PREDICTING STARTUP SUCCESS

Maryam Ghaseri
August 2020

INTRODUCTION

- Use startup performance to predict success / failure
- Venture capital point of view

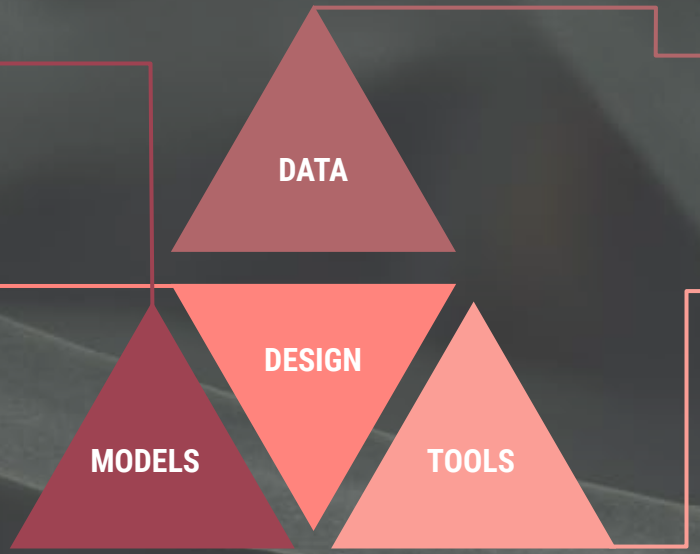
METHODOLOGY

MODELS

Classification

DESIGN

Success / Failure



DATA

CrunchBase Pro

TOOLS

Python, Sklearn, Pandas, Seaborn, Matplotlib, Pickle, SQL, Flask

DATA AND DESIGN

DATA

CRUNCHBASE PRO DATA

- Snapshot of 2015
- Founded in the last 10 to 15 years
- Five relational databases
- 60k+ companies

TARGET COMPANIES

- **At least one round of funding**

DATA CLEANING

- Removed outliers
- Valid / non-missing date founded
- Valid / non-missing date of first funding

DESIGN
(VC POINT OF VIEW)

SUCCESS

- Acquired
- IPO

FAILURE

- Closed
- Operating - assumptions:
  - Founded in the past five years
  - No funding in the past three years *

* The average months between fundings is 12 to 18 for tech companies.

FEATURES AND MODELING

FEATURES / FEATURE ENGINEERING

FUNDING

- Funding amount ($US)
- Months until first funding
- Num. funding rounds

LOCATION

- Country
- U.S States

INDUSTRY

Examples:

- Software
- Healthcare
- Biotechnology
- E-commerce

MODELS AND METRIC

BEST MODEL

- Logistic Regression *
- Grid search cross validation

METRIC

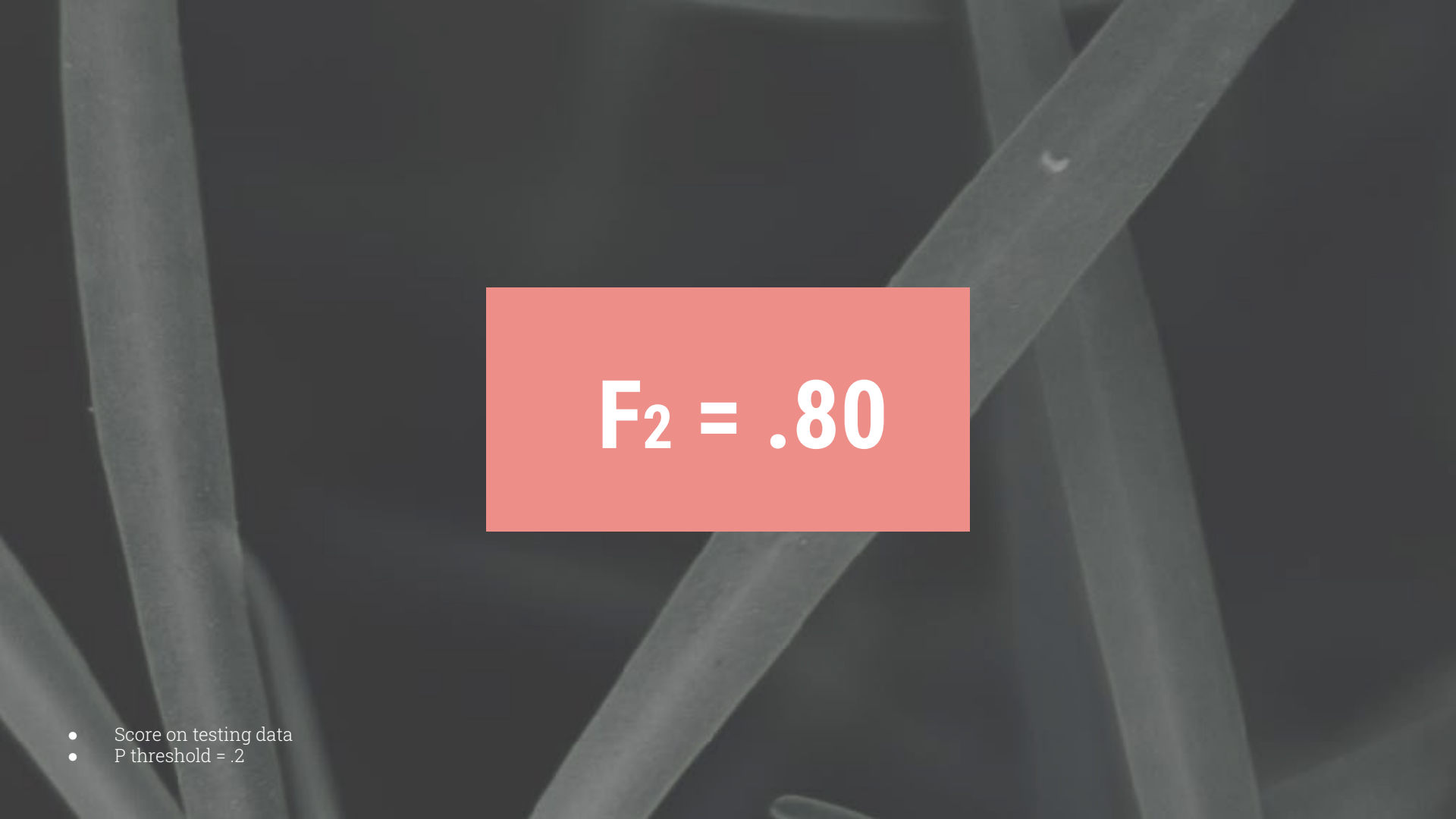
- F_2 - more weight on **recall**

* See the appendix for performance of the other models tested.

RESULTS

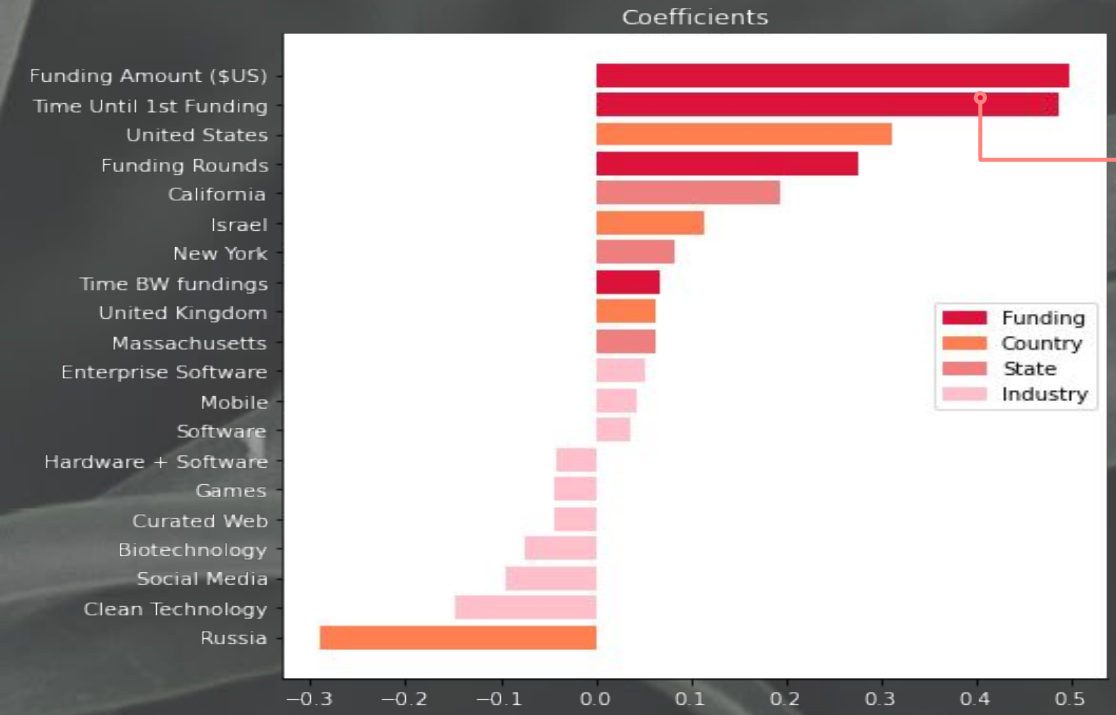
SUCCESS PROBABILITY

37%



- Score on testing data
- P threshold = .2

FEATURE IMPORTANCE

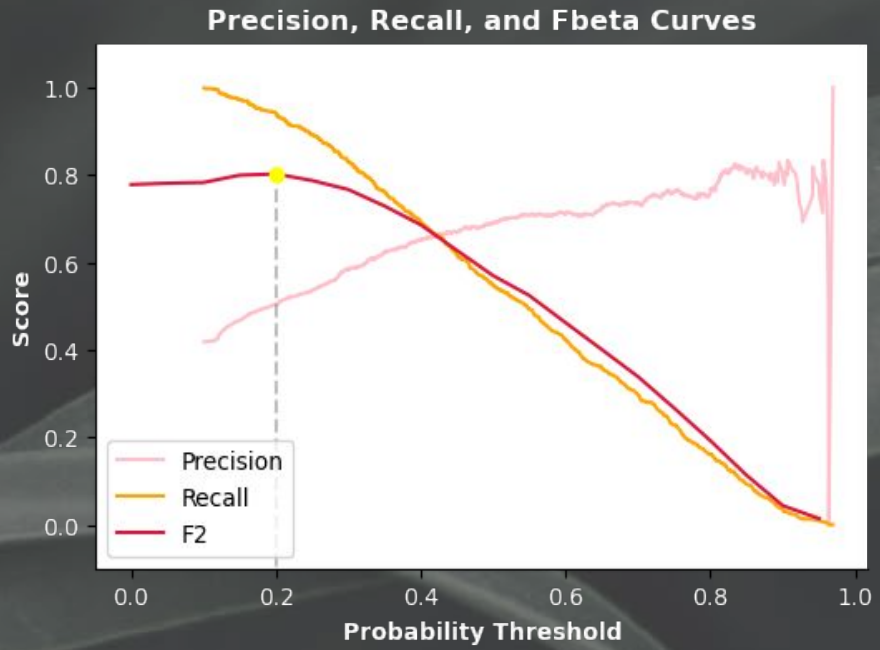


FUNDING MATTERS!!

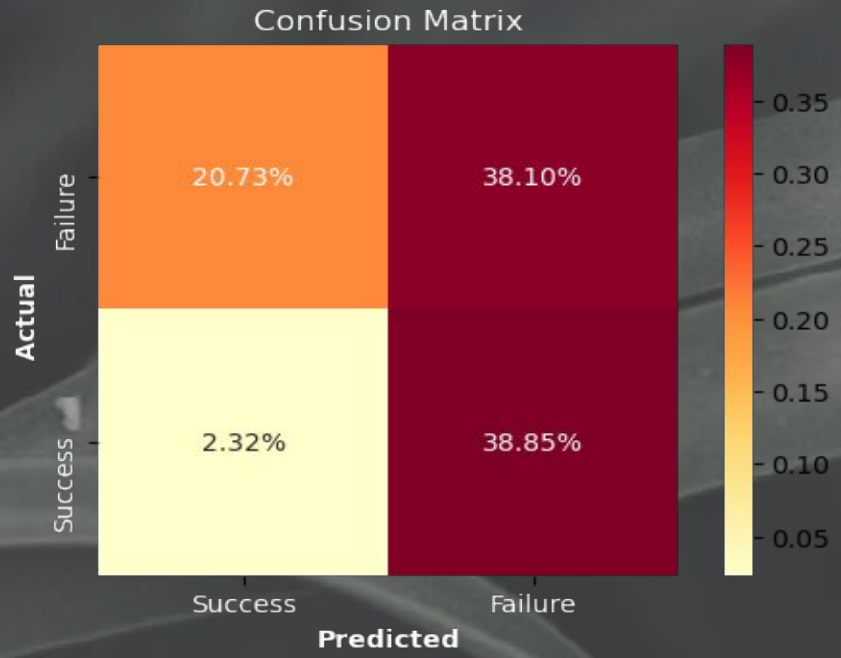
OTHER IMPORTANT FEATURES:

- United States
- California
- Software

LOGISTIC REGRESSION



LOGISTIC REGRESSION

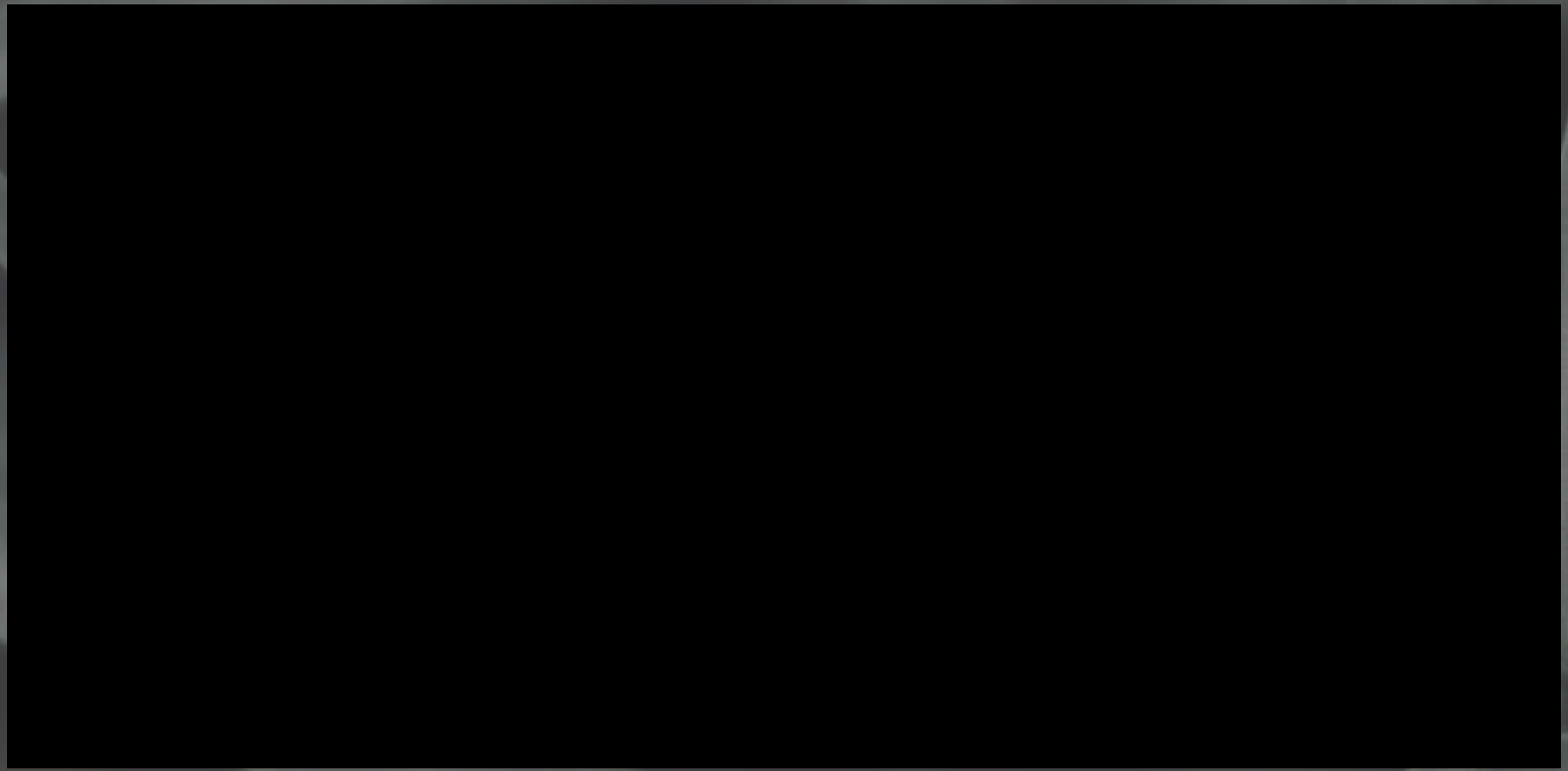


SAMPLE SUCCESSFUL PREDICTIONS

	FOUNDED AT	1st FUNDING DATE	FUNDING AMOUNT	FUNDING ROUNDS	STATUS	MODEL PREDICTION
AMAZON	7/5/1994	7/1/1995	$8M	1	IPO	**Successful**

IMPLEMENTATION

FLASK IMPLEMENTATION

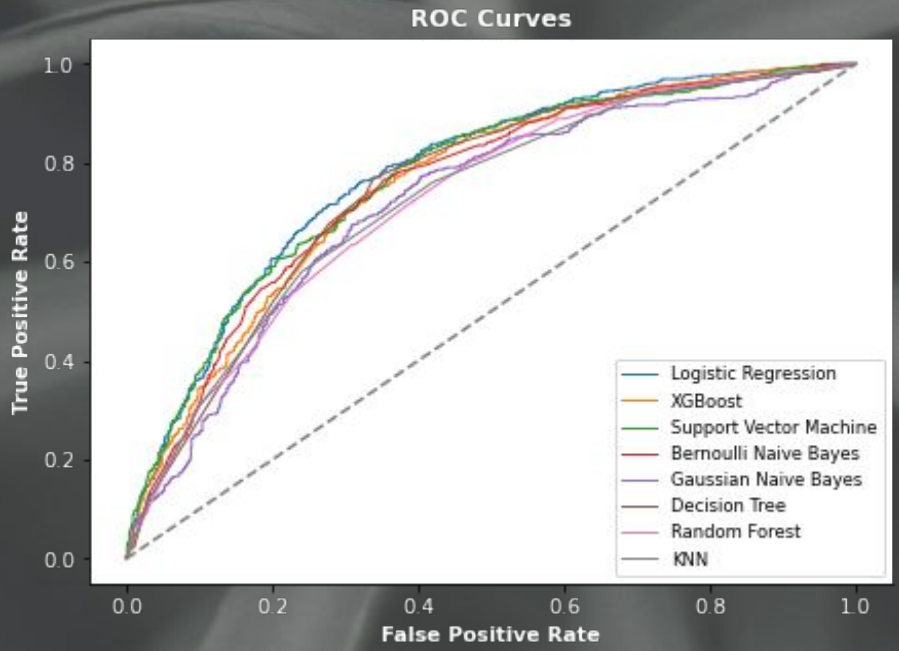


THANKS!

Questions?

APPENDIX

MODEL COMPARISON



MODEL COMPARISON

	F_2	ROC AUC
Logistic Regression - Tuned	**0.8040**	**0.7795**
Logistic Regression	0.8020	0.7790
XGBoost	0.7953	0.6596
Bernoulli Naive Bayes	0.7935	0.7542
Random Forest	0.7915	0.5695
Decision Tree	0.7912	0.5277
Support Vector Machine	0.7868	0.7684
K Nearest Neighbors	0.7867	0.7264
Gaussian Naive Bayes	0.7777	0.7185

SAMPLE DECISION TREE