DETECTING KICKSTARTER PROJECT SUCCESS

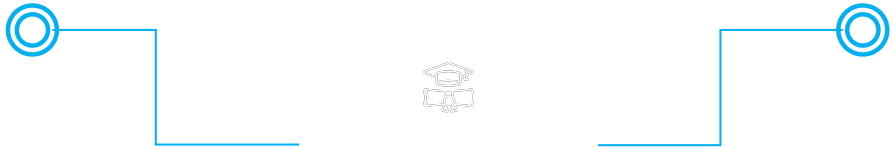
Giovanni Piva - Roberto Vicentini

DIPARTIMENTO MATEMATICA

WHAT IS IT?

Online platform that allows creators to seek financial support from a community of backers who are interested in their projects

Based on the principles of crowdfunding with "All or Nothing" model

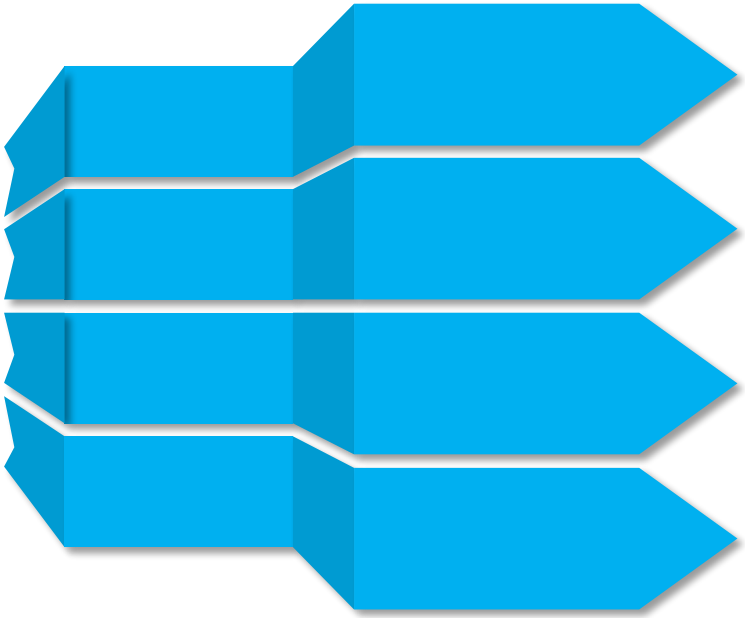


KICKSTARTER

Backers, not only provide financial support, but also become part of the project's journey and gain access to unique rewards and experiences

A project is considered successful if it meets or exceeds its funding goal within the specified campaign timeframe.

WHAT ARE WE GOING TO DO?



01 DATA PRESENTATION AND PREPROCESSING

02 CLEANED DATASET EXPLORATION

03 PREDICTION MODELS

04 COMPARISON AND CONCLUSION

DATA PRESENTATION

We started collecting a Dataset containing data of **378661** Kickstarter projects from all over the world.

VARIABLES MEANING	
Name	Meaning
ID	Project ID
Name	Project Name
Category	Project's Subcategory
Main Category	Project's Main Category
Currency	Currency for the funding
Deadline	Project's Deadline
Goal	Project's Goal
Pledged	Total Money Pledged by the Project
State	Final State of the Project
Backers	Final Project's Backers
Usd Pledged	Conversion in US dollars of the pledged column (conversion done by kickstarter)
Usd Pledged Real	Conversion in US dollars of the pledged column (conversion from Fixer.io API)
Usd Goal Real	Conversion in US dollars of the goal column (conversion from Fixer.io API)

OUR GOAL

The goal is to predict whether an European Kickstarter project will be successful.

SUCCESS

FAILURE

WHY EUROPE?

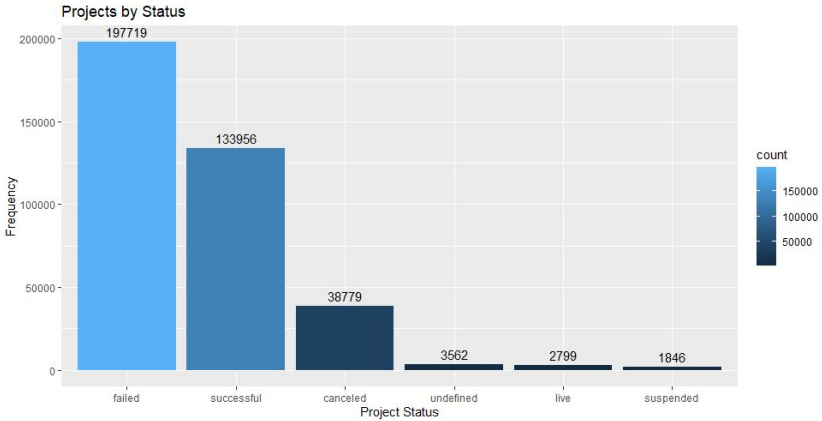
01 WE ARE EUROPEAN

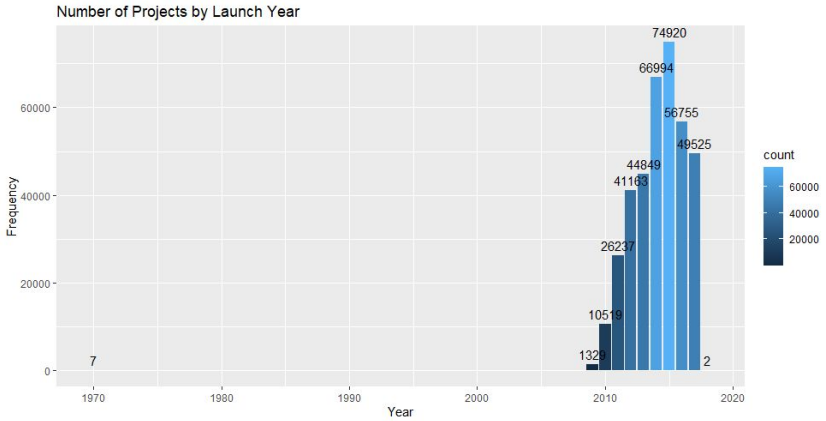
02 EU DIVIDED BY COUNTRY INFORMATION

03 USA PROJECTS AS FINAL TEST

DATA CLEANING

- AFTER REMOVING ALL **NaN** VALUES-





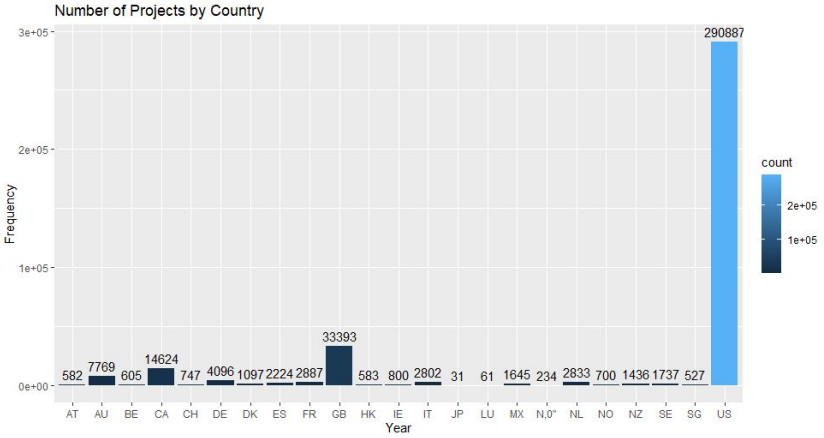
Status

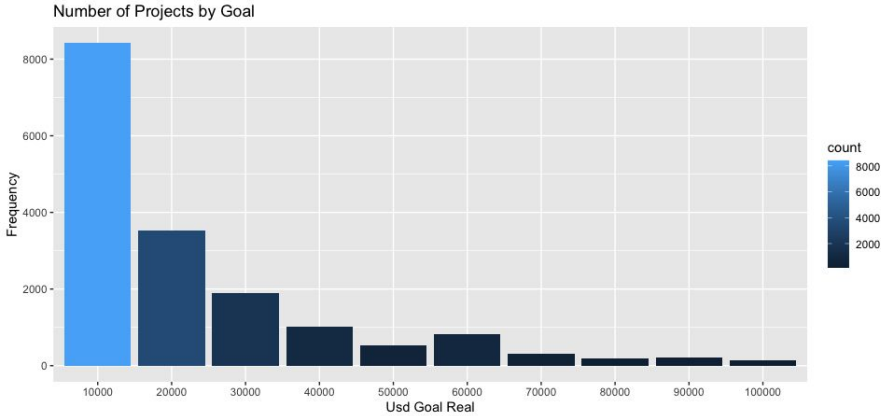
Transform all the variables different from *Success* as *Failure*

Year

Remove outliers of 1970 and 2018

DATA CLEANING





Country

Remove all nation **not** in Europe

Goal

Maintain only Kickstarter project from **1000$** to **100k$**

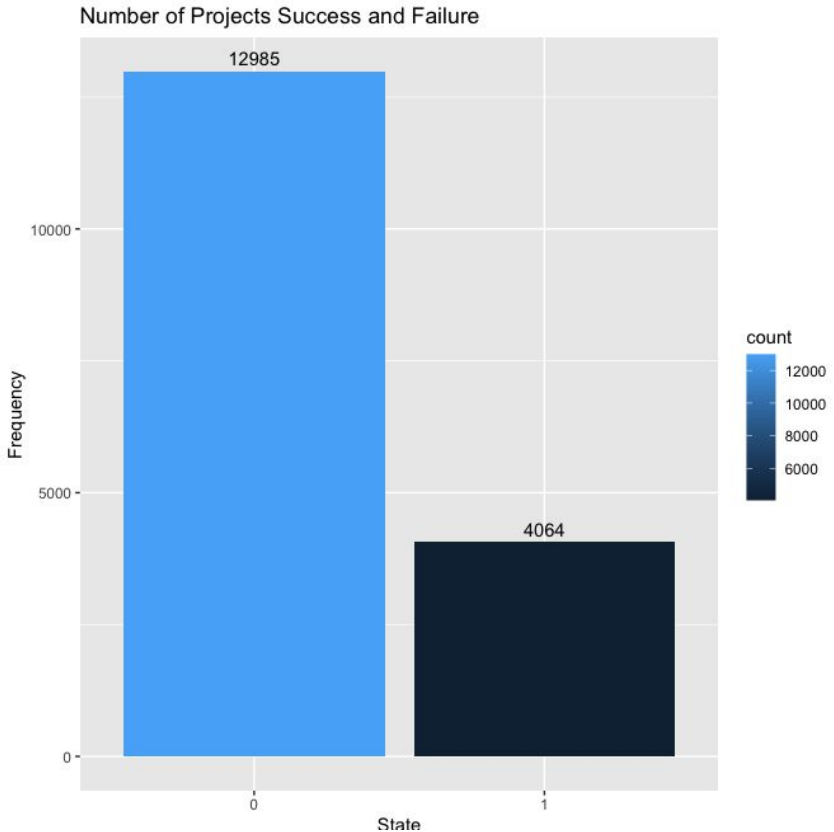
DATA EXPLORATION

01 Success and Failure Balance

02 Success Rate **by Country**

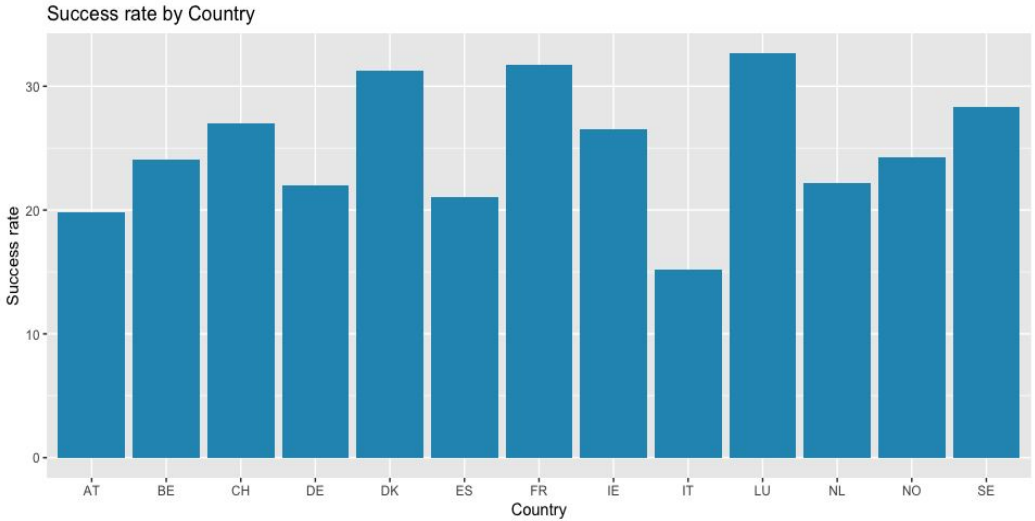
03 Number of Projects and Success Rate **by Year**

04 Goal and Success Rate **by Category**



DATA BALANCE CHECK

The classes **Success** and **Failure** are not balanced.
76% vs **24%**

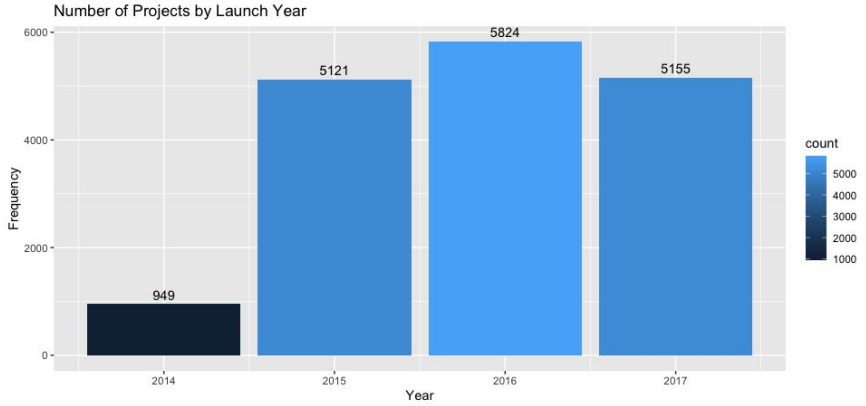


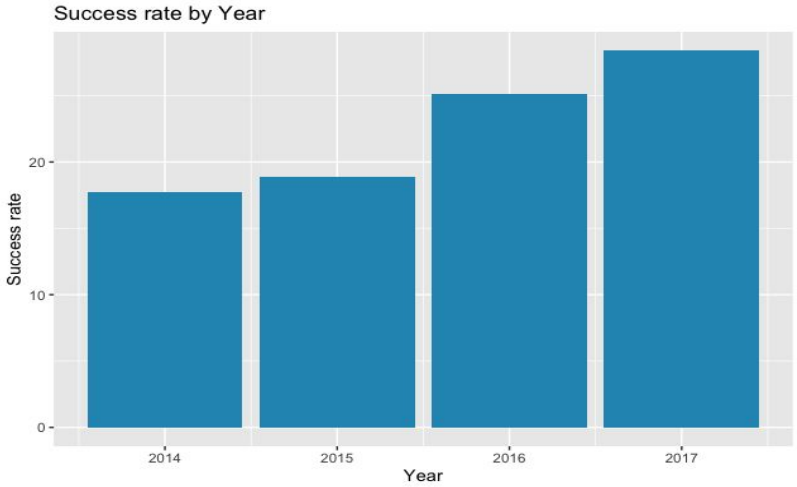
COUNTRY INSIGHT

What Country has the higher Success Rate?

YEAR INSIGHTS

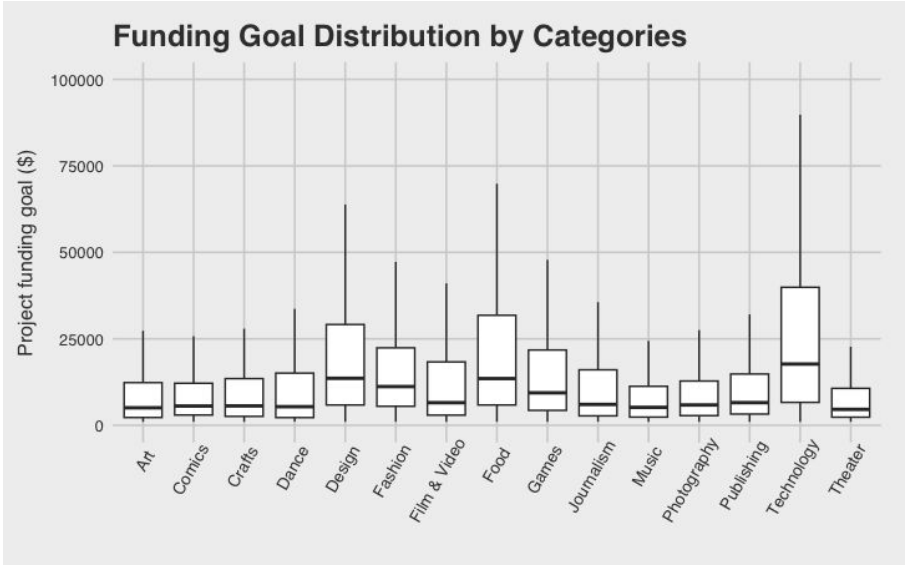
- What Year has the highest number of Launched Projects?
- What Year has the highest number of Success Rate?

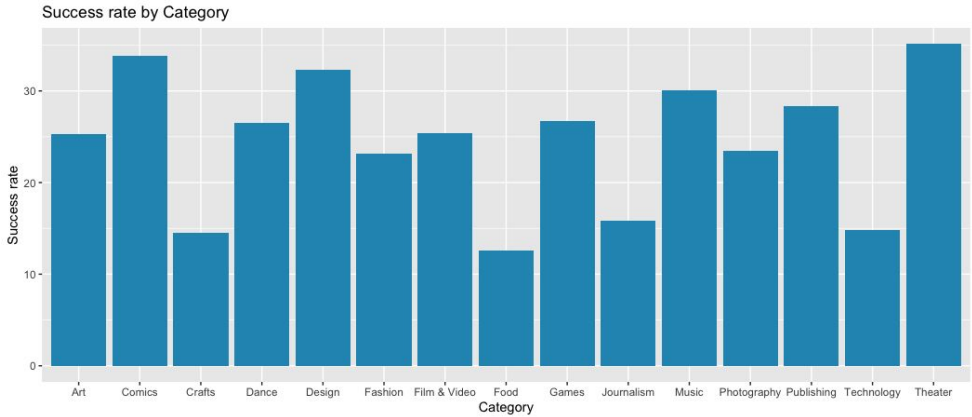




CATEGORY INSIGHTS

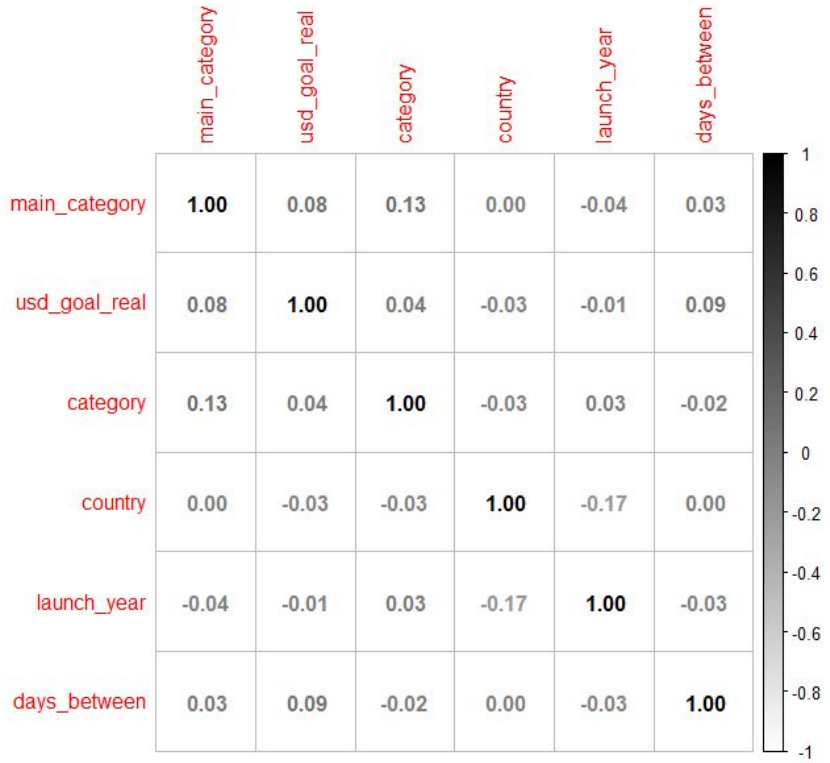
- What Category requires the highest Funding Goal?
- What Category has the highest number of Success Rate?





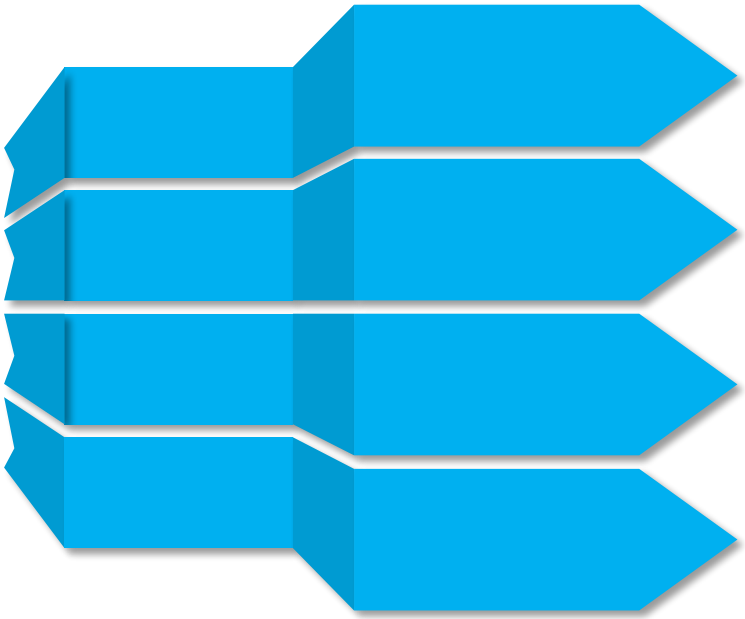
DATA ANALYSIS

We analyze the correlation between pairs of different features that will be used for our predictions.



MODEL EVALUATION

-METRICS-



01 ACCURACY

02 PRECISION AND RECALL

03 ROC CURVE

04 ERROR RATE AND FALSE NEGATIVE RATE

GLM:
GENERAL SETTINGS

01 **AIM:** we are estimating the probability of a kickstarter project to be a success

02 **VIF:** we consider the Variance Inflation Factor for the remotion of the collinearity

03 **STEPWISE:** possibility to apply the Stepwise Selection approach for including and excluding iteratively the covariance inside the model

04 **BALANCING:** we have also trained aor model on the balanced dataset

SIMPLE LOGISTIC REGRESSION

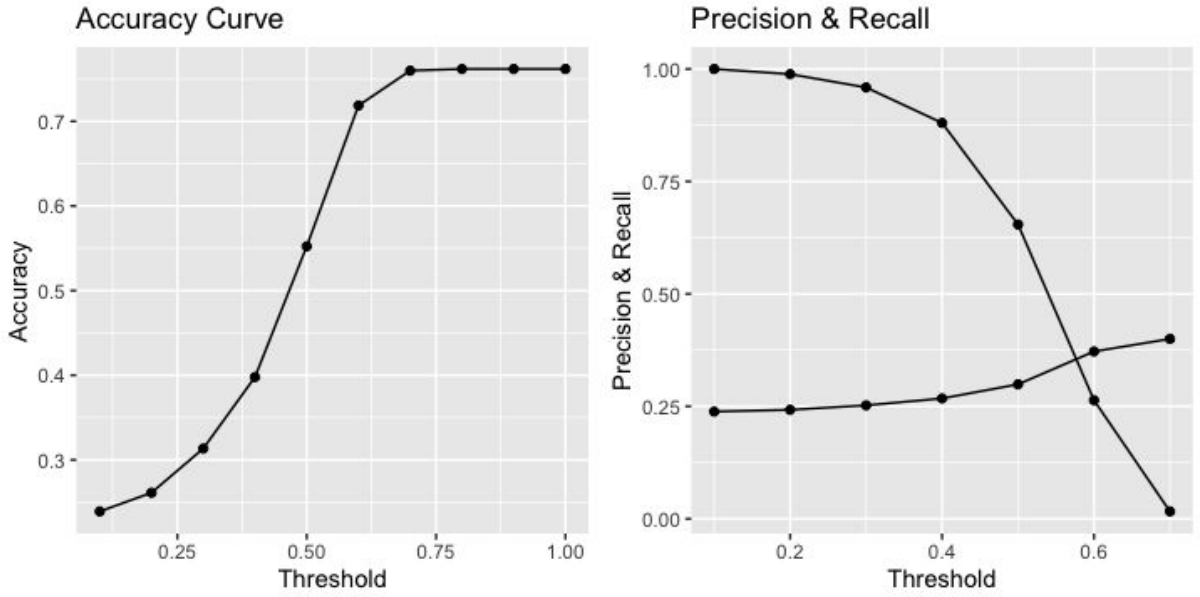
LR Unbal: High accuracy but fails to detect any positive instances, resulting in zero recall and precision.

LR Bal: Better in terms of recall and precision.

LR Weight: Similar to previous, showing that class weighting helps improve the model's ability to handle imbalance data.

	Err. Rate	Acc	Recall	Prec.	FN Rate
LR Unbal.	0.238	0.761	0	0	0
LR Bal.	0.281	0.718	0.236	0.371	0.628
LR Weight	0.272	0.727	0.211	0.373	0.626

LOGISTIC REGRESSION: GRAPH



Balanced Simple Logistic Regression

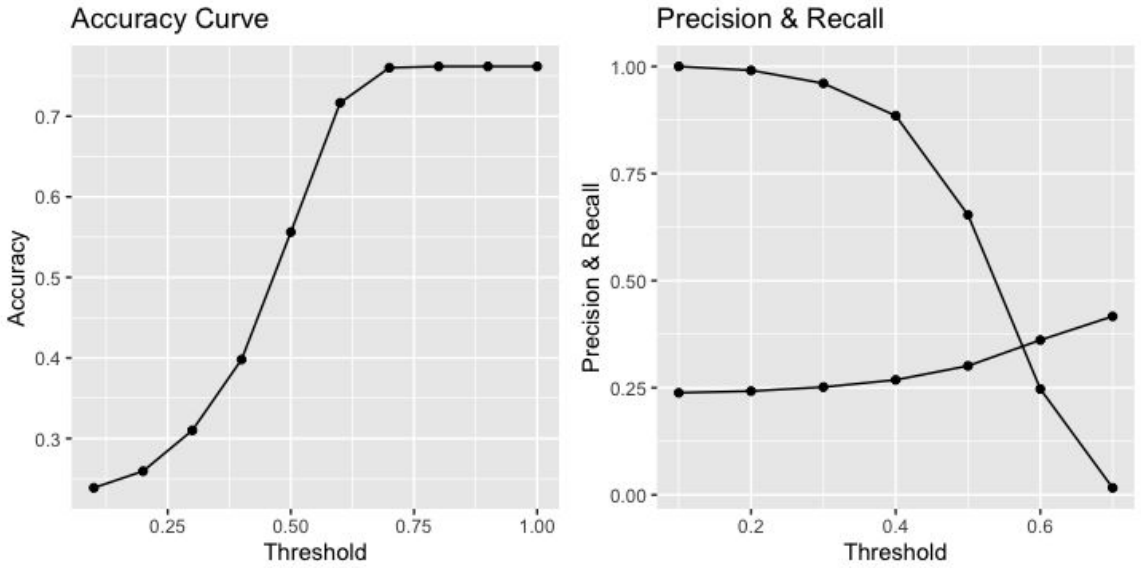
STEPWISE LOGISTIC REGRESSION

Stepwise Unbal: High accuracy but fails to detect any positive instances, resulting in zero recall and precision.

Stepwise Bal: Small improvement in recall and precision. However, there is still room for improvement in handling imbalance classes.

	Err. Rate	Acc	Recall	Prec.	FN Rate
Stepwise Unbal.	0.238	0.761	0	0	0
Stepwise Weights Bal.	0.282	0.717	0.281	0.376	0.623

STEPWISE LOGISTIC REGRESSION: GRAPHS



Stepwise Balanced and Weights

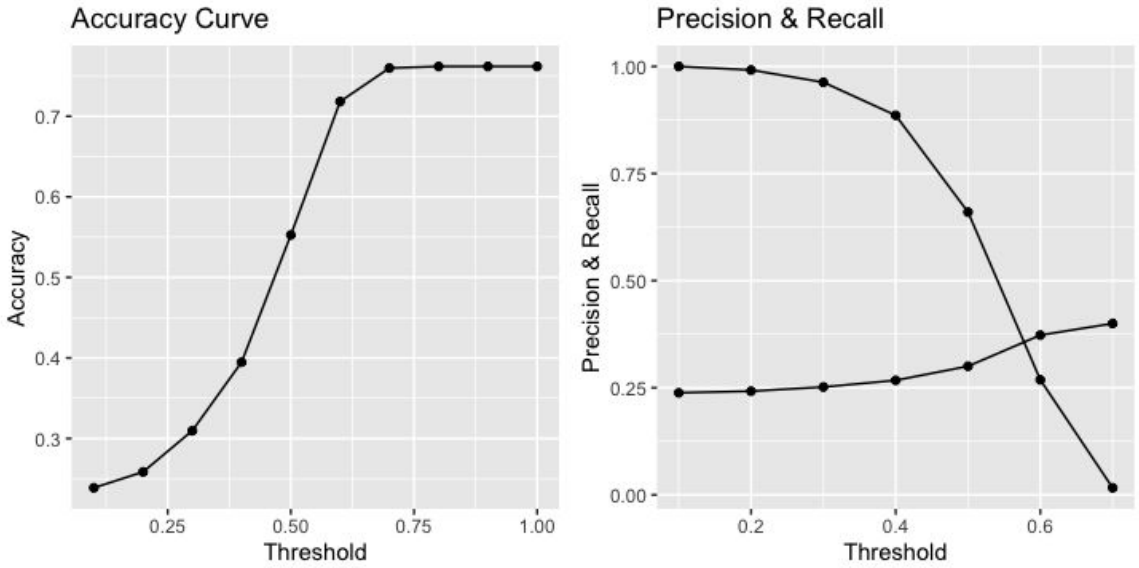
LINEAR DISCRIMINANT ANALYSIS

LDA Unbal: High accuracy but lacks any positive class detection. It appears to be biased towards the majority class.

LDA Bal: Small improvement in recall and precision. But precision is still low.

	Err. Rate	Acc	Recall	Prec.	FN Rate
LDA Unbal.	0.238	0.461	0	0	0
LDA Bal.	0.281	0.718	0.268	0.372	0.627

LINEAR DISCRIMINANT ANALYSIS: GRAPH



LDA Balanced

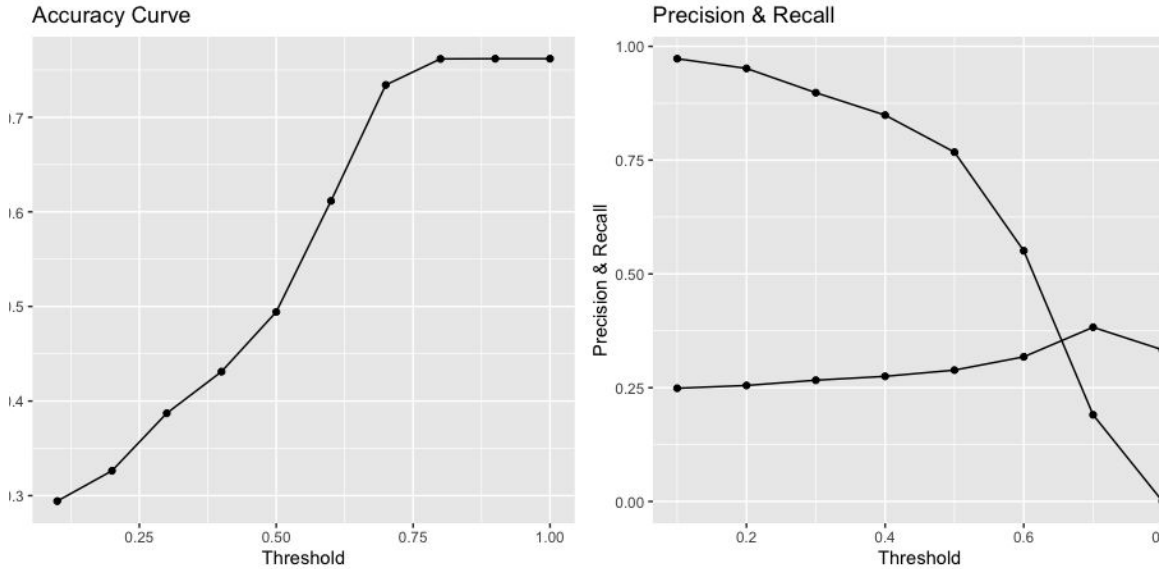
QUADRATIC DISCRIMINANT ANALYSIS

QDA Unbal: High accuracy but lacks any positive class detection. It appears to be biased towards the majority class.

QDA Bal: Small improvement in recall and precision. The accuracy is lower, suggesting a trade-off similar to other balanced models.

	Err. Rate	Acc	Recall	Prec.	FN Rate
QDA Unbal.	0.238	0.761	0	0	0
QDA Bal.	0.388	0.611	0.550	0.317	0.682

QUADRATIC DISCRIMINANT ANALYSIS: GRAPH



QDA Balanced

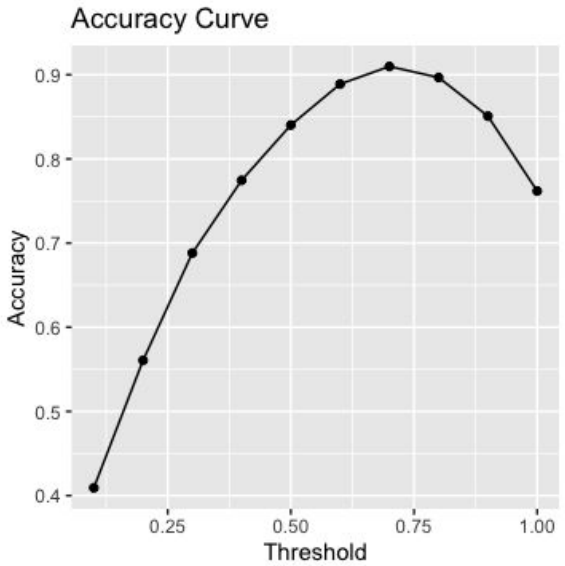
RANDOM FOREST

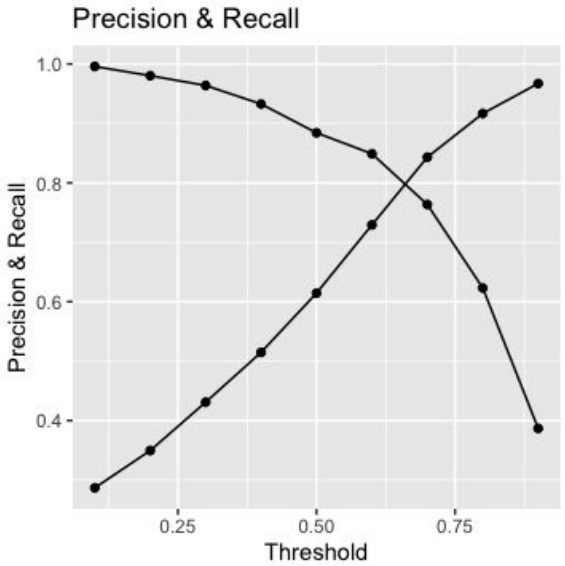
RF Unbal: Good accuracy, recall and precision. It could be a good trade-off for our goal.

RF Bal: Outperforms all other models in this comparison, achieving high accuracy and significantly improving both recall and precision for both classes, indicating its effectiveness in handling class imbalance.

	Err. Rate	Acc	Recall	Prec.	FN Rate
RF Unbal.	0.23	0.769	0.09	0.60	0.393
RF Bal.	0.11	0.88	0.84	0.73	0.27

RANDOM FOREST: RESULTS





Balanced Random Forest

IMPORTANCE

WHICH ARE THE MOST IMPORTANT FEATURES?

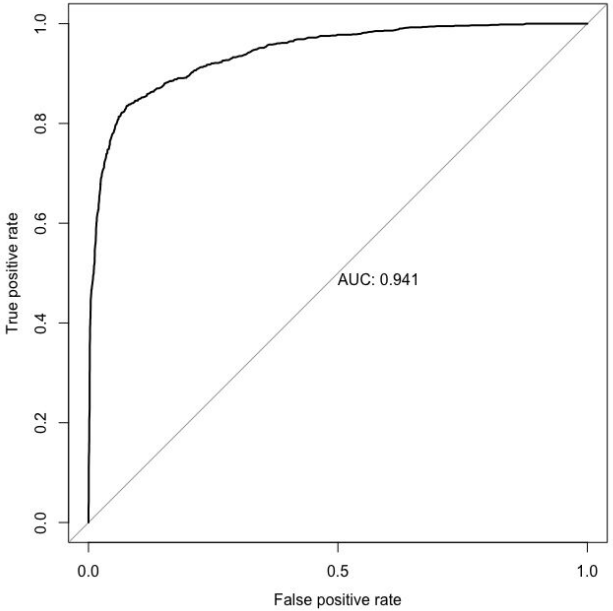
usd_goal_real	category	days_between	country	main_category	launch_year
1268,2214	729,5915	724,2666	538,5848	378,2932	216,1669

FINAL COMPARISON

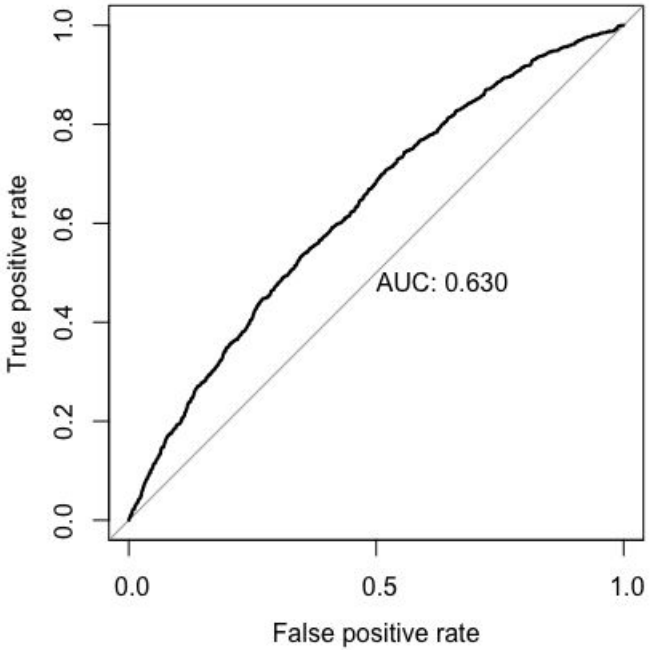
	Err. Rate	Acc	Recall	Precision	FN Rate
LR Unbal.	0.238	0.761	0	0	0
LR Bal.	0.281	0.718	0.236	0.371	0.628
LR Weight	0.272	0.727	0.211	0.373	0.626
Stepwise Unbal.	0.238	0.761	0	0	0
Stepwise Wight	0.282	0.717	0.281	0.376	0.623
LDA Unbal.	0.238	0.461	0	0	0
LDA Bal.	0.281	0.718	0.268	0.372	0.627
QDA Unbal.	0.238	0.761	0	0	0
QDA Bal.	0.388	0.611	0.550	0.317	0.682
RF Unbal.	0.23	0.769	0.09	0.60	0.393
RF Bal.	**0.11**	**0.88**	**0.84**	**0.73**	**0.27**

Table 5.1: Results at **0.6** Threshold

FINAL COMPARISON

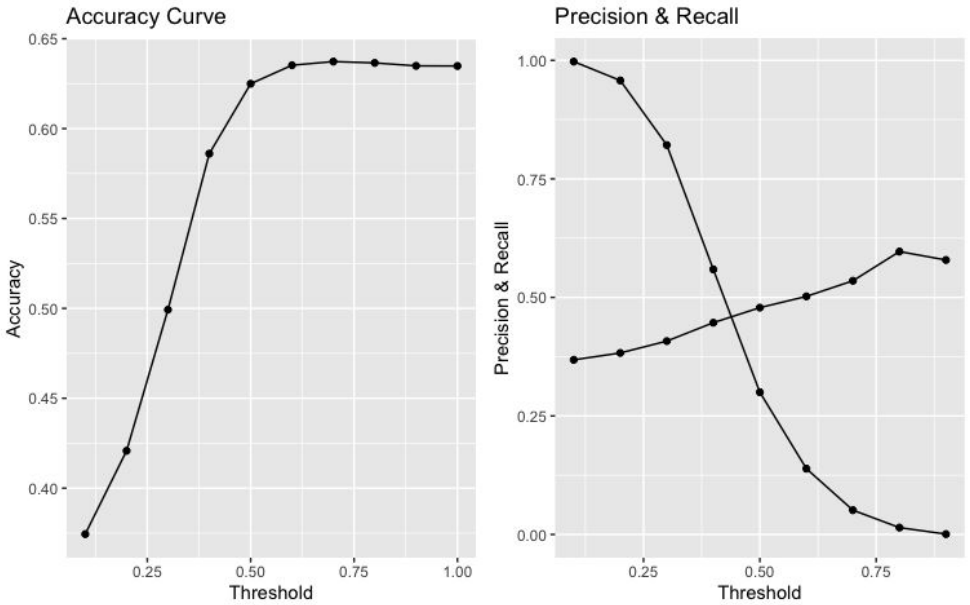


RANDOM FOREST

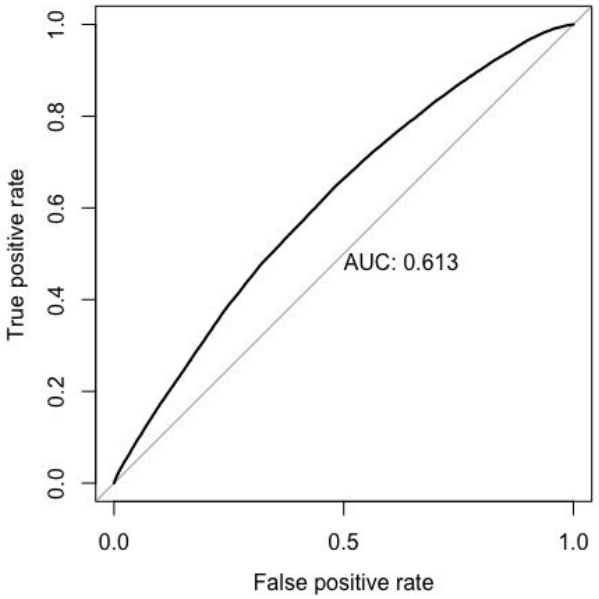


QDA Balanced

USA TEST



USA RESULT RF



USA ROC Curve

THANK YOU