**Career Path in the Data Science and Machine Learning Industry**

-A study based on Kaggle Data Science Survey 2017

Zhou Yihui

013799500

San Jose State University

**Overview**

In this project, the paper focuses on the 2017 Kaggle data science and machine learning survey1, which gathers answers from more than 16,000 respondents who work in the field to answer questions on how the data science job is like, what are the requirements to enter the field, and what skillsets are required at job. The paper provides methods to analyze and explore datasets and has stored its findings in a jupyter notebook.

**Interest declaration and dataset selection**

Prior to my education at San Jose State University, I worked as a private banker at JPMorgan in Hong Kong, independently covering more than 10 ultra-high-net-worth individuals and families where my daily responsibilities include making investment decisions, pitching investment ideas to clients and maintain client relationship.

Early this year I decided to move to the bay area for family reasons and also because of some trends I have spotted earlier in my personal development.

I am always interested in communicating with my clients and I learn a ton from them. I am also interested in observing client behaviors and try to simplify my daily job. Here are a few interesting findings during my job. I observed that clients are easier to accept my proposal if I call them in the late afternoon or after lunch. When I try to cold call in person, in an entire building, I would prefer to start from the top floor instead of the ground floor because every attempt I make brings me close to this strenuous journey of prospecting. I built a few bond tables to automatically refresh and link with Bloomberg terminal to send to client without me monitoring and cross checking which adds large probability of human error into the automation. The core of the private bank business is indeed client service – when a client calls the office, the phone must be picked up after 3 seconds. Junior analysts are required to sit through lunch and late evening to pick up the phones! It is so hard to imagine that. And in fact, many of the calls are not urgent nor significant at all, simply inquiring about bond price or stock price. So I suggested my senior to have automated line system to direct client to the right person instead. They said no.

On the investment side, the trend in the financial market is changing so fast. As I recall, the recent several sharp drawdowns all have something to do with automation strategies, automated trading, or passive investment strategies like ETFs. The traditional way of bottom up fundamental research, the classic way still taught at most business schools, is taking less a portion in the market and will continue its declining trend. I deeply feel the challenge from data and computer science but as a traditional salesperson, I feel deeply helpless too. So instead of waiting to be replaced by machines one day (maybe that will not happen), I decided to make a change first. I would like to find the truth underneath the sales behaviors I observed, the patterns of the algorithm trading and the why and how data is going to disrupt the financial industry. So the dataset, Kaggle Survey challenge 2017 is perfect fit for me at this stage to get some insights from the survey conducted from insiders in the data science industry.

**Importance of dataset**

Given the popularity of the data science field, I constantly hear doubts and noises around the subject matter. I think there are mainly three aspects that add to the complexity. First, data science is an ever-evolving industry and it is challenging for anyone to fully grasp what is going on in the field. Second, data science jobs are indeed popular and but still in much less demand than software engineers. Third, I have made up my mind to transition my career but I am not completely sure which track I should continue. So I hope that by studying the Kaggle survey dataset I could get further insights on the matter.

**Initial purpose and aspirations**

Setting clear goals is of utmost important to the success of a data science project.

In the book *Think like a data scientist*2, Brian proposes that the main goal for a data science project should be set at the beginning of the project, as in many other fields. We need to identify a customer for the project. In this case, the project is designed and executed for exploring the machine learning/data science career track, so customers would be students like me who are looking for career transition. We could apply filters to the goals to set solid project goals by asking: (1) What is possible? (2) What is valuable? (3) What is efficient?

In this case, as previously I have identified my major concerns and confusions, I would like to set my goals to clear the confusion and help me make better career decisions. In this project, I would like to achieve below three goal: first, to present an overview of the data science industry including job titles, responsibilities, education background; second, to find out incentives of entering the data science field including salaries, career path, motivations behind other people in the industry; third, to find out the necessary tools to enter this field including which language to use, and required skill sets.

**Dataset Studies**

According to the book2, *Think like a data scientist*, the second step is to explore available data There are three ways how data is accessed: through a file system, in a database, or behind an application programming interface (API).

**Source**

According to the book2, *Think Like a Data Scientist*, the 3rd step is data wrangling, which refers to the process of taking data and information in difficult, unstructured, or otherwise arbitrary formats and converting it into something that conventional software can use. In this case, the data is already stored in clear format in csv otherwise I would need to adopt programming languages, or a set of tools to do the wrangling process.

The source of the dataset is from Kaggle, a well-known platform for data science. Kaggle organizes data science and machine learning industry wide surveys and in 2017 Kaggle received more than 16,000 replies.

The full data set could be found here: <https://www.kaggle.com/kaggle/kaggle-survey-2017/home>

**Description of & structure of the data Layout**

The dataset contains five files. According to website1, below is the brief introduction of the five datasets:

(1) Schema.csv: a CSV file with survey schema. This schema includes the questions that correspond to each column name in both the multipleChoiceResponses.csv and freeformResponses.csv.

(2) MultipleChoiceResponses.csv: Respondents' answers to multiple choice and ranking questions. These are non-randomized and thus a single row does correspond to all of a single user's answers.

(3) FreeformResponses.csv: Respondents' freeform answers to Kaggle's survey questions. These responses are randomized within a column, so that reading across a single row does not give a single user's answers.

(4) ConversionRates.csv: Currency conversion rates (to USD) as accessed from the R package "quantmod" on September 14, 2017

(5) RespondentTypeREADME.txt: This is a schema for decoding the responses in the "Asked" column of the schema.csv file.

**Domains of attribute values**

The fourth step in the book2, preliminary assessment calls my alert with regard to dealing with problems within data such as outliers, biases, precision, specificity, or any number of other inherent aspects of the data. Descriptive analysis is one good way to examine the validity of data. In fact, the project applied descriptive analysis on some data columns and I noticed that some of the answers provided by respondents are not valid answers. For example, one of the columns in the multiple-choice responses is age. I examined the range of the answers in age; surprisingly, the min answer of that is 0 and the max if 100, which would not make much sense for the machine learning industry.

Most of the survey answers are contained in the multipleChoiceResponses.csv and freeformResponses.csv file.

In the multiple-choice questions, the responses are chosen from a set of answers. The values of the various attributes are mostly in string and numeric forms.

In the freeform responses, the responses are generated entirely by the surveyed individuals. The answers from this part is more sparse and random than the multiple choice part, so the data study will mostly be based on the multiple choice dataset.

In the multiple-choice dataset, there are 16,716 number of respondents.

**Notes on exploration of the dataset**

After preliminary assessment of the data, next I decided to explore more about the data set and noticed that dealing with missing values are quite important in generating valid results from data set.

**Questions to ask of the dataset**

What are you hoping to achieve? Now that you should have a better understanding.

The paper would like to tackle and dig into the dataset from different segments and perspectives to solve my questions about a career in data science.

The first question I would like to ask of the dataset is: “What are the backgrounds of the respondents who have entered the industry already?” So the paper starts from fundamental questions of the respondents, including countries, gender, education, and salaries. A quick review of the dataset showed that, majority of the answers come from respondents come from the United States, followed by India; instead China ranked only top 6. The United States also rank top as the highest paying country. As Contrast, India ranked at bottom in the compensation list. Another sharp contrast with regard to background and salaries is gender. Among the respondents, only 17% of them are female. When it comes to salaries, the contrast is obvious. Whereas the median and lower quantile of the genders pose no difference; male salary distribution is heavily right skewed with many outliers on getting high salaries. The age distribution in the industry is worth noting too: most people cluster in the age range 25-30. Respondents also come from a quite diverse education background with computer science, mathematics or statistics, and Engineering as top three common education background. Masters dominate this field. One interesting thing to note is that salaries between doctors, masters, and bachelors in the field does not vary as expected.

The second question I would like to get an answer from the dataset is: “What are the data science job about? What techniques are used?” One of the top fields in the data science field is machine learning. According to the results from the survey, the top three most used ML techniques are Logistic Regression, Decision Trees/Random Forest, and Support vector Machines (SVMs). The dataset also provides forecast on next year’s most popular ML technique: deep learning and the tool used, without doubt, is Tensorflow. When it comes to programming languages, the diverging trend is visible between top two popular languages: Python and R. R is more popular with data visualization, and to analyst jobs. Python is more popular with data scientist, machine learning and software engineer jobs. A few top linked tools with Python are: jupyter notebook, SQL, Tensorflow, Unix, and Amazon Web Services. The mostly commonly used libraries in Python are Sklearn, Pandas, Numpy, Tensorflow and Keras.

The third question is: ”How do people in this industry feel about that job? How do I transition into the field? ” Among the respondents, the job title with highest satisfaction is data scientist, consistent with what I have expected. One trend from the respondents is that many respondents come to the data science from various previous roles such as researchers with salaries change up 20% or more. The biggest challenge people in the industry face is dirty data, which is consistent with what I have learned before that 80% of data scientist job is about data wrangling and preprocessing. One last thing I noticed from the dataset is that, most respondents comment that the way of finding a job in the data science field is through friend, family, former colleague, or internal recruiters.

**Summary**

**What you have learned about the dataset**

Throughout the project, I learned three things. First, I learned what data science job in general is about. From results of survey, 80% of their job involves data cleaning and wrangling. The rest 20% is about generating insights form dataset.

Second, I learned about how to get a job in data science and what skill sets are needed. Personal and professional network is the dominating way to get job interviews and learn about position openings. Three kinds of skills are needed for data science job, including domain knowledge, math and statistics, and computer science knowledge.

Third,I learned that given the rise of big data and the many solid knowledge required for data scientist job, it is worth career transitioning. As my aspiration resides in becoming data scientist, I have made my goal clear and firm that I will learn more solid and fundamental knowledge in computer science.

**About being a data scientist**

Data scientist takes up many skills and many of the skills are interdisciplinary. According to the data science Venn chart, math& statistics knowledge, hacking skills, and substantive domain knowledge make up the major three areas of data science skillsets. Data analysis is indeed an iterative process. As data scientist need to state the question, build the mode, interpret the results and communicate findings.

**About what skills you might need for the future**

Data is the new oil for the future in the digital economy and brings huge monetization potential

big data technologies. Within the data sphere there are many different forms of data, including structured data, semi-structured and unstructured data. Skill sets to analyze data and transform them is really important.

Data preprocessing skills are very important too as it is ranked as the most challenge job in data scientist job. In this project, the data set used is already, to some extent, preprocessed and saved in csv files which has reduced the difficulty in analysis. The data sets pose traits of tidy data too such as that each variable measured is in one column, and tables are linked using the same column. With regard to the big data infrastructure, there are many applications and concepts such as data warehouse and data pipelines. Within each application/concept, there are many systems involved.

Database language is another area I need to focus on. SQL query is surprisingly, a very important language working with a large amount of data.

**Reflection on the process**

Stating the right questions are of utmost importance and there are many types of right questions including descriptive, exploratory, inferential, predictive, causal, and mechanistic, just to name a few. Throughout the process, I feel this process is in fact not that easy.

References:

1. Kaggle Machine Learning and Data Science Survey 2017,<https://www.kaggle.com/kaggle/kaggle-survey-2017>

2. Godsey, B. (2017). Think Like a Data Scientist. Shelter Island, NY: Manning Publications Co.