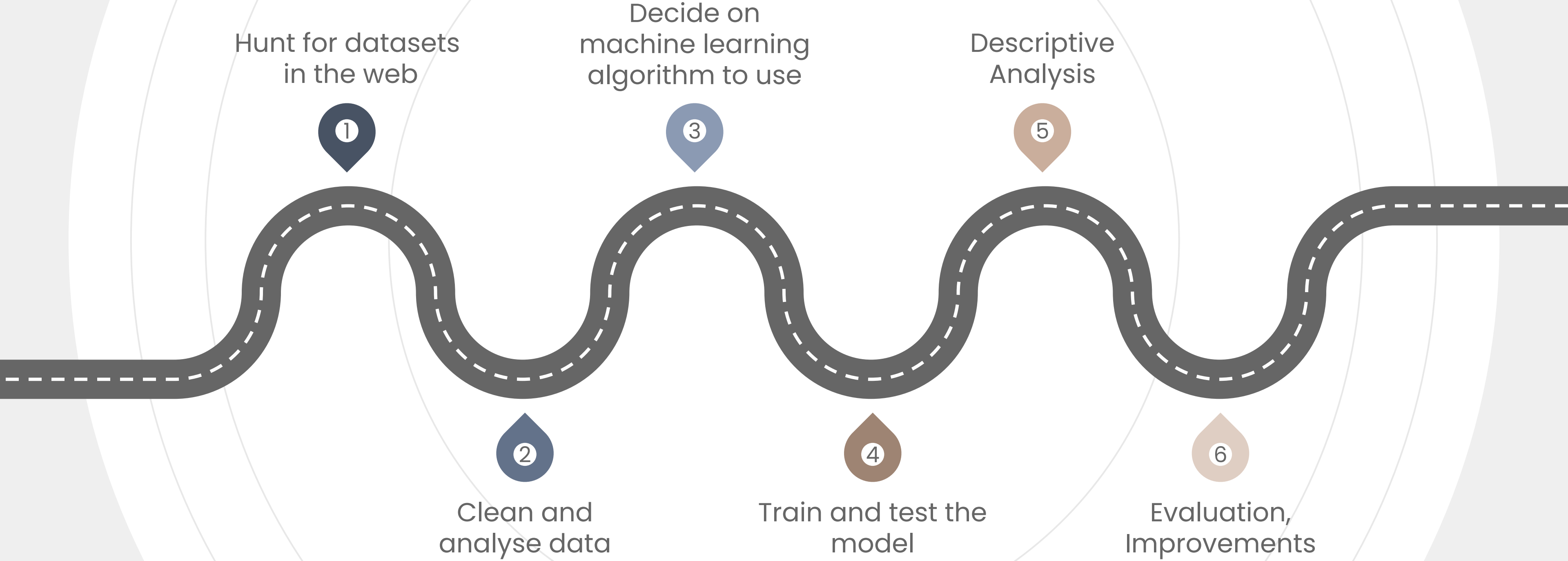
Business Case

Profiling Customer through Credit risk assessment

Use Case

Automation of classifying customers into their loan grade

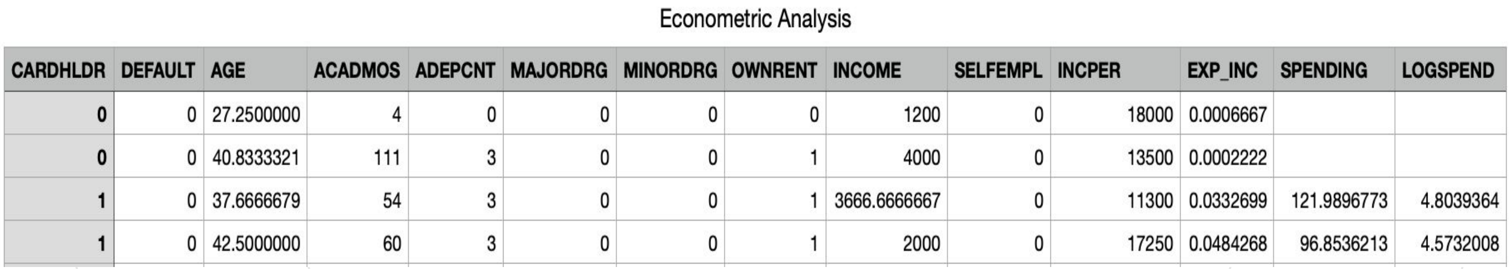
Project Plan Route



1 Data Extraction

Rejected dataset

Credit Card Data from book "Econometric Analysis"



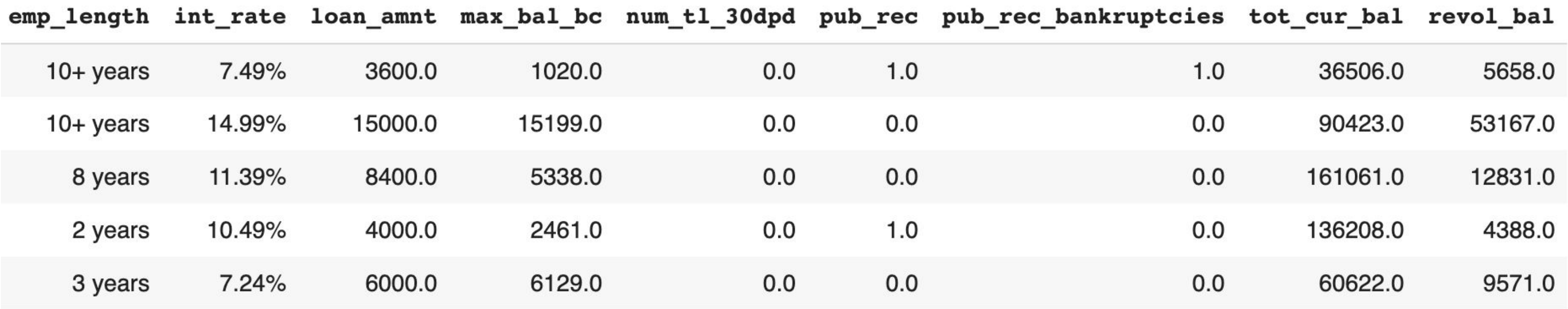
Econometric Analysis

CARDHLDR	DEFAULT	AGE	ACADMOS	ADEPCNT	MAJORDRG	MINORDRG	OWNRENT	INCOME	SELFEMPL	INCPER	EXP_INC	SPENDING	LOGSPEND
0	0	27.2500000	4	0	0	0	0	1200	0	18000	0.0006667		
0	0	40.8333321	111	3	0	0	1	4000	0	13500	0.0002222		
1	0	37.6666679	54	3	0	0	1	3666.6666667	0	11300	0.0332699	121.9896773	4.8039364
1	0	42.5000000	60	3	0	0	1	2000	0	17250	0.0484268	96.8536213	4.5732008

- No provision of data dictionary - description of the column heads - hence there is a risk of misinterpretation
- Lack many relevant and crucial data, exp. Interest rate on loans, loan amount, employment length, etc

Employed Dataset

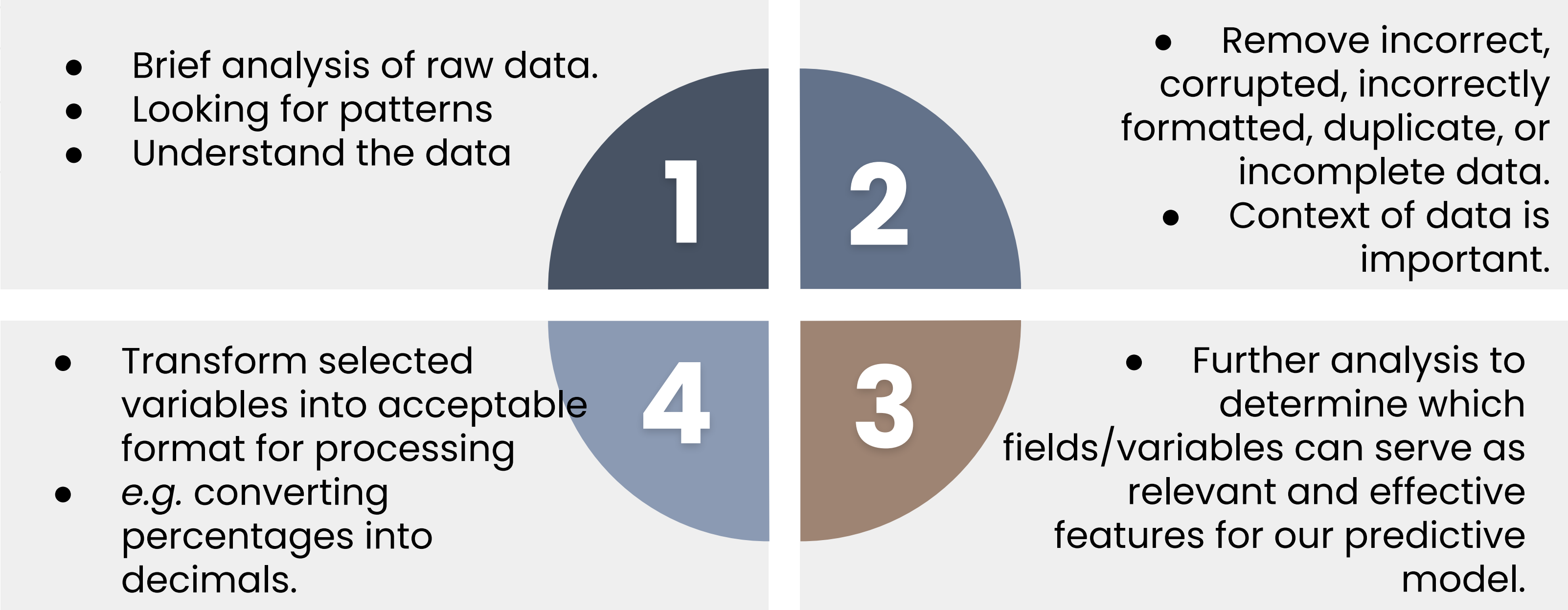
Lending Club Loan Data



emp_length	int_rate	loan_amnt	max_bal_bc	num_tl_30dpd	pub_rec	pub_rec_bankruptcies	tot_cur_bal	revol_bal
10+ years	7.49%	3600.0	1020.0	0.0	1.0	1.0	36506.0	5658.0
10+ years	14.99%	15000.0	15199.0	0.0	0.0	0.0	90423.0	53167.0
8 years	11.39%	8400.0	5338.0	0.0	0.0	0.0	161061.0	12831.0
2 years	10.49%	4000.0	2461.0	0.0	1.0	0.0	136208.0	4388.0
3 years	7.24%	6000.0	6129.0	0.0	0.0	0.0	60622.0	9571.0

- With more than 100 columns and a detailed data dictionary, this dataset is much more comprehensive - able to analyse and filter the crucial information
    - **num_tl_30dpd** - Number of accounts currently 30 days past due (updated in past 2 months)
    - **pub_rec** - Number of derogatory public records
    - **pub_rec_bankruptcies** - Number of public record bankruptcies

Clean, Analyse & Transform Data



1

- Brief analysis of raw data.
- Looking for patterns
- Understand the data

2

- Remove incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data.
- Context of data is important.

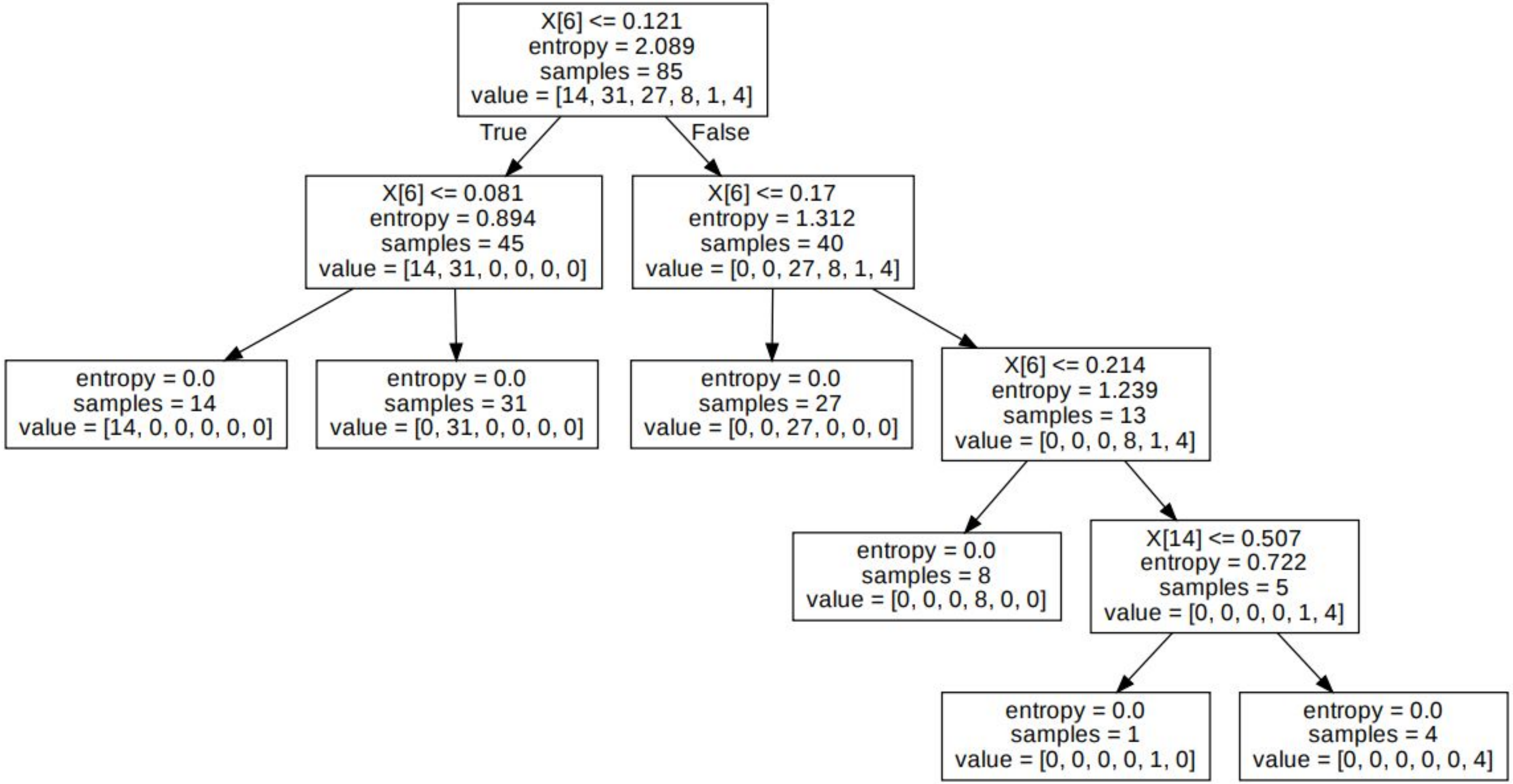
3

- Further analysis to determine which fields/variables can serve as relevant and effective features for our predictive model.

4

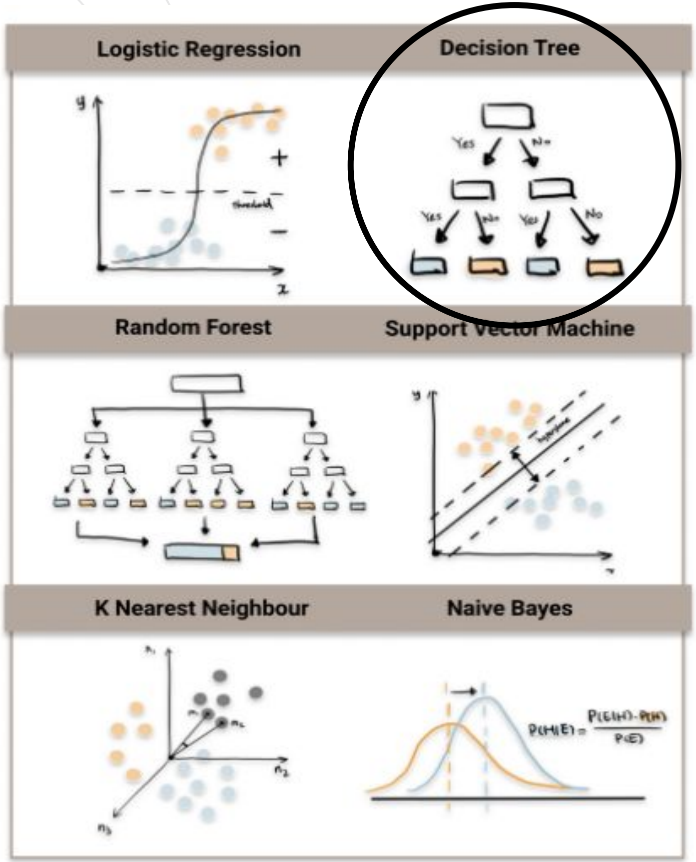
- Transform selected variables into acceptable format for processing
- *e.g.* converting percentages into decimals.

With int_rate and only using 0.01% of data for training



3

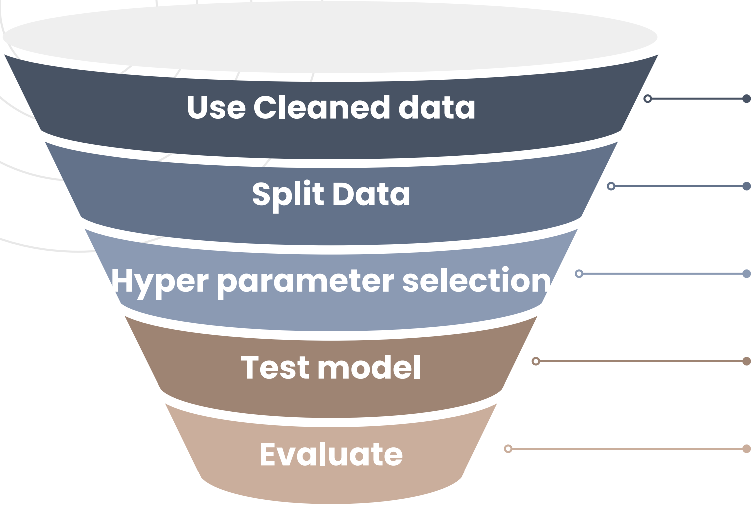
Machine Learning Algo



Used **Decision Tree** algorithm for multiclass classification for this case

Note: Possible to use other algorithms but due to time constraints we went with this

4 Decision Tree Learning



Use Cleaned data: Use the data cleaned during the previous step

Split Data: Split data to training and testing data

Hyper parameter selection: Conduct experiment to determine which parameter values yields the best accuracy

Test model: Test model with parameters selected

Evaluate: What we can do to improve accuracy

Our Model Accuracy

40%

5

Descriptive Analysis

Discovered that column **int_rate significantly affects** the **output**. Giving us very high accuracy for the model

WITH int_rate	**WITHOUT** int_rate
- Accuracy minimum **90%**	- Accuracy **stable** around **40%**

6

Evaluation, Improvements

What could we do better

- ★ Deal with the NaN values better
- ★ Used multiple machine learning algorithms and evaluate which has the best accuracy, No Free Lunch Theorem (David Wolpert)
- ★ Find better relationships/trends in the data to give more insights

Thank you for listening!

Richard Lee

Sherri Chuah

Shirlyn