Data Science in Industry

Software Engineer Salaries in Washington, DC Area
- Average Base Pay: $81,360/yr (7% below national average)
- Additional Cash Compensation: Average $7,049
- Range: $1,618 - $18,353

Data Scientist Salaries in Washington, DC Area
- Average Base Pay: $92,158/yr (12% below national average)
- Additional Cash Compensation: Average $9,627
- Range: $3,195 - $22,206
MODERN DATA SCIENTIST

Data Scientist, the newest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand what a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS
- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE
- Computer science fundamentals
- Scripting language e.g., Python
- Statistical computing packages e.g., R
- Databases: SQL and NoSQL
- Behavioral algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and HIVE/PG
- Custom reducers
- Experience with cloud like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS
- Passionate about the business
- Curious about data
- Influenced by authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION
- Able to engage with senior management
- Story telling skills
- Translate data driven insights into champions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g., Flare, (3D), Tableau
What is the hardest part of data science?
What is the hardest part of data science?

1. Dealing with people 😊
2. Figuring out what questions to ask (domain knowledge)
3. Getting the data for those questions
4. Organizing the data
5. Dealing with missing data
6. Training supervised Machine Learning
How to become successful?
Association Rules

1. Gather all frequent item sets (specified by support % of occurrences)
2. From those frequent items consider each possible combination and calculate the supports.
3. The most frequent rules that match our support level are presented.
Basic idea behind rule generation

- If \{A,B,C,D\} is a frequent itemset, candidate rules:
  - ABC → D
  - ABD → C
  - ACD → B
  - BCD → A
  - A → BCD
  - B → ACD
  - C → ABD
  - D → ABC
  - AB → CD
  - AC → BD
  - AD → BC
  - BC → AD
  - BD → AC
  - CD → AB
Association rule terminology

- Confidence: $\text{Confidence} (x \rightarrow y) = \frac{\text{support}(x \cup y)}{\text{support}(x)}$
- Lift: $\text{Lift} (x \rightarrow y) = \frac{\text{support}(x \cup y)}{\text{support}(x) \times \text{support}(y)}$
  - If the rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.
- Conviction: $\text{conv}(X \Rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)}$
  - interpreted as the ratio of the expected frequency that $X$ occurs without $Y$ (that is to say, the frequency that the rule makes an incorrect prediction)
  - if $X$ and $Y$ were independent divided by the observed frequency of incorrect predictions.
Jupyter Example
Clustering

K-Means will give you clusters that you can label!
Putting it together

1. From clusters you can label them which allows you to engineer statistically the association rule rolle ups.
2. From there you can see all the rules and target a dependent variable
3. Set your decision tree, random forests, or neural network to target this dependent variable and independent variables
4. Update with feedback from the field and you don’t have to worry about changing any code. Plus what code/variables would you change? The time savings of ML 😊.
Questions 😊