Crunchbase Funding Prediction Model

By Daiki Minaki

Objective

Predict whether or not a company will get funded in the following year based on data at a given time.

Problem

There Are A Lot Of Startups Out There.

Can We Predict The Ones Worth Funding?

About The Data

Following Datasets Provided by Crunchbase (Up to 2015):

	Shape	Numeric Data	Text Data	Datetimes
Company Data	(51146, 14)	2	9	3
Investment Rounds Data	(212810, 18)	1	16	1
Organization Data	(606064, 16)	0	16	0
People Data	(605630, 15)	0	15	0
Acquisition Data	(18968, 18)	1	15	2
IPO Data	(1259, 13)	2	8	3

Approach

1. Data Cleaning & Wrangling
2. Exploratory Data Analysis
3. Random Forest Classifier
4. Model Evaluation
5. Analyze Results

Data Cleaning & Wrangling

Goal For Final Dataset:

- **Features and Target Are All Numeric Values**
- **Each Row Represents One Fiscal Quarter Of A Company**
- **Features Summarize All Important Data From Original Datasets**
- **Target Shows Whether Company Is Funded In Coming Year**

Processed Data

Final Dataset

- About 1,600,000 Rows
- 111 Numeric Features
- 2 Text Features (HashingVectorizer)

Features Categories

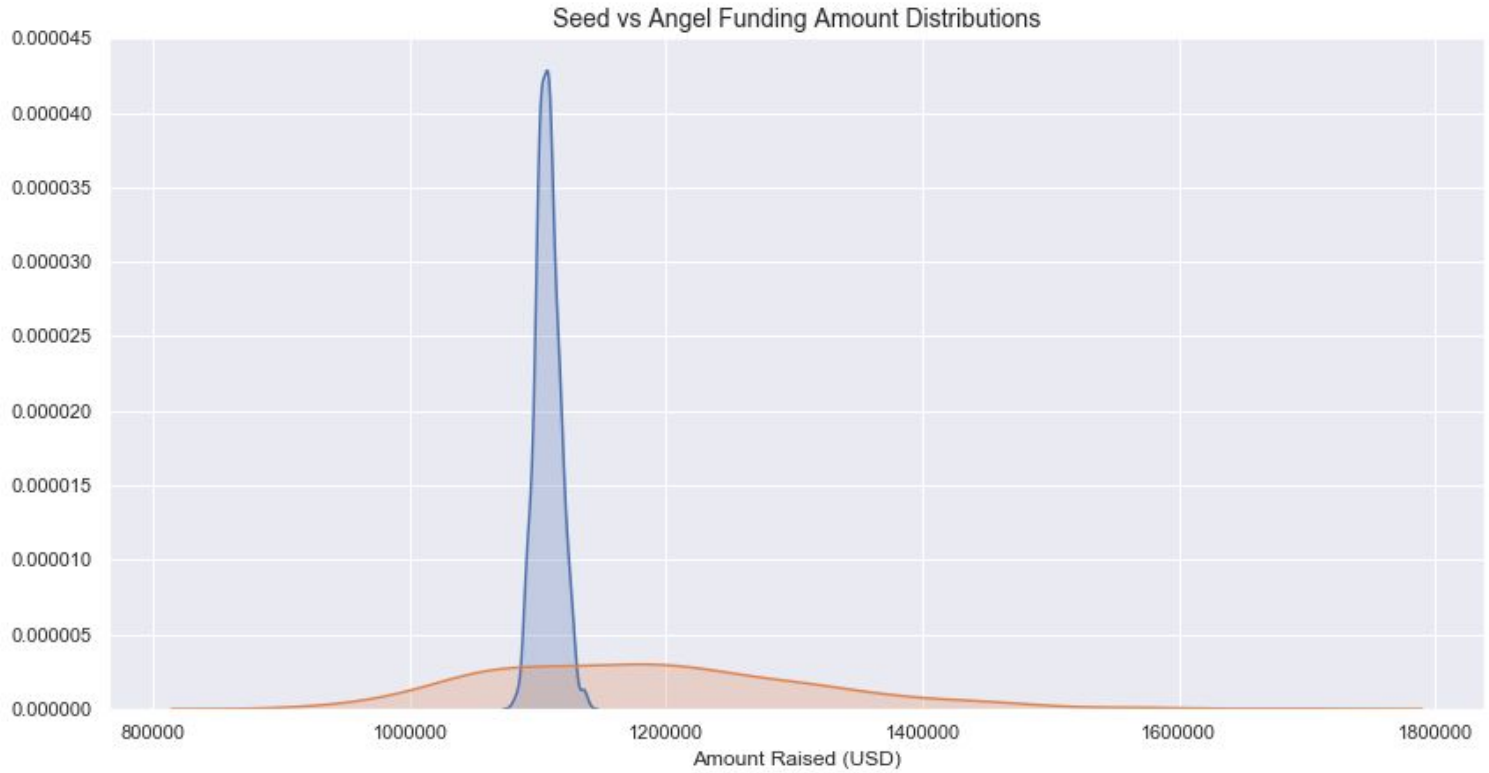
- Company Features
- Investment Round Features
- Investor Features
- Macro Features

Exploratory Data Analysis (EDA)

1. Investment Analysis
2. Investor Analysis
3. Category Analysis

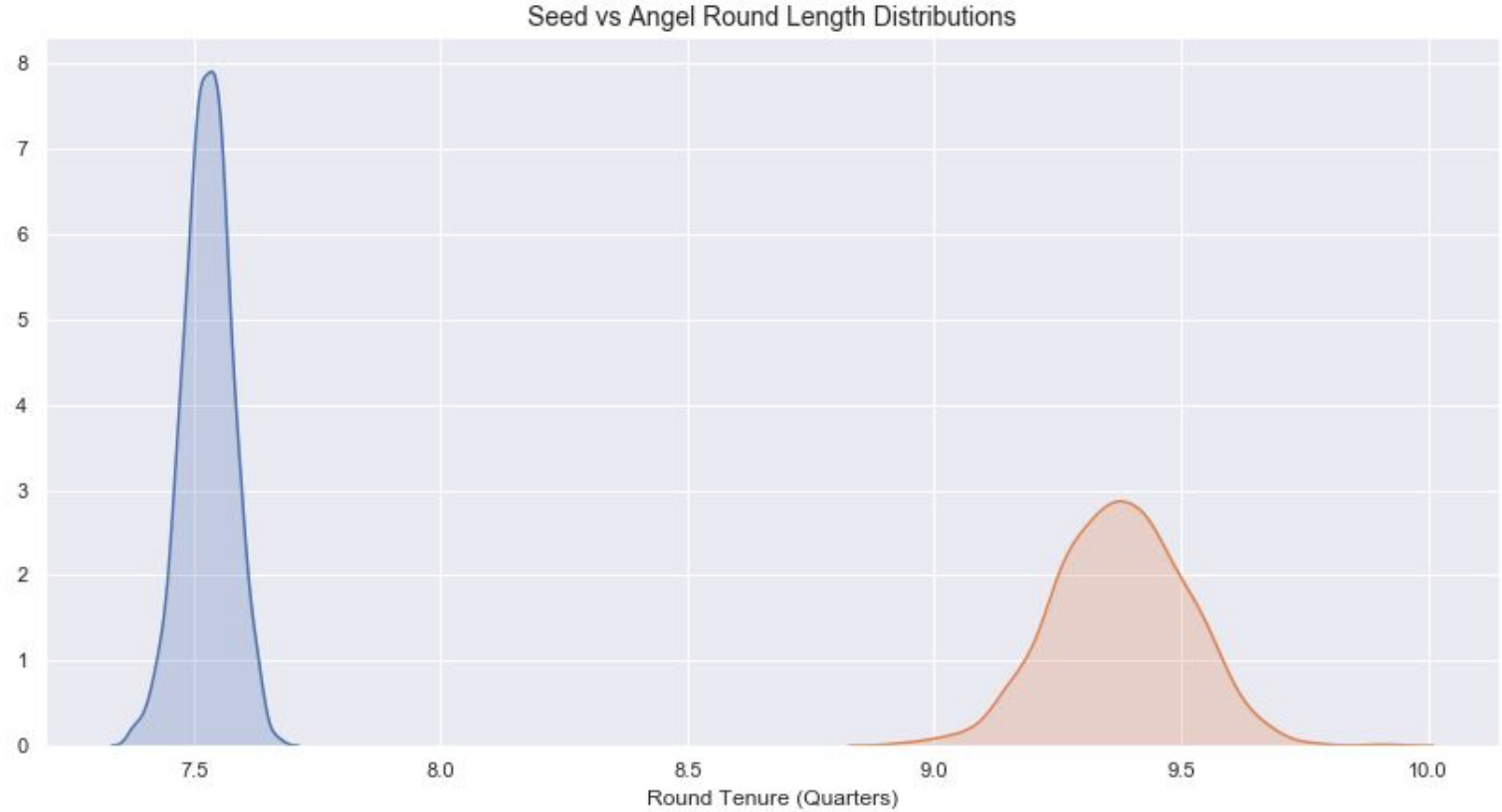
EDA: Investment Analysis

Analyzing Round Amounts (Other Rounds in Notebook)



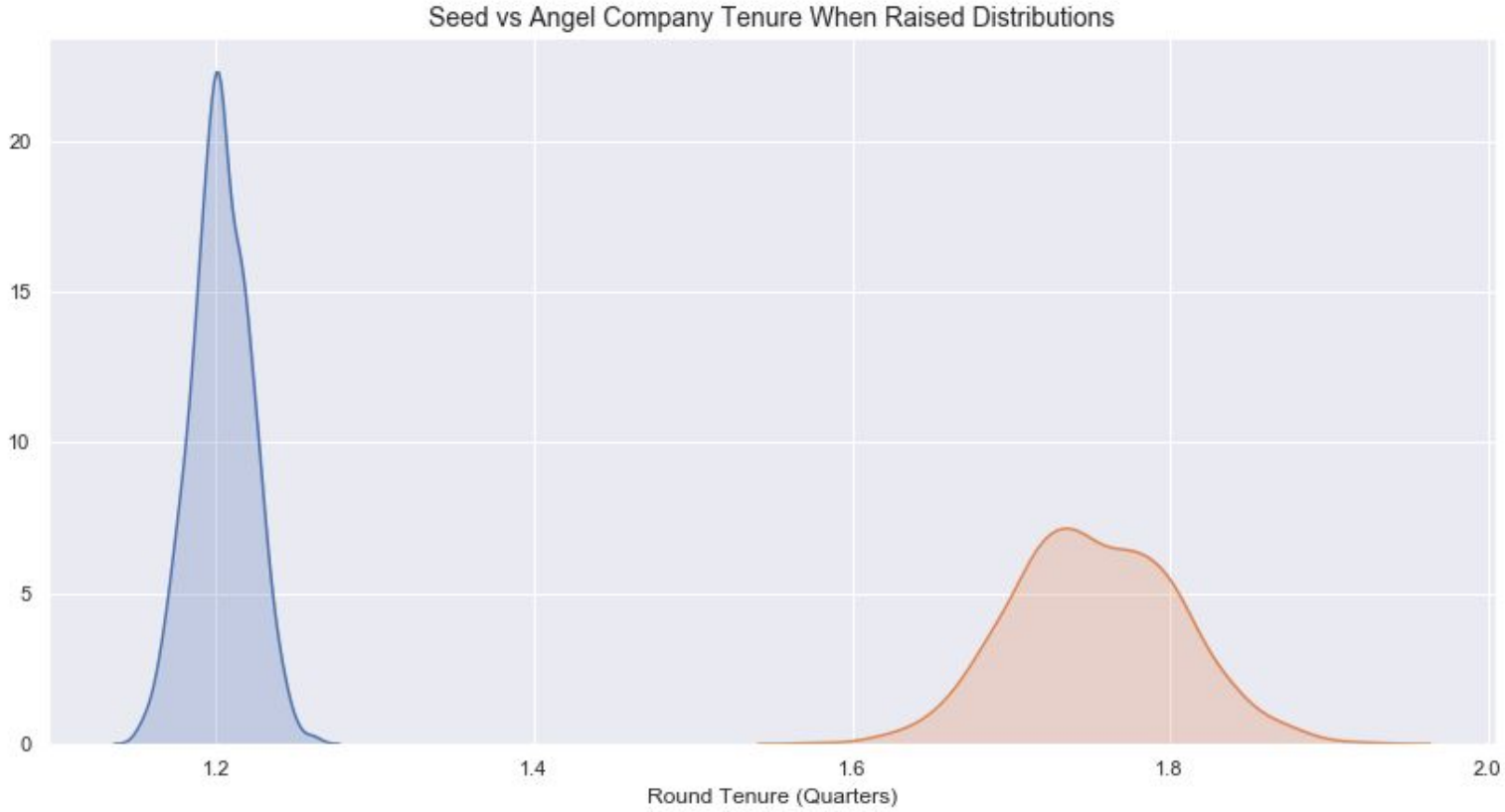
EDA: Investment Analysis

Analyzing Round Lengths (Other Rounds In Notebook)



EDA: Investment Analysis

Analyzing Company Tenure At Funding (Other Rounds In Notebook)



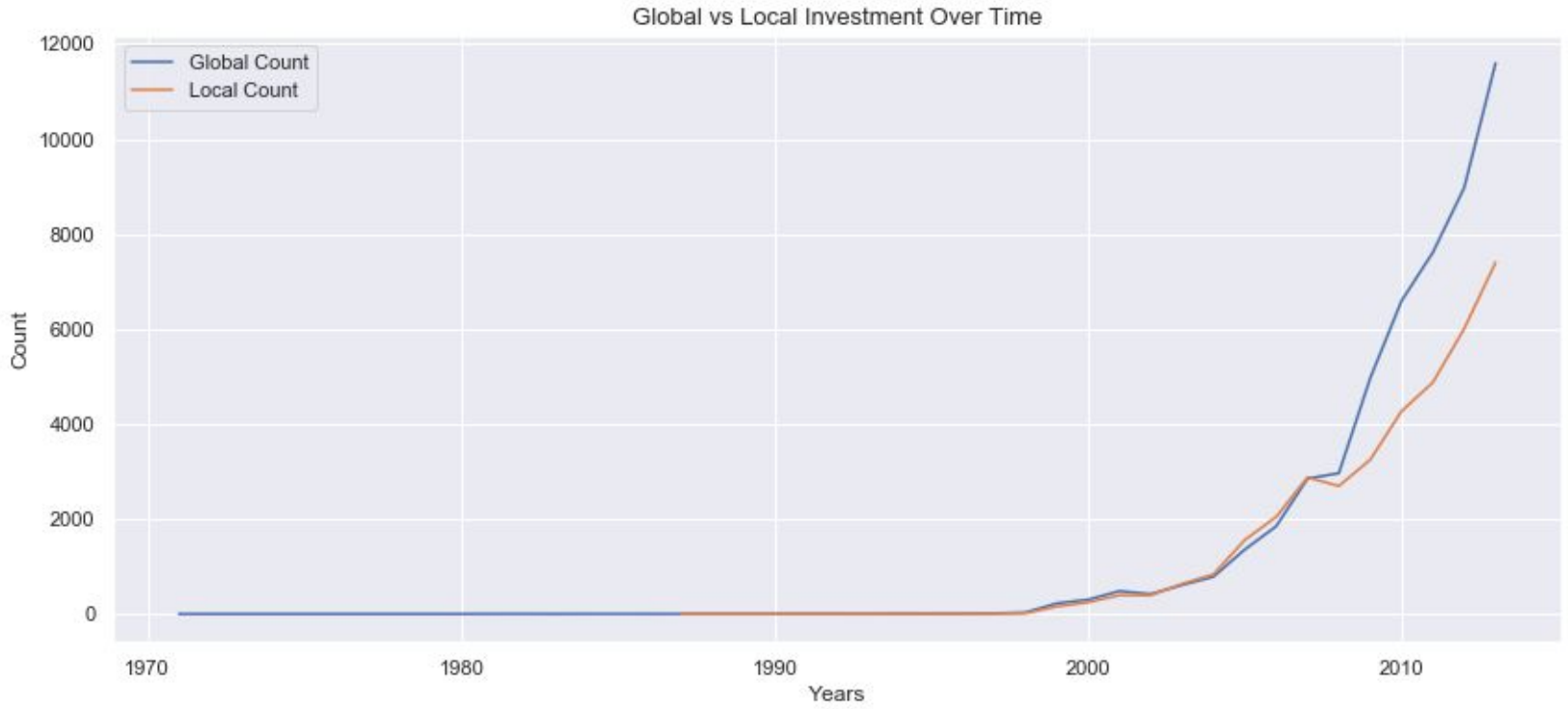
EDA: Investment Analysis

Key Observations (Full Analysis In Notebook)

- **Funding Amount by Angels Have A Higher Variance Than Seeds.**
- **Later Rounds** Typically Have **Higher Variance** in Amount & Tenure.
- **Round Length For Venture Rounds** Tends to be **between 10 - 15 Quarters.**
- **Equity Funding is Greater Than Debt Before IPOs but Less After.**

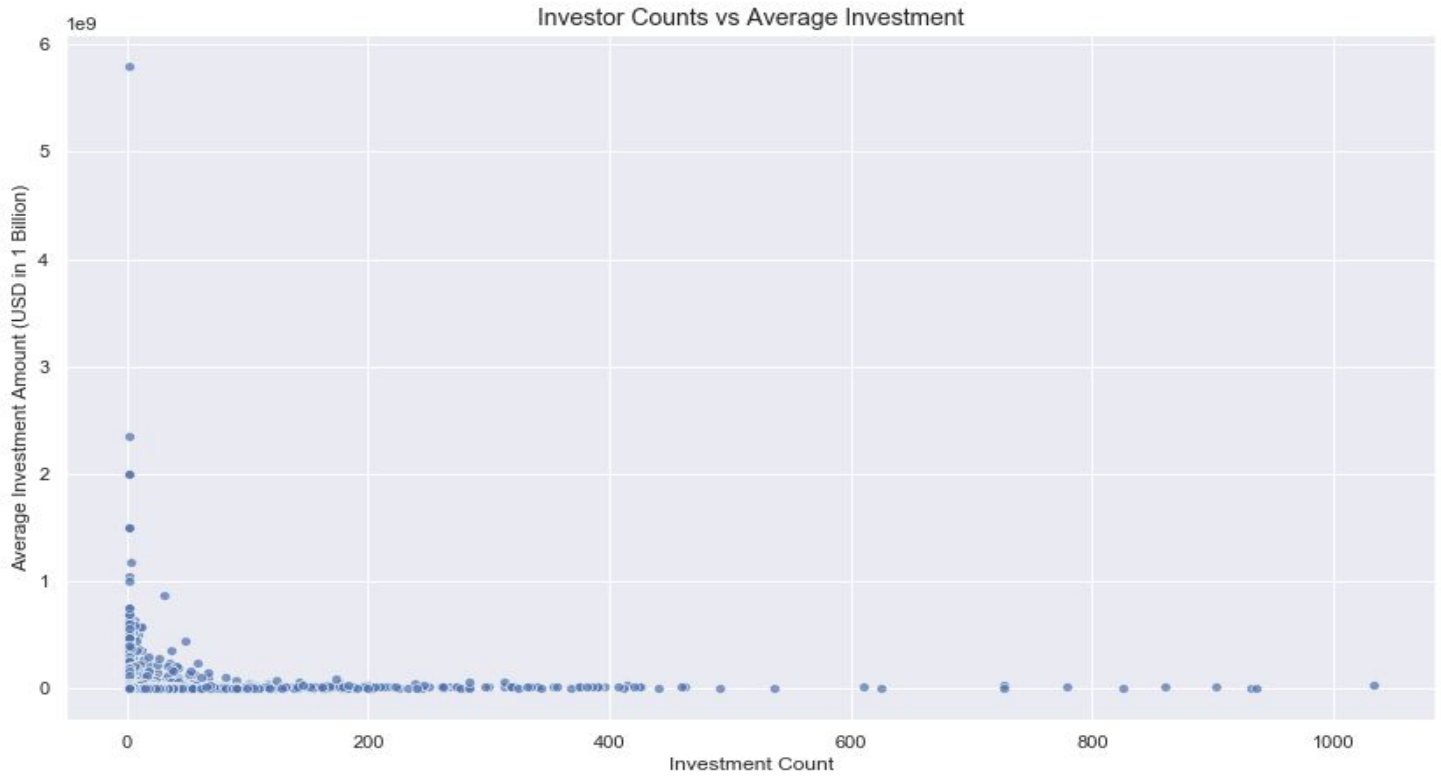
EDA: Investor Analysis

Analyzing Globalization of Investing (Full Analysis In Notebook)



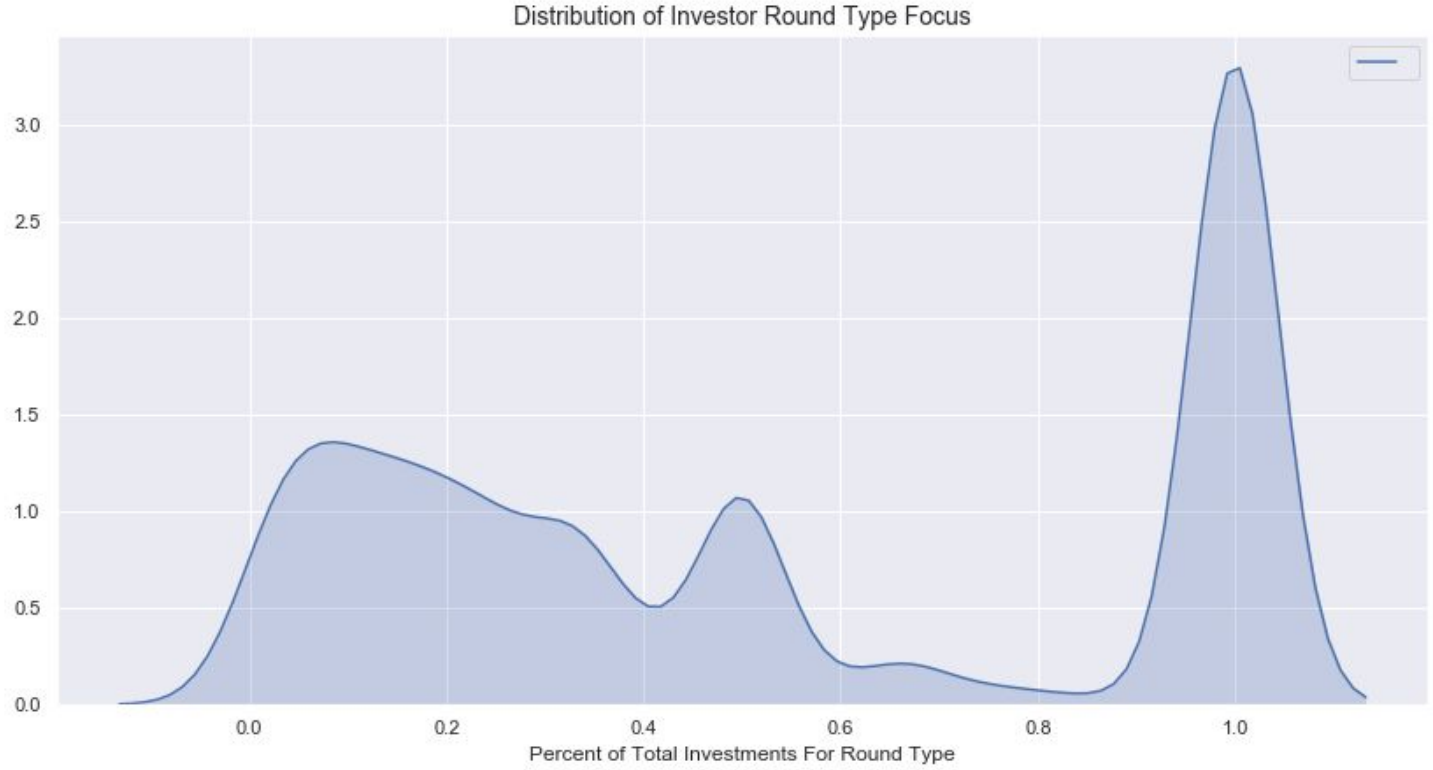
EDA: Investor Analysis

Analyzing Differences In Investor Strategy (Full Analysis In Notebook)



EDA: Investor Analysis

Do Investors Focus On Specific Round Types? (Full Analysis in Notebook)



EDA: Investor Analysis

Key Observations

- **There Are More Gaps in Data of Global Investors**
- **Difference in Invested Amount** of Global and Local Investors **is NOT statistically significant**
- **Investors Do NOT Focus** On Just **Many Small** or **Few Large Investments But Have Varying Strategies Overall.**
- **Most Investors Focus On More than One Round Type.**
- **Seed Investors** Tend to **Focus on Seed Funding**.

EDA: Category Analysis

Analyzing Funded Categories Over Time

Top Categories in the 1980s

Top Categories in the 2010s

EDA: Category Analysis

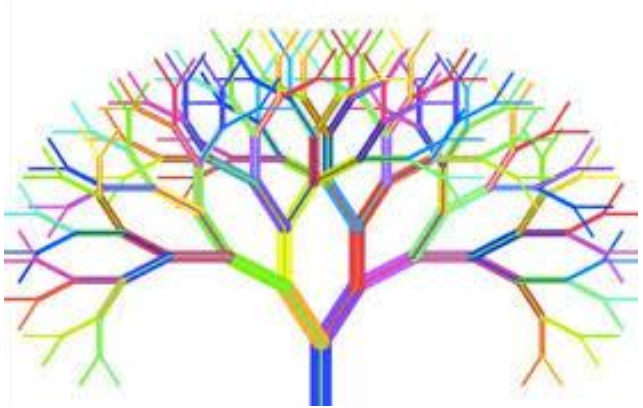
Analyzing Funded Categories Over Time (Full Analysis In Notebook)

- **1980s:** Manufacturing, Services, Designers, Automotive, Technology
- **1990s:** Software, Technology, Curated Web, Service, Internet
- **2000s:** Software, BioTech, Enterprise Software, Mobile, Curated Web
- **2010s:** Software, Enterprise Software, Mobile, Curated Web, Commerce

Choosing The Model

Why Random Forest Classifier?

- Flexible
- Prevents Overfitting
- High Feature & Sample Count
- No Scaling Required



Training The Model

Input Data After HashingVectorizer

- **111** Numeric Features
- **3000** HashingVectorizer Features

Pipeline

- **FeatureUnion**
  - **Imputer**
  - **HashingVectorizer**
  - **SelectKBest**
- **RandomForestClassifier**

Hyperparameter Tuning

***Default Parameters* Yielded Best Results!**

Model Evaluation

Accuracy: 0.94

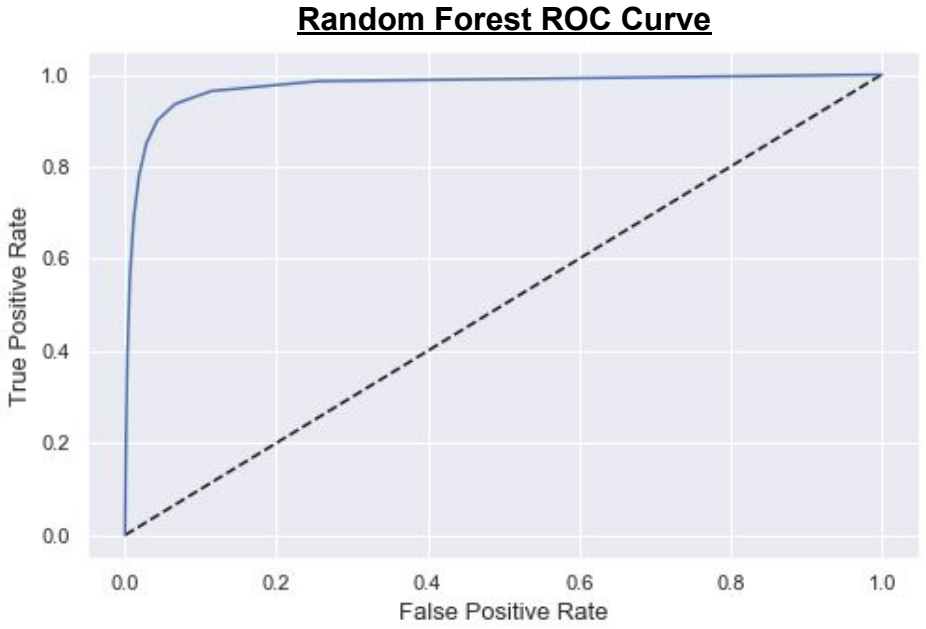
Avg CV Score: 0.89

Avg Precision: 0.94

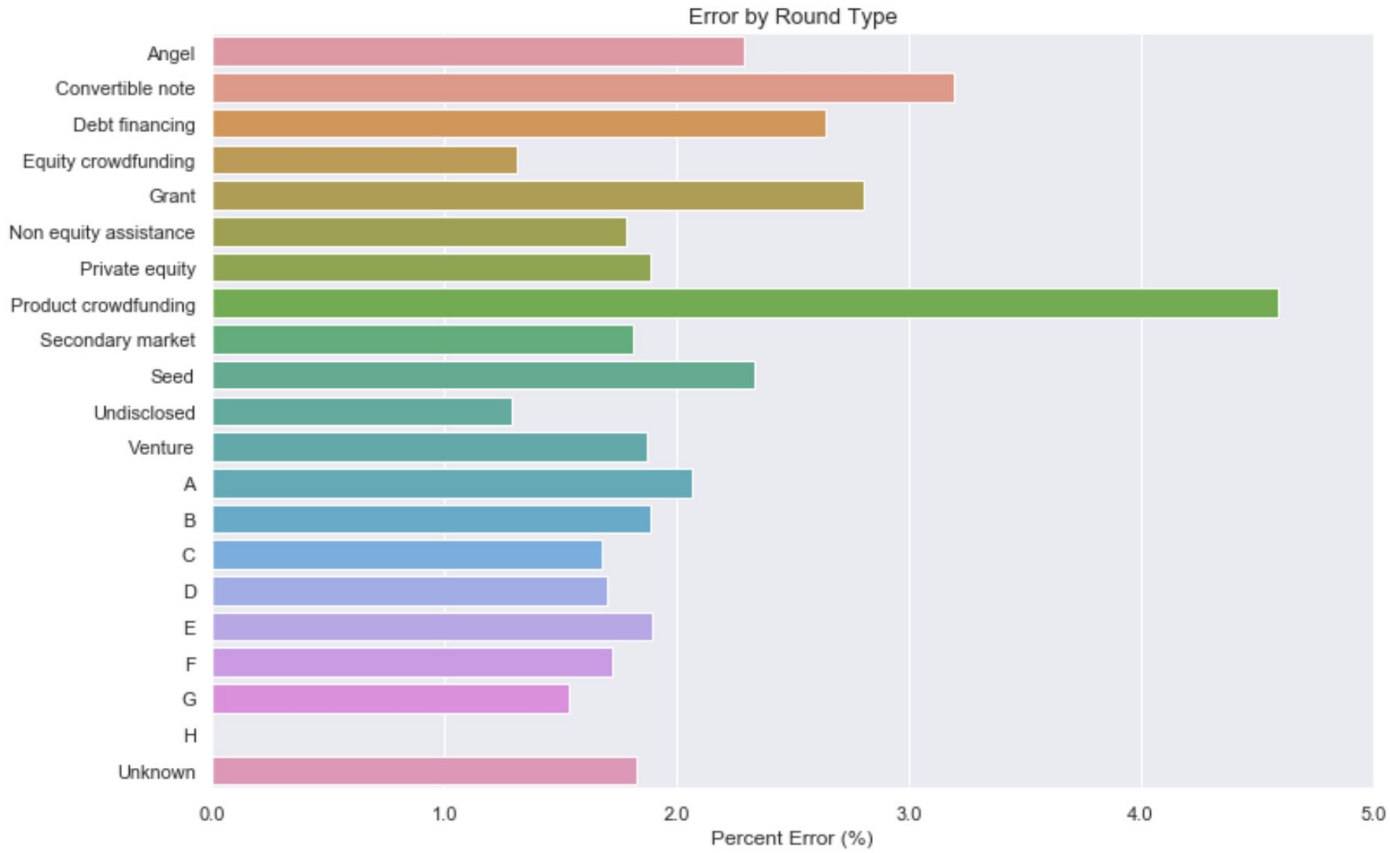
Avg Recall: 0.94

ROC Curve: *Figure on Left*

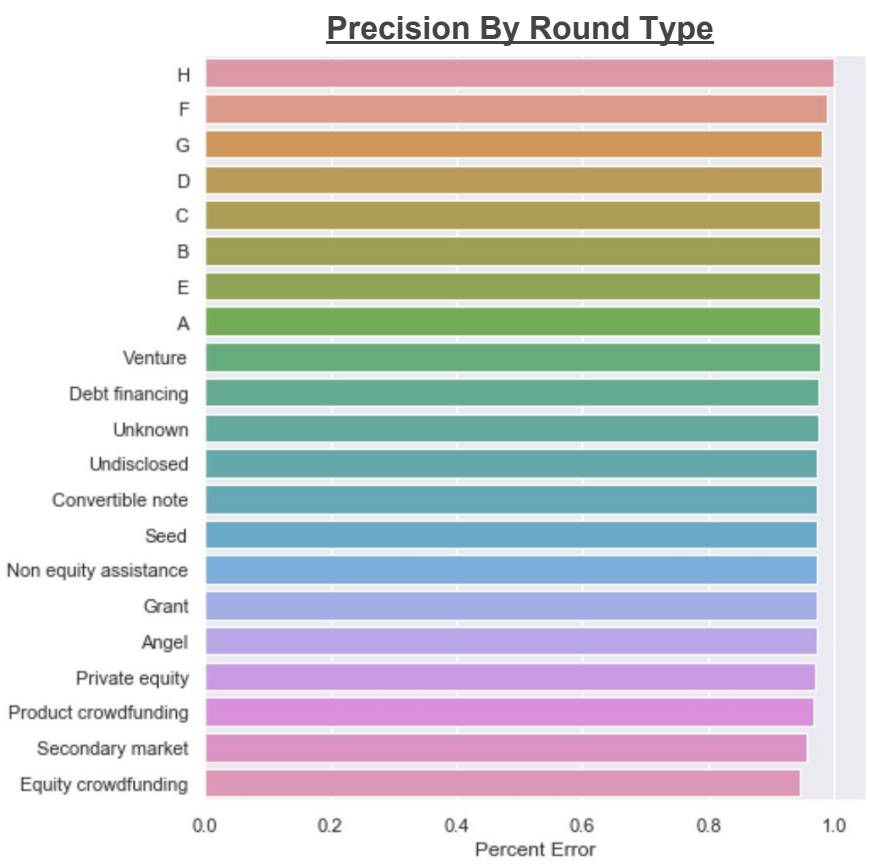
AUC Score: 0.975

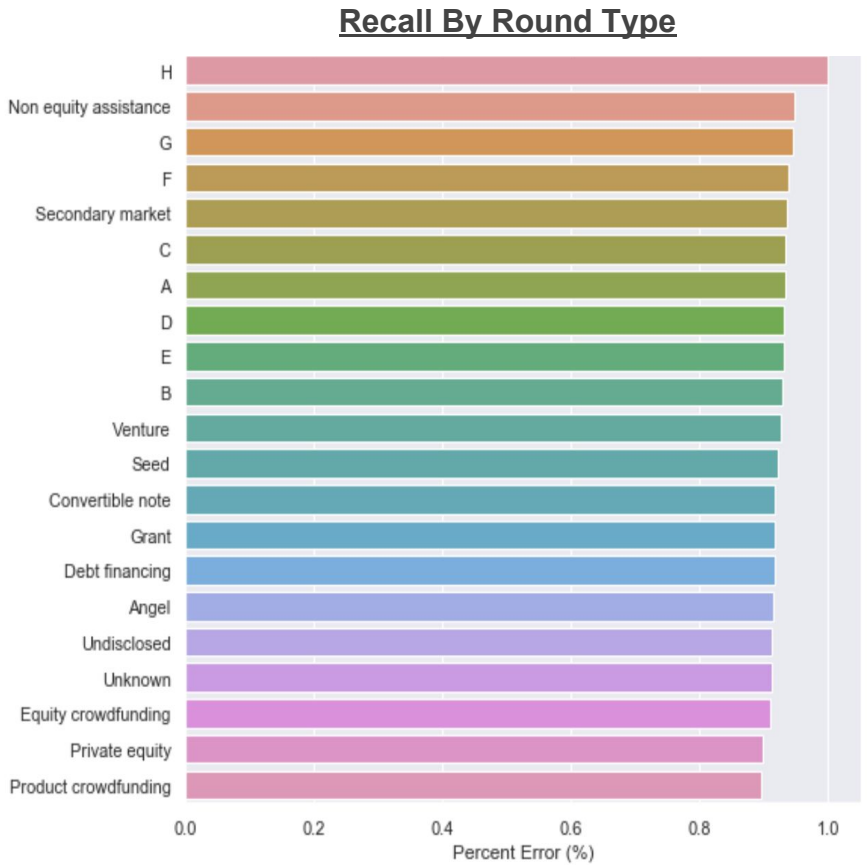


Error by Funding Type



Precision & Recall by Round Type





Analyzing Results

Model Had Some Problems With Recalling Funded Targets.

Precision: 91% of Predicted Funded Were Actually Funded.

Recall: 78% of Actual Funded Rounds Were Identified as Funded.

	precision	recall	f1-score	support
0.0	0.95	0.98	0.96	380907
1.0	0.91	0.78	0.84	93510
micro avg	0.94	0.94	0.94	474417
macro avg	0.93	0.88	0.90	474417
weighted avg	0.94	0.94	0.94	474417

Precision Or Recall?

In this case, Precision IS *more important* than Recall.

We would rather fund a company and be correct than fund all the correct companies.

(If not fund then at least find)

Outcome

We Were Able To Build a Random Forest Classifier That Predicts Company Funding At *About 94% Accuracy*.

References and Resources

GitHub Repository
https://github.com/daikiminaki/Springboard/tree/master/Capstone_Project_Crunchbase_Funding

Detailed Report

Email
dminaki95@gmail.com/