DATA SCIENCE

Data Scientist (n): (Pretty close to the actual description) Person who is better at statistics than any software engineer and better at software engineering than any statistician.

1. REQUIREMENTS

1.1 Computer Science aspect

- Programming language: Python or R
- Command line and Shell Scripting
- Basic understanding of OS (CPU, GPU)
- SQL: Basic querying, indexing, database understanding
- Analytics platform: JupyterLab, Anaconda, pySpark
- Libraries: pandas, numpy, matplotlib, nitk, sklearn, seaborn, R-caret, tensorflow, keras
- Version Control: GitHub/Bit Bucket
- Docker/Kubernetes

1.2 Stats and Math Aspect

- Basic stat i.e. mean, median, mode, hypothesis testing, bias, variance, percentiles, p-value, f-test, t-test
- Probability distributions, Bayesian statistics, Linear algebra, Calculus
- Pearson's and Spearman's correlation, K-means clustering, Chi-mean-square, Generator functions, Linear Regression, Logistic regression, PCA (principle component analytics), Linear Discriminant analytics, Resampling methods, Subset Selection methods – Forward selection, Backward Selection etc.. f-dimension reduction, Support Vector Machines

1.3 Domain/Business Knowledge:

- Understanding business/industry
- Practicality comes from experience
- Understanding end customer
- Define metrics that matter
- Ask good questions
- Make it actionable
- Translate for non-technical audience
- Story telling

1.4 Software Dev:

- Model deployment/Productionization
- Version control
- Agile process/Sprint
- DevOps
- Code optimization

1.5 Machine learning and Data Science:

- AI, Machine learning models: Supervised, Unsupervised Reinforcement, NN, CNN, Deep Learning, Data Prep/Exploration
- Feature engineering (the most important step at this stage)

2. DIFFERENT KINDS OF ROLES AT THE WORKPLACE:

- DS (Data Scientist): Build experiments and predictive models
- DE (Data Engineer): Build efficient data pipelines
- MLE (Machine Learning Engineer): Build ML pipelines, deploy models and optimize them
- DA (Data Engineer): Build reports, visualization
- BA (Business Analyst): Build reports, visualization (with management experience)

3. DATA SCIENCE PROCESS

- Scoping (Most important step)
- Data Exploration (Most fun)
- Data Preparation (Most time consuming)
- Feature Engineering (Most fun and need skills)
- Modeling (High skill job but relatively quick to do)
- Automation and Deployment (Needs CS skills)
- Maintenance (Something nobody wants to do)

3.1 Scoping:

- Understanding the Business Problem
- Defining the Return on Investment (Reduce cost, Increase revenue..)
- Collect the Voice of the Customer
- Additional Data
- Fill out Documentation

3.2 Data Exploration:

- Understand all datasets
- Ask relevant questions
- Perform EDA
- Brainstorm over data
- Don't figure it out all on your own. Seek help from the experts
- Avoid analysis paralysis

3.3 Data Preparation:

- Understand and Define inputs
- Conceptualize the train and test split
- Build the train and test blocks
- Check Validity of all datasets

3.4 Feature Engineering: "Feature engineering is an art"

- Transform columns to useful features
- Feature Validation
- Document any findings, concerns or choices made

3.5 Modeling: (Most mistakes happen here)

- Document the experiment and approach
- Build Model Pipeline
- Tune Hyper Parameters
- Document the performance metrics
- Any other optimizations

3.6 Automation and Deployment:

- Implement in a process and optimize for resources
- Automate Health Checks or Performance reports
- Model Update cadence and framework
- Work with client or client success manager to implement the model
- Stress test the model for performance
- Combine all documentation and create master reference file

4.1 BLOGS AND RESOURCES

- Kaggle.com
- Univ Irvine Database
- Analytics Vidhya
- Machine Learning Mastery
- Towards Data Science
- Reddit: Data Science and Machine Learning Channel
- Data Science Central
- datatau.net
- Hackernews
- Machine Learning (Coursera) by Andrew Ng
- Deep Learning by Ian Goodfellow

4.2 PATH TOWARDS ANALYTICS

- Decide what you want to do/be? DE, DS, BA, MLE? Go in a certain direction, analyze your skillsets, strengths, weaknesses
- Simply Google for a day or two and decide which references you want to follow accordingly
- Be theoretically equipped with AI/ML/DL space. Understand the math behind every technique.
- Work on your CS skills such as Python, R, Command line, SQL. No need to memorize, just get the basics right.
- Build your portfolio i.e. Start a Kaggle project. Form a Team, pick a dataset and build data reports as well as dashboards. Upload your work on GitHub. You need to work on Real data (if not, pick datasets from the internet). Teach each other.

5. PROJECTS AT THE INDUSTRY LEVEL

5.1 SERVICE ANALYTICS

Industry Consumer Appliance: Fortune 500

Objective:

- Reduce truck rolls using historical customer complaints using the model scoring
- Predict parts to be carried by the tech agent when visiting the household

What worked well?

- The model is currently in production being used by all Americans contacting the call center
- The reason for success was having it well scoped from the start, kept it very simple, lots of eyes on the project, having the right roles in the team
- Dashboard to show ROI updated every day automatically

Challenges:

- Translating machine learning to english
- Disagreement on the ROI and right metrics to measure performance
- Not being used as being advised by the call center agents. Competing priorities.

Algorithm and Evaluation: Ridge Regression Classifier, Naive Bayes, Natural Language Processing- Tokenization, TF-IDF etc

Data size: Duration: 2003-2020. Close to a quarter of million rows, mostly text

Evaluation metric: Precision and Custom metrics

Team: DS, MLE, BA, SME

5.2 PREDICT VIRALITY

Industry: Media

Objective: Predict if an article is going to be viral so that they can decide ad strategy accordingly

What worked well?

- Model was performing well
- Updates every 5 minutes

Challenges:

- The end users i.e. editors were not bought in
- Not integrated in their system

Algorithm and Evaluation: NLP, Time series, Ensemble

Data size: RSS feed going back a few months, live data stream

Evaluation metric: R squared

Team: DS, SME

5.3 CUSTOMER CHURN

Industry: Cable, Banking

Objective: Predict the likelihood of someone leaving the service.

What worked well?

- Well known problem in the industry
- Agreement on the metrics
- Ready to take actions i.e. run campaigns
- Model predicts at 4 times higher likelihood for top 50% of population
- Data was pretty clean

Challenges:

- You need to know the data in and out, especially the way it was generated
- Sometimes it is hard to define churn
- Building training data is hard
- Explaining the model is HARD

Algorithms and Evaluation: Random Forest, XGBoost

Data size: Close to a billion rows, all rows concatenated

Evaluation Metric: Accuracy, AUC RDC

Team: DS, SME, DE, CSM

5.4 IOT- REMOTE DIAGNOSTIC

Industry: Consumer Appliances
Objective: Identify failure before the refrigerator actually fails

What worked well?

- The client sent 4 refrigerators and a pair of washer and dryer
- Promising R&D results
- Architecture was able to manage the flow of data

Challenges:

- Legal challenges for storing data
- Not a lot of machines out in the wild yet
- Data size is huge

Algorithm and Evaluation: Pattern matching algorithms, tons of visualization

Data size: Billions and billions of rows from sensors

Evaluation Metric: Accuracy

Team: DS, SME, DE, CSM, Engineers, Stakeholders

5.5 MANUFACTURING AI

Industry: Consumer Appliances

Objective: Identify if all components are present in a dishwasher before shipping

What worked well?

- Quick pilot results
- The platform allowed image processing quickly using proper storage
- Potential savings of millions of dollars

Challenges:

- Implementations requires more than data science
- Annotated data

Algorithms and Evaluations: Deep learning- CNN + advanced filters using Tensor Flow

Data size:

- 1000s of images in pilot
- Millions in production

Evaluation Metrics: Accuracy- Need over 99% accuracy to beat human

Team: DS, SME, DE, CSM, Engineers, Stakeholders, Hardware experts