CREDIT CARD APPROVAL:

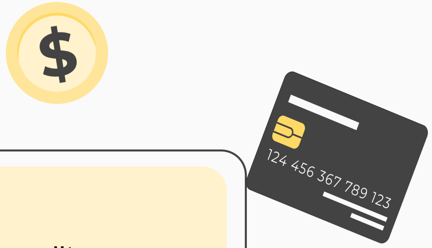
PREDICTIVE ANALYSIS

By: Manuel Iglesias

PROJECT OUTLINE

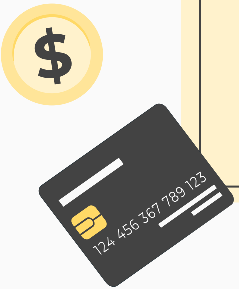
For our project, we used a predictive model to determine the likelihood of a person's credit card application being denied or approved.

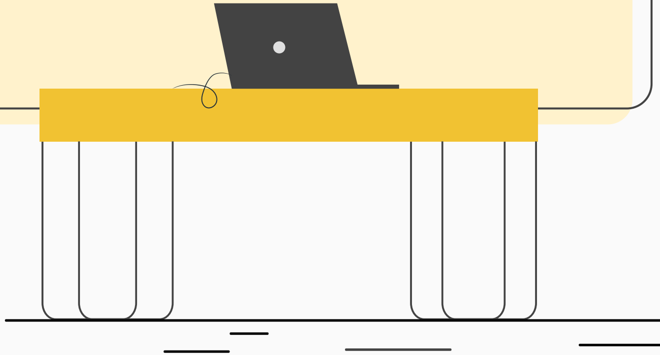
1. Download the dataset from Kaggle
2. Using Google Collab Clean the DataSet and save it
3. Use Logistic Regression
4. Compare the classification model & diagram
5. **Analyze the data & determine valuable insights**



WHY?

- **For financial institutions:** Our model enables informed decisions on credit applications, reducing losses from defaults and enhancing risk management.
- **Value of Predictors**: Provides insight into which factors are most critical in the application review process, improving criteria for decision-making.
- **For potential credit card applicants**: Helps them understand which factors to focus on for application approval, or when it's best to avoid applying.
- **Overall Benefit**: Aids both lenders and applicants in navigating the credit application process more effectively, highlighting key aspects to consider for both applying and approving credit cards.





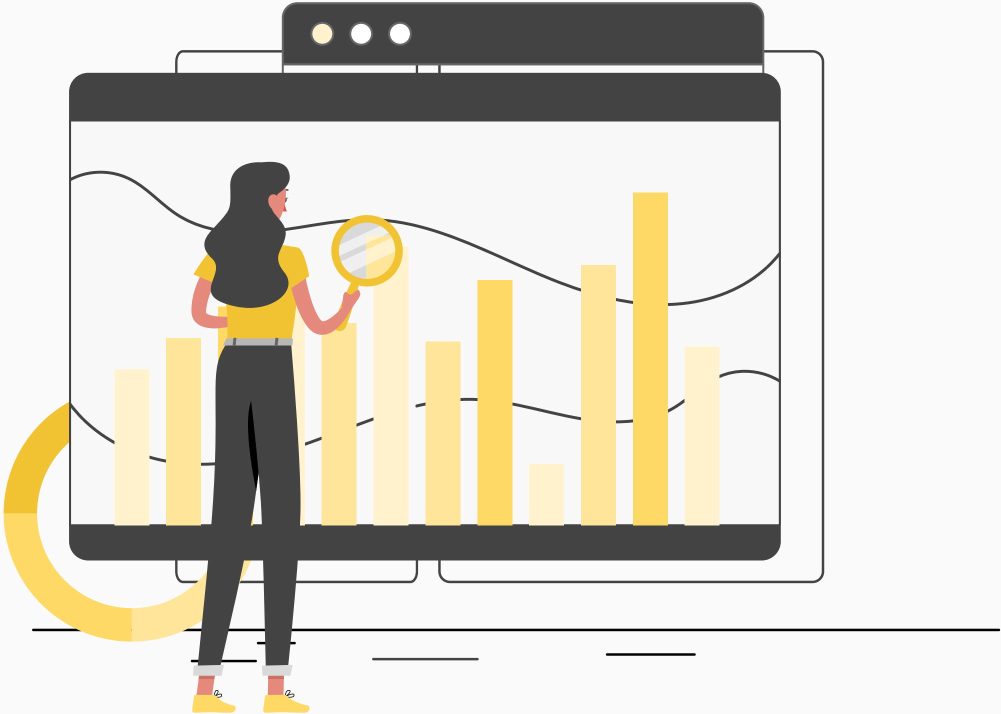
DATA AND DIDA FRAMEWORK

Dataset Description:

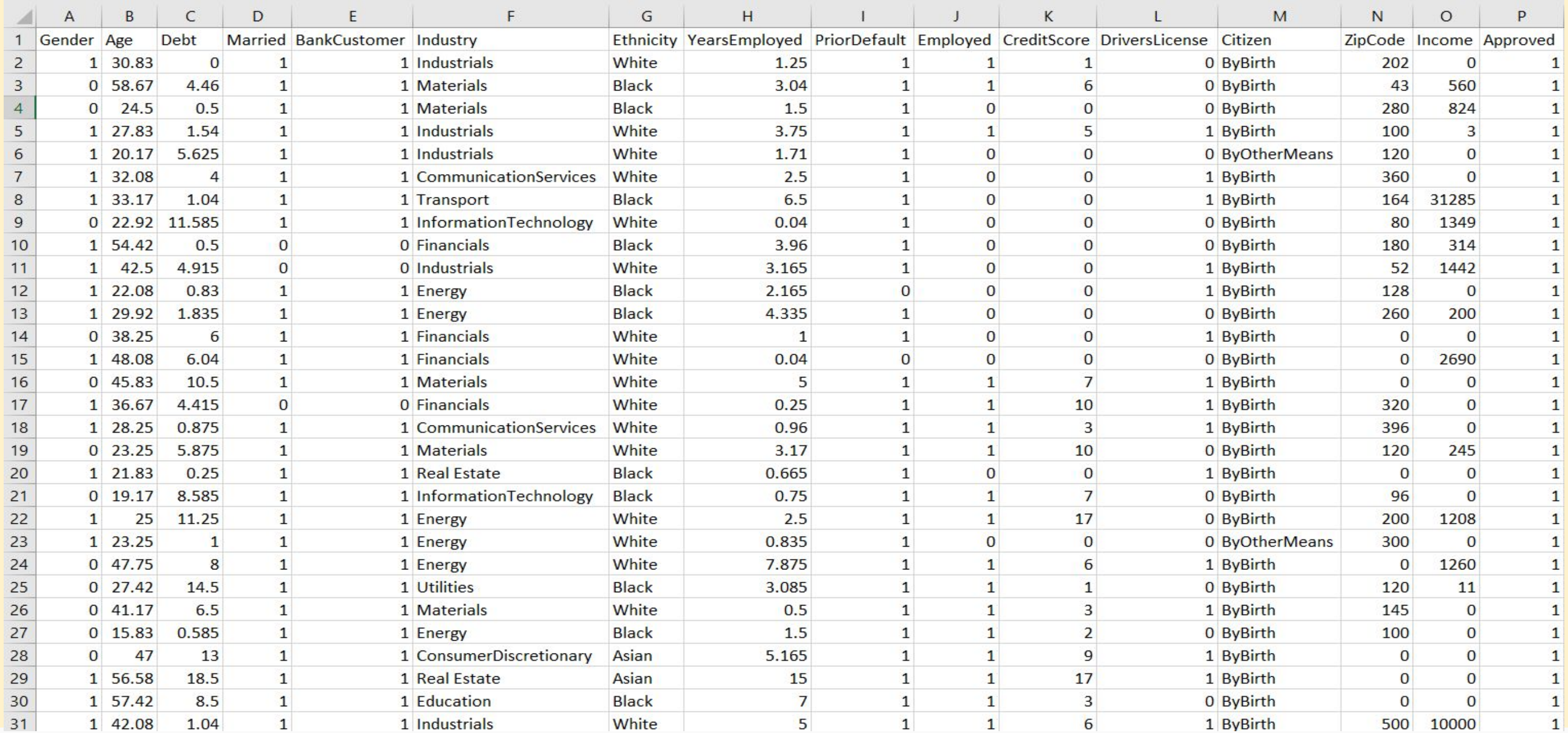
- 690 Individuals with diverse backgrounds
- Mix of categorical and continuous variables and is divided among 16 different variables

DIDA Framework:

- **Data**: Age, Gender, Income, Outstanding Debt, Years Employed, Prior Defaults. etc.
- **Insights**: Calculate the probability of a person's credit card application being approved based on the provided data
- **Decision**: Approval Threshold Setting
- **Advantages**: Risk Assessment, Customer Segmentation, Bias Reduction



DATA EXPLORATION: SAMPLE OF DATASET



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Gender	Age	Debt	Married	BankCustomer	Industry	Ethnicity	YearsEmployed	PriorDefault	Employed	CreditScore	DriversLicense	Citizen	ZipCode	Income	Approved
2	1	30.83	0	1	1	Industrials	White	1.25	1	1	1	0	ByBirth	202	0	1
3	0	58.67	4.46	1	1	Materials	Black	3.04	1	1	6	0	ByBirth	43	560	1
4	0	24.5	0.5	1	1	Materials	Black	1.5	1	0	0	0	ByBirth	280	824	1
5	1	27.83	1.54	1	1	Industrials	White	3.75	1	1	5	1	ByBirth	100	3	1
6	1	20.17	5.625	1	1	Industrials	White	1.71	1	0	0	0	ByOtherMeans	120	0	1
7	1	32.08	4	1	1	CommunicationServices	White	2.5	1	0	0	1	ByBirth	360	0	1
8	1	33.17	1.04	1	1	Transport	Black	6.5	1	0	0	1	ByBirth	164	31285	1
9	0	22.92	11.585	1	1	InformationTechnology	White	0.04	1	0	0	0	ByBirth	80	1349	1
10	1	54.42	0.5	0	0	Financials	Black	3.96	1	0	0	0	ByBirth	180	314	1
11	1	42.5	4.915	0	0	Industrials	White	3.165	1	0	0	1	ByBirth	52	1442	1
12	1	22.08	0.83	1	1	Energy	Black	2.165	0	0	0	1	ByBirth	128	0	1
13	1	29.92	1.835	1	1	Energy	Black	4.335	1	0	0	0	ByBirth	260	200	1
14	0	38.25	6	1	1	Financials	White	1	1	0	0	1	ByBirth	0	0	1
15	1	48.08	6.04	1	1	Financials	White	0.04	0	0	0	0	ByBirth	0	2690	1
16	0	45.83	10.5	1	1	Materials	White	5	1	1	7	1	ByBirth	0	0	1
17	1	36.67	4.415	0	0	Financials	White	0.25	1	1	10	1	ByBirth	320	0	1
18	1	28.25	0.875	1	1	CommunicationServices	White	0.96	1	1	3	1	ByBirth	396	0	1
19	0	23.25	5.875	1	1	Materials	White	3.17	1	1	10	0	ByBirth	120	245	1
20	1	21.83	0.25	1	1	Real Estate	Black	0.665	1	0	0	1	ByBirth	0	0	1
21	0	19.17	8.585	1	1	InformationTechnology	Black	0.75	1	1	7	0	ByBirth	96	0	1
22	1	25	11.25	1	1	Energy	White	2.5	1	1	17	0	ByBirth	200	1208	1
23	1	23.25	1	1	1	Energy	White	0.835	1	0	0	0	ByOtherMeans	300	0	1
24	0	47.75	8	1	1	Energy	White	7.875	1	1	6	1	ByBirth	0	1260	1
25	0	27.42	14.5	1	1	Utilities	Black	3.085	1	1	1	0	ByBirth	120	11	1
26	0	41.17	6.5	1	1	Materials	White	0.5	1	1	3	1	ByBirth	145	0	1
27	0	15.83	0.585	1	1	Energy	Black	1.5	1	1	2	0	ByBirth	100	0	1
28	0	47	13	1	1	ConsumerDiscretionary	Asian	5.165	1	1	9	1	ByBirth	0	0	1
29	1	56.58	18.5	1	1	Real Estate	Asian	15	1	1	17	1	ByBirth	0	0	1
30	1	57.42	8.5	1	1	Education	Black	7	1	1	3	0	ByBirth	0	0	1
31	1	42.08	1.04	1	1	Industrials	White	5	1	1	6	1	ByBirth	500	10000	1

VARIABLES & LISTS

DEPENDENT VARIABLE

Approved

RVAR LIST

Industry, Ethnicity, Citizen, Drivers License, Zip code

CVAR LIST

Gender, Married, Bank Customer, Prior Default, Employed

NVAR LIST

Age, Debt, Years Employed, Credit Score, Income

R DUMMIES

Gender_1, Married_1, BankCustomer_1, PriorDefault_1, Employed_0

DATA MINING TECHNIQUES

A

Logistic Regression

Why:

The application of Logistic Regression enables:

- Enhanced data manipulation
- Deeper insight into variable distribution and interrelationships.
- Key to our analysis is the examination of predictor coefficients
  - their influence on approval probabilities
- Use K- fold cross validation to prevent overfitting

B

Classification Tree

Why:

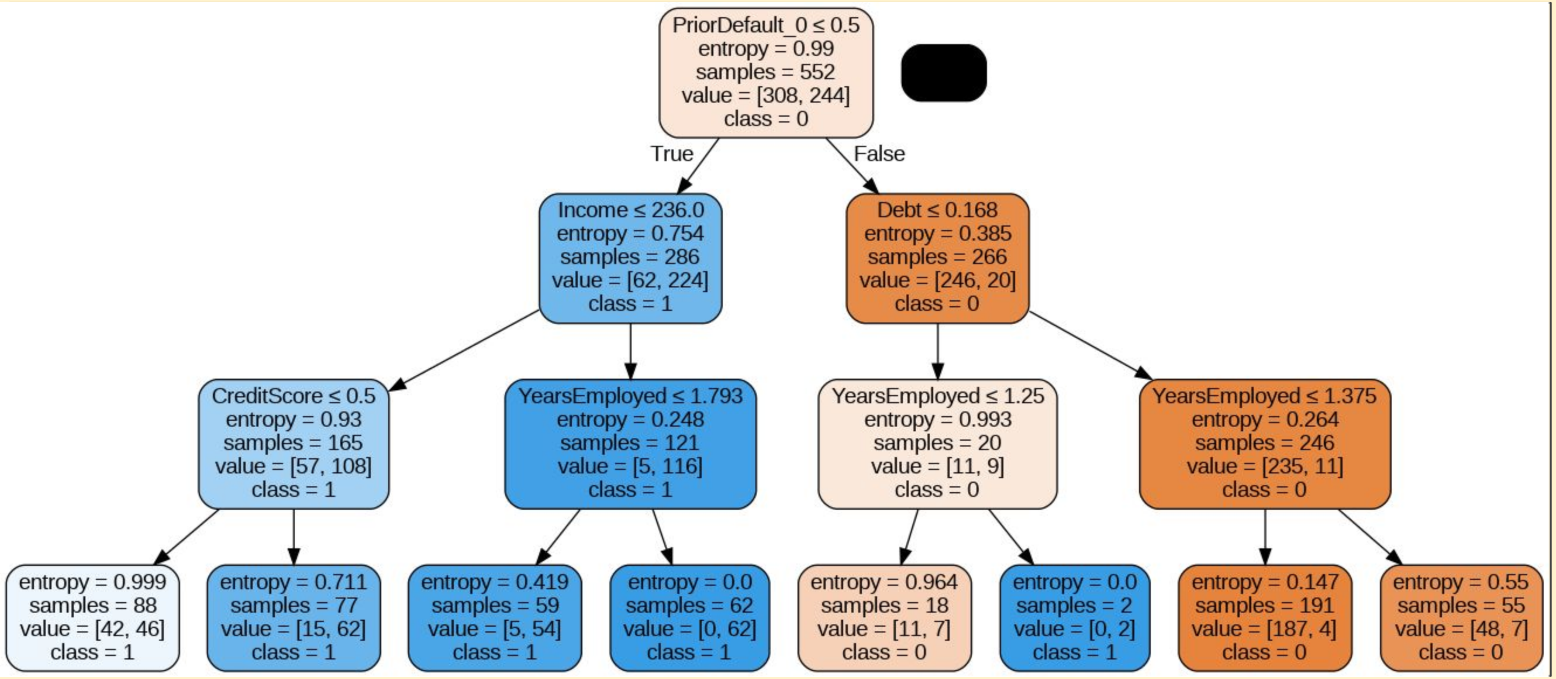
The application of the Classification Tree enables:

- It adeptly handles both numerical and categorical data types.
- Effectively captures non-linear relationships within the dataset.
- Ideally suited for binary classification tasks, such as approving or rejecting credit card applications.
- Highlights the importance of various predictors, aligning with our project's objective to assess each variable's impact on the likelihood of credit application approval.

VARIABLES

VARIABLE NAME	alpha	kfolds	N candidates	Min alpha	Max alpha	Max iter	Max depth	Min depth	Test part size
LOG	10	5	1000	0.001	100	2000			0.2
TREE		5					8	1	0.2

BEST PRUNED TREE



CLASSIFICATION TREE STATS

3

Tree levels

0.933

Roc_auc_score

LEAF INTERPRETATIONS - USEFULNESS

LEAF NODE ID = 4

- Path = ['BankCustomer_0 <= 0.5', 'Income > 236.0', 'YearsEmployed > 1.7925000190734863']
- sample = 62
- value = [0, 62]
- class = 1

PROBABILITY = 100

LEAF NODE ID = 6

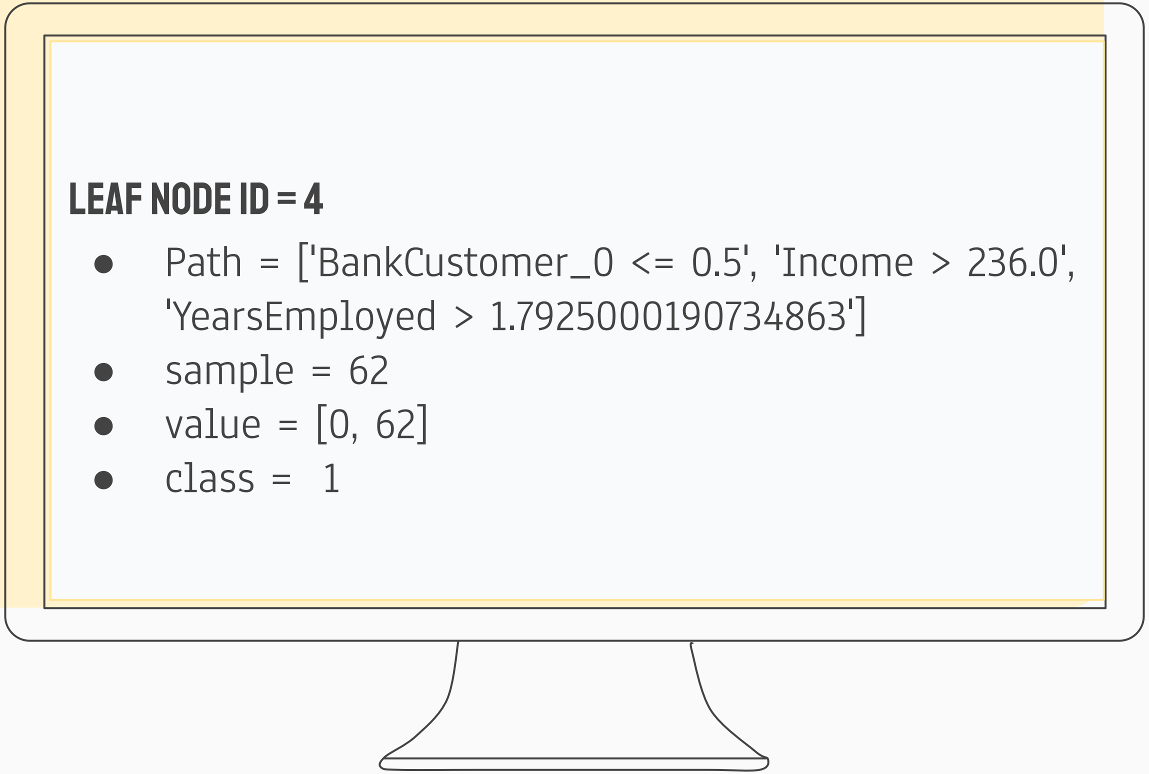
- Path = ['BankCustomer_0 > 0.5', 'Debt <= 0.16750000417232513', 'YearsEmployed > 1.25']
- sample = 2
- value = [0, 2]
- class = 1

PROBABILITY = 100

LEAF NODE ID = 3

- Path = ['BankCustomer_0 <= 0.5', 'Income > 236 'YearsEmployed <= 1.7925000190734863']
- sample = 59
- value = [5, 54]
- class = 1

PROBABILITY = 0.91



LEAF NODE ID = 4

- Path = ['BankCustomer_0 <= 0.5', 'Income > 236.0', 'YearsEmployed > 1.7925000190734863']
- sample = 62
- value = [0, 62]
- class = 1

MOST USEFUL AND SIGNIFICANT RULE

LOG REGRESSION

ALPHA = 0.1

Age	-0.185558
Debt	-0.093025
YearsEmployed	0.343885
CreditScore	0.686659
Income	1.966801
Gender_0	0.042923
Married_0	1.957422
BankCustomer_0	-2.377422
PriorDefault_0	-3.285441
Employed_1	0.338168
Intercept	1.263581

ALPHA = 0.01

Age	-0.185097
Debt	-0.094837
YearsEmployed	0.337581
CreditScore	0.685287
Income	2.019125
Gender_0	0.041300
Married_0	2.501183
BankCustomer_0	-2.923406
PriorDefault_0	-3.303809
Employed_1	0.346560
Intercept	1.270219

ALPHA = 0.001

Age	-0.185076
Debt	-0.095033
YearsEmployed	0.337022
CreditScore	0.685157
Income	2.024419
Gender_0	0.041202
Married_0	2.556318
BankCustomer_0	-2.978744
PriorDefault_0	-3.305608
Employed_1	0.347365
Intercept	1.270888

LOG REGRESSION OPTIMAL

ALPHA = 0.3012973

Age	-0.184025
Debt	-0.088207
YearsEmployed	0.347703
CreditScore	0.697204
Income	1.862477
Gender_0	0.035825
Married_0	1.314176
BankCustomer_0	-1.730639
PriorDefault_0	-3.257112
Employed_1	0.314951
Intercept	1.258711

0.954

ROC AUC SCORE

0.301

OPTIMAL ALPHA

MOST IMPORTANT VARIABLES

	CREDIT SCORE	PRIOR DEFAULT 0	YEARS EMPLOYED	INCOME	EMPLOYED _1
ALPHA = 10	✓	✓	✓	✓	✗
ALPHA = 33	✓	✓	✓	✗	✗
ALPHA = 52	✓	✓	✗	✗	✗
ALPHA = 95	✓	✗	✗	✗	✗
ALPHA = 112	✗	✗	✗	✗	✗

CLASSIFICATION TREE VS LOGISTIC REGRESSION

CLASSIFICATION TREE

0.933

LOGISTIC REGRESSION

0.9549

A higher AUC means more accurate, meaning that the logistic regression is better at predicting credit score acceptance.

In log regression the most important measurement was credit score whilst in the classification tree the most effective path starts with whether the user is a bank customer or not.

FINAL INSIGHTS

- **Logistic Regression Model:** Accurately predicts credit card application outcomes.
- **Key Factors:** Employment status, income, years employed, no prior defaults, credit score.
- Classification Tree Insight: Effective rule - approve if not a bank customer, income > 236, years employed > 1.79.
- **Application:** Aids financial institutions in risk assessment, decision-making efficiency, credit limit adjustments, targeted marketing.
- **Performance:** 88% accuracy, 70% average probability for predicted approvals.

QUESTIONS?

THANK YOU!