upGrad

#LifeKoKaroLift

PG Diploma in DS

24-10-2020

upGrad

Lecture On : Investment Case Study

Instructor : Dr Reena Duggal

upGrad

Today's Agenda

1. Discussing the Premise of the Assignment
2. The objectives for the assignment
3. Pre-Modelling Steps
4. Analysis Steps and Checkpoints
5. Q & A

What is Spark Funds?

You work for Spark Funds, an asset management company. Spark Funds wants to make investments in a few companies. The CEO of Spark Funds wants to understand the global trends in investments so that she can take the investment decisions effectively.

Business objective: The objective is to identify the best sectors, countries, and a suitable investment type for making investments. The overall strategy is to invest where others are investing, implying that the 'best' sectors and countries are the ones 'where most investors are investing'.

Spark Funds has two minor constraints for investments:

1. It wants to invest between **5 to 15 million USD** per round of investment
2. It wants to invest only in **English-speaking countries** because of the ease of communication with the companies it would invest in.

Investment Case Study

upGrad

Investment type analysis

Comparing the typical investment amounts in the venture, seed, angel, private equity etc. so that Spark Funds can choose the type that is best suited for their strategy.

Country analysis

Identifying the countries which have been the most heavily invested in the past. These will be Spark Funds' favourites as well.

Sector analysis

Understanding the distribution of investments across the eight main sectors. (Note that we are interested in the **eight 'main sectors'** provided in the mapping file. The two files — companies and rounds2 — have numerous sub-sector names; hence, you will need to map each sub-sector to its main sector.)

Steps to proceed with the Case Study

There are four major parts that are needed to be done for this case study:

1. Data understanding/Exploration
2. Data cleaning (cleaning missing values, removing redundant columns etc.)
3. Data Analysis
4. Recommendations(Checkpoints)

Pre-Modelling Steps: Data Exploration

upGrad

Data Understanding/Exploration

df = pd.read_csv

1. Read the data to Python dataframe
2. Loading data using encoding
    - Try using different encoding formats
    - Use **chardet** library to detect encoding format (Hint:ISO-8859-1)
3. Explore/understand data: .info(), .describe(), .head(), .tail(),.shape
    - This will give you a sense of what type of dataset you are dealing with
4. Unique Values check: Use of .unique() function

Pre-Modelling Steps: Data Cleaning

Numeric
Mean (no outliers)
Median (outliers)
Categorical
Mode

Advanced upGrad
Nearest Neighbour
Emplength → Age

- Data Cleaning
  1. Redundant Columns: Use info from Data Exploration steps
  2. Check the percentage of missing values.
  3. Remove all those with very high missing percentage.
  4. For columns with less missing percentage: perform data cleaning steps for both columns and rows
  5. Null Value treatment: Decide if you need to drop null values or impute dummy data into them
  6. Checking out the distribution to impute mean, median or mode.
  7. Check for data consistency to avoid running into issues while merging especially conversion to correct format for strings. Example: convert to lowercase

More than 90% missing values

10000 10 rows Drop
100 00 1000 rows

Pre-Modelling Steps: Data Analysis

upGrad

- Aggregation of data
  - Use .mean(), .median() type of functions to extract the best average metric for the purpose
  - Mean vs Median example
  - Join(Merge) Operation to combine dataset from different data frames
  - Avoid use of loops while writing code

Joining Tables

upGrad

Company

permalink	name	homepage_url	category_list	status	country_code	state_code	region	city	founded_at
/Organization/-Fame	#fame	http://livfame.com	Media	operating	IND	16	Mumbai	Mumbai	NaN
/Organization/-Qounter	:Qounter	http://www.qounter.com	Application Platforms|Real Time|Social Network...	operating	USA	DE	DE - Other	Delaware City	04-09-2014
/Organization/-The-One-Of-Them-Inc-	(THE) ONE of THEM,Inc.	http://oneofthem.jp	Apps|Games|Mobile	operating	NaN	NaN	NaN	NaN	NaN
/Organization/0-6-Com	0-6.com	http://www.0-6.com	Curated Web	operating	CHN	22	Beijing	Beijing	01-01-2007
/Organization/004-Technologies	004 Technologies	http://004gmbh.de/en/004-interact	Software	operating	USA	IL	Springfield, Illinois	Champaign	01-01-2010

Master frame

Rounds

company_permalink	funding_round_permalink	funding_round_type	funding_round_code	funded_at	raised_amount_usd
/organization/-fame	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	B	05-01-2015	10000000.0
/ORGANIZATION/-QOUNTER	/funding-round/22dacff496eb7acb2b901dec1dfe5633	venture	A	14-10-2014	NaN
/organization/-qounter	/funding-round/b44fbb94153f6cdef13083530bb48030	seed	NaN	01-03-2014	700000.0
/ORGANIZATION/-THE-ONE-OF-THEM-INC-	/funding-round/650b8f704416801069bb178a1418776b	venture	B	30-01-2014	3406878.0
/organization/0-6-com	/funding-round/5727accaeaa57461bd22a9bdd945382d	venture	A	19-03-2008	2000000.0

Wide-format

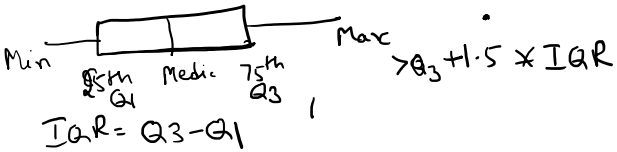
Mapping

category_list	Automotive & Sports	Blanks	Cleantech / Semiconductors	Entertainment	Health	Manufacturing	News, Search and Messaging	Others	Social, Finance, Analytics, Advertising
NaN	0	1	0	0	0	0	0	0	0
3D	0	0	0	0	0	1	0	0	0
3D Printing	0	0	0	0	0	1	0	0	0
3D Technology	0	0	0	0	0	1	0	0	0
Accounting	0	0	0	0	0	0	0	0	1

Walkthrough of Checkpoints

Outliers Detection

1) IQR (Box plot)



2) > 95th percentile
< 5 percentile

pd.melt(wide-df, id_vars (don't want to change), value_vars)

0	1	2	3	4
City	Name	Red	Yellow	Black
NY	Tom			

value_vars = df[2:]

id_var = df[0:1]

setdiff (columns, value_vars)

City, name

Points to remember

- The entire assignment is divided into checkpoints to help you navigate.
- For each checkpoint, you are advised to fill in the tables into the spreadsheet provided in the download segment(Investment.xls).
- Need to submit your insights in a ppt file. Sample PPT is provided. The structure is a suggestion; make sure not to exceed 10 slides.
- Convert the PPT in PDF format for submission. You need to submit a PDF.
- A single ZIP file is needed to be submitted with one Jupyter Notebook, one excel sheet and a PDF file(ppt).
- Don’t forget to comment the code properly as it carries separate marks.

Doubts??

upGrad

	Mean	Median
Seed	2	1.5
Angel	5	3.5
Venture	10	8.5
Private Eq	20	18.5

5m — 15m

Poll Questions

Q-1: When should we use Median to impute missing values rather than mean?

a) When there are so many data points
b) When you have so many missing values
c) When the data is having outliers
d) Can use both mean and median

Q-2: Should we drop the variable if we have 30% missing values

a) Yes
b) No
c) Depends how critical the variable is

upGrad

#LifeKoKaroLift

Thank You!