DAVID GUIX - MACHINE LEARNING PROJECT

STARTUP FUNDING

TABLE OF CONTENTS

OVERVIEW

AUTHOR

David Guix Sánchez

GITHUB REP:

https://github.com/DavidGuix/Startup-Funding-Machine-Learning/

DATASETS:

investments_VC.csv -> From Crunchbase database with companies up to 2015. Source: Kaggle

dragons_den_dataset.xlsx -> Scraped from the internet with companies from the British TV show 'Dragon's Den' since 2005. Source: Wikipedia

Shark Tank Companies.csv -> Dataset with companies from the US TV show 'Shark Tank' first 4 seasons. Source: Data World

Shark Tank Companies S05_to_S11.xlsx-> Dataset with companies from the US TV show 'Shark Tank' from 5th to 11th season. Source: Kaggle

This project analyzes the needs of Next Level, a startup accelerator that is seeking to leverage a machine learning model that will precisely classify companies that will be funded by Venture Capital (VC). In an effort to model this problem, I have collected a dataset of startups around the world, founded from 1980. In addition, I have created a secondary dataset for predictions, based on data from the TV shows 'Dragons Den' and 'Shark Tank'.

BUSINESS PROBLEM

Next Level is a small fictional startup accelerator that is low on capital, currently they have 100k USD to invest in interesting startups. Because of their limited funds, they are looking for a way to better filter companies in the hopes of making the most of their investments. They are aware of the following statistics: 75% of venture-backed startups fail. Under 50% of businesses make it to their fifth year. 33% of startups make it to the 10-year mark. Only 40% of startups actually turn a profit. Given this knowledge, Next Level is targeting startups that they believe have the best opportunity at receiving funding from VC, a sure-fire way for investment profits. For this purpose, they have hired me to create a model classify whether or not a startup will be funded by a VC.

DATA DESCRIPTION

In order to help Next Level, I have used a Kaggle dataset with information on 54,000 companies sourced from Crunchbase to train our model.

In addition, I have created a secondary dataset with info from two TV shows where startups present their projects to a group of real investors.

DATA ISSUES

Because certain values possessed overly predictive power, they were dropped from the models. Those columns are as follows: round_A, round_B, round_C, round_D, round_E, round_F, round_H

The sum of all of these provides our target, the ‘venture’ column.

NON NUMERICAL VARIABLES

The original dataset has several columns categorical and date columns. In order to provide the models with more relevant data to predict, I have created synthetic columns:

- ‘from_founding_to_last_funding' -> number of day from founding date to last funding

- ‘from_founding_to_funding' -> number of days from founding to first funding

- ‘from_first_to_last_funding' -> number of days from first funding to las t funding

- ‘math_expectation_country’ -> math expectation of success per country

- ‘math_expectation_market' -> math expectation of success per market type

- ‘status' -> transformed from categorical ‘closed’, ‘open’, or ‘sold’ to 0, 1 or 2

- ‘attractiveness’ -> divided the venture column in 3 classes: low (0-50k), average (50k-2 m) and high (2m onwards)

SHARKS AND DRAGONS DATA

I have obtained this data from different sources, and manually standarized info such us the category (market).

In addition a used get_dummies on the category column and a ‘multiple_entrepreneur’ column with a binary data 0 and 1

APPROACH

EDA

I have done a data cleaning process on both sets of data. Please see my repository on EDA for further details.

https://github.com/DavidGuix/Startup-Funding-EDA/

PROJECT ASSUMPTIONS

Based on my knowledge on the topic and the goal of the project (predict the future success of getting VC funding of startups), I focus on predicting the ‘venture’ column as my target.

SHARKS AND DRAGONS

I believe the addition of this data is relevant to this project. At the end of the day these 2 TV shows represent exactly the problem I would like to solve: will a startup get a deal when they face a group of investors???

MODELS

I have faced a difficult task trying to predict the success of a startup, hence I have used 3 different attacking paths to tackle this issue.

REGRESSION

I have tried several regression algorithms to predict the final funding amount by Venture Capital. Linear Regression, SVR, RandomForestReg and a Neural Network.

Random Forest presented the best metrics, however the Neural Network showed much better predictions through generalization.

CLASSIFICATION

CLASSIFYING ATTRACTIVENESS

I have tried several regression algorithms to predict the class of attractiveness, defined by Venture Capital funding. Ensemble models, SVM, Decisión Tress, RandomForest and a Neural Network.

XGBoost presented the best metrics, however the Neural Network showed much better predictions through generalization.

CLASSIFYING DEAL OR NO DEAL

I have tried several regression algorithms to predict the class of attractiveness, defined by Venture Capital funding. Ensemble models, SVM, Decisión Tress, RandomForest and a Neural Network.

XGBoost presented the best metrics, however the Neural Network showed much better predictions through generalization.

ONE ALGORITHM (OR ALGORITHM OF POWER)

None of the previous results using a single type of algorithm was good enough to help deciding whether to invest money or not in a startup. This gave me the idea of combining the 3 results using a home made algorithm that would unite the results of all the previous algorithms. I like to call this the One Algorithm, or Algorithm of power in reference to the books of Tolkien.

The purpose of the algorithm is to unite them all and give as a result, a recommendation from 0 to 5 on whether is advisable to invest or not in the startup given.

In order to implement for test purposes the Sharks and Dragons algorithm, we will need to make up some numbers on how much the entrepreneur evaluates the startups and what % wants to give away in return for the VC investor. In the real world, we will get this info directly by the startup that wants to receive funding.

CONCLUSION

The conclusions based on a random sample of 20 startups are truly positive, as the combination of the 4 algorithm recommend with more than 4 stars, startups that do receive a huge level of investment in the future (high level of attractiveness.

Also, all the startups with 0 USD in venture funding got les than 1.5 stars in rating.

NEXT STEPS

Get more key data:

- Add history data of successful startups at their early stages.

- Consider adding data based on scalability capacity

- Consider not only the country of origin but also the countries they are operating or planning to

- Consider the business model and the team behind the startup as key factors to success

Predicting whether a company would get Venture Capital funding is a complex problem, and expanded data collection would greatly benefit the precision of the model.

For More Information:

See the full modeling process in the modeling notebook or review this presentation.

For additional info, contact me at davidguixsanchez83@gmail.com

Repository Structure

```
├──── README.md

├──── archive

|    ├──── ML Startups.ipynb

├──── data

|    └──── investments_VC.csv

|    └──── dragons_den_dataset.xlsx

|    └──── Shark Tank Companies.csv

|    └──── Shark Tank Companies S05_to_S11.xlsx

├──── documents
```

```
|    ├──── STARTUP FUNDING MACHINE LEARNING PDF.pdf

|    ├──── STARTUP FUNDING MACHINE LEARNING MEMORIA.pdf
```