CSCE 5300 - Introduction to Big Data and Data Science

Project Title:

The Predictive and Successful Funding of Young or Early Stage Enterprises

Team 2 Members
Deepak Adimoolam
Venkateswarlu Amireddy
Richard Correia
Arun Kumar Reddy Kandula

Introduction:

At the beginning of our preparation, we considered all of the datasets that were made available. There were others that also fascinated us, such as the weather, flight and social media datasets, however, in the end we simply gravitated more towards this dataset focused on the funding/investing of relatively young companies. We think it is because many of the team members have aspirations of possibly starting a company or joining a startup venture, and felt that this would be a great topic to explore and analyze. As a result, we are excited to go through this journey and discovery, of applying the tools that we learn in class, to analyze the funding dataset.

Motivation:

Properly funding companies is a primary source for economic activity and innovation. There are many groups or organizations out in the market that raise funds and hope to invest in companies that will succeed and provide an acceptable return on the investment. However, like anything, there are many pitfalls to finding the right companies and industries/technologies to invest in. This project provides an opportunity to evaluate three data files that contain information on companies and industries that received investments between the 1990's and 2015. The data is broken out by regions, technology/industries, invested amounts, rounds of funding and their status (folded, still operating, acquired or went public). By performing the proper processing of the data and modeling the data, we think we can come up with some proper strategies to increase the success rate of choosing the right companies, industries, and locations to invest in. This is such a precursor to keep investments flowing from investors to companies deserving of the investments. Doing so, creates a dynamic environment in the economy to drive investment and growth.

Objectives:

The data files will need to be cleaned up and connected/merged to be able to do a complete analysis. Additionally, through the analysis, we want to identify the industries/technologies and regions to invest in, and determine if there is any discrimination between types of round investing and size of investments on the success of the company. One of our big goals is to create a model that can help us predict a path to success (acquired, IPO, and operating) vs shutting down.

Integration:

We are considering using either HDFS or AWS to store and process the data. Any modeling would likely be done through SPARK. A couple of the algorithms that we are considering for the machine learning part are classification or k-means since the predicted outcomes are going to be based on four different classes.

MLlib is Spark's machine learning library that offers scalable ML algorithms for classification, regression, clustering, and collaborative filtering.: Apache Spark is a fast and general-purpose cluster computing framework that provides in-memory processing capabilities.

Spark's distributed processing engine enables ML algorithms to perform computations on large datasets in-memory, resulting in faster processing times compared to MapReduce. integrating Hadoop components with ML algorithms allows for more efficient data processing, faster model training, and better insights from large and complex datasets.

Significances:

We see this as an exciting project. The impact of targeted and successful investments into the right companies and industries, as well as regions, can have profound implications on finding additional sources of money and delivering market momentum. That is why it is so important to develop a process that focuses on analyzing data over a long period of time (15 to 29 years) and create a predictive model that is focused on the actions needed to take a company from a fledgling idea/ startup to a full-blown success. It is interesting to note that small businesses, especially new ones, can become the lifeblood of an economy.

Features, deliverables, uniqueness, and milestones. Technical features of the project:

One of the key features is to deliver a process and model that can be easily updated and reused as new data becomes available on investments.

Deliverables	**Dates**
Proposal	Feb 10
Decide on platform to store data	TBD
Create necessary directories and folders	TBD
Store data	TBD
Begin early-stage cleanup of data	TBD
Connect the data files	TBD
Technical Features	
EDA	TBD
Visualize the data	TBD
Investigate several potential ML models	TBD
Break up the data for training and testing	TBD
Use hotkey to change categorical data to numerical	TBD
Apply min-max and normalization techniques	TBD
Apply oversampling as necessary	TBD
Measure model performance based F1 score and accuracy	TBD
Select model	TBD
review and detail results	TBD
Process data to understand trends and commonality, or centers of momentum	TBD
Visualize the data	TBD

Uniqueness:

As part of this project we are trying to analyze trends of the growing number of startups based on their funds , user base, market capitalisation and market base how well they would be performing or going to an IPO or being acquired in the near future when there is no sufficient capital funds for the organization to be stable and to be continuing with how it started off.

Our platform analyzes data over a long period of time (15 to 29 years) and create a predictive model that is focused on the actions needed to take a company from a fledgling idea/ startup to a full-blown success. This unique predictive model will help Investors to take data driven decisions to maximize their return on investment.

Visualizations:

We will create several bars and scatter plots on the data to gain some sense of trends, dominant categories/classes and correlations. It is important to visually see the data, vs blindly following what seems to be hot at the time.

Hadoop Components:

Hadoop MapReduce : It works on the principle of single master Node and multiple worker nodes,

- Client submits the job to master node, master splits each job into tasks and assigns it to worker nodes.
- Operated on key/value pairs, the input file is split into multiple blocks.
- Mapper transforms and filters the input data
- Reducers aggregate the mappers output

HDFS : It works on Client server architecture principle and is highly fault tolerant and designed to be deployed on low cos commodity hardware, used in applications which have large datasets.

- one Name node and multiple data nodes make up the HDFS architecture
- client can interact or write to Data Nodes as well
- All the metadata related information is placed in namenodes
- Files are split into one or more blocks and are stored in DataNodes

Machine Learning Algorithms:

Our objective is to predict a path to success for investors. As we already have the labels for the companies (folded, still operating, acquired or went public), based on which we can determine if the industry/companies were successful. As we have a labeled dataset we would mainly focus on supervised

machine learning models, as each model has its own significance we will be exploring multiple models but not limited to Multinomial Logistic Regression, Decision Trees for better explainability, because investors as putting in money and will expect reason for predictions made by the model rather than blindly trusting the prediction because their money is at stake, we will also explore advanced ML algorithms like Bagging and Boosting. We might also look into unsupervised models like Kmeans clustering to see if unsupervised learning is sufficient to fit the data.

Data Workflow:

We have 3 data files related to multiple features of industry type, funding rounds, etc. we will merge all the information from these data sources and load it into HDFS/S3 and move the data into Spark environment for data analysis. After fetching the data from HDFS to Spark environment we will do the data cleaning and handle the missing values, followed by EDA to understand the information in the available features, which is the most important step in any Data Analysis. Next step is Data Analysis, where we will be extracting meaningful insights from the data and show it to the investors with some Data Visualizations.

We can integrate this Big Data part with Data Science by building ML models to make predictions of the future investments. For this we will use the cleaned dataset from the spark environment and implement Feature Engineering if the existing features are not sufficient to fit the data. Once all the features sufficient to fit the data are decided we will split the train and test dataset and start the training process. followed by evaluating the trained model on the test dataset, if the accuray on the test data is less we will tune the hyperparameters and retrain the model. Once the model produces predictions with acceptable accuracy, our model will be ready for deployment in practical applications. Figure 1 is the workflow diagram explaining all the steps involved in our end-end data model.

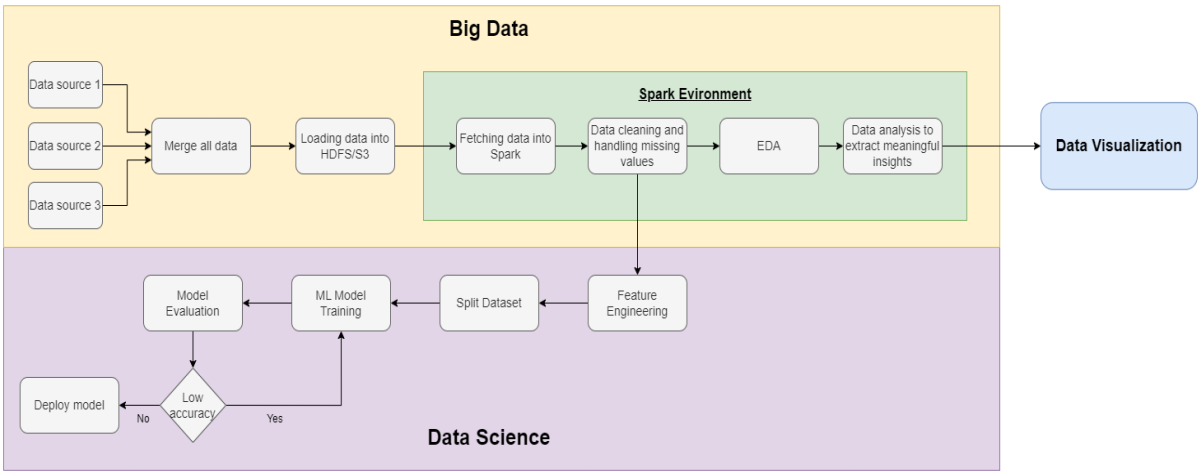

Figure 1: Data Workflow diagram

References

https://phoenixnap.com/kb/hadoop-mapreduce

https://labelyourdata.com/articles/how-to-choose-a-machine-learning-algorithm

https://towardsdatascience.com/explainable-ai-xai-with-a-decision-tree-960d60b240bd

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3167812

https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/