Data scientists under 🎧

(Let’s help them together)

Sharath Kumar Dhamodaran

October 15, 2021
Disclaimer
...my expectations
...my expectations

Mathematics
Set Theory, Calculus, Linear Algebra

Machine learning
Supervised, Unsupervised, Reinforcement, Deep Neural Nets, Nonlinear Optimization

Statistics
Probability, Descriptive, Inferential, Bayesian, Stochastic
But the reality...

Machine learning
Supervised, Unsupervised, Reinforcement, Deep Neural Nets, Nonlinear Optimization

Mathematics
Set Theory, Calculus, Linear Algebra

Statistics
Probability, Descriptive, Inferential, Bayesian, Stochastic

Computing
Databases, Logging, Deployment, Memory Management, Parallel Computation
But the reality...

**Mathematics**
- Set Theory,
- Calculus,
- Linear Algebra

**Statistics**
- Probability,
- Descriptive,
- Inferential,
- Bayesian,
- Stochastic

**Computing**
- Databases,
- Logging,
- Deployment,
- Memory Management,
- Parallel Computation

**Machine learning**
- Supervised,
- Unsupervised,
- Reinforcement,
- Deep Neural Nets,
- Nonlinear Optimization

**Relationship & Communication**
- Advising,
- Negotiating,
- Managing expectations,
- Story Telling
But the reality...

Business & Science

Business Functions,
Problem Framing, Design Thinking
Ethics, Bias, Fairness, Interpretability

Mathematics
Set Theory, Calculus, Linear Algebra

Statistics
Probability, Descriptive, Inferential, Bayesian, Stochastic

Machine learning
Supervised, Unsupervised, Reinforcement, Deep Neural Nets, Nonlinear Optimization

Computing
Databases, Logging, Deployment, Memory Management, Parallel Computation

Relationship & Communication
Advising, Negotiating, Managing expectations, Story Telling
But the reality...

Coaching & Mentoring
Share Best Practices, Staff Development

Machine learning
Supervised, Unsupervised, Reinforcement, Deep Neural Nets, Nonlinear Optimization

Mathematics
Set Theory, Calculus, Linear Algebra

Statistics
Probability, Descriptive, Inferential, Bayesian, Stochastic

Business & Science
Business Functions, Problem Framing, Design Thinking, Ethics, Bias, Fairness, Interpretability

Computing
Databases, Logging, Deployment, Memory Management, Parallel Computation

Relationship & Communication
Advising, Negotiating, Managing expectations, Story Telling


But the reality...

Technology Toolbox

- Big Data
  - Hadoop, Spark

- Computing
  - Python, R, GIT, Julia

- Cloud
  - AWS, GCP, Azure

- Frameworks
  - TensorFlow, PyTorch, Keras

- BI
  - PowerBI, Tableau

- Databases
  - SQL, NoSQL

Machine learning
- Supervised, Unsupervised, Reinforcement, Deep Neural Nets, Nonlinear Optimization

Statistics
- Probability, Descriptive, Inferential, Bayesian, Stochastic

Coaching & Mentoring
- Share Best Practices, Staff Development

Business & Science
- Business Functions, Problem Framing, Design Thinking
- Ethics, Bias, Fairness, Interpretability

Computing & Communication
- Databases, Logging, Deployment, Memory Management, Parallel Computation
- Advising, Negotiating, Managing expectations, Story Telling
Data scientists that do 70% of these are the best of the best.
My best friend: Data Engineer
My best friend: Data Engineer

Data Pipelines
source (raw) -> destination (analytics-ready)
My best friend: Data Engineer

Data Pipelines
source (raw) -> destination (analytics-ready)

Build and oversee data pipelines
Build pipeline monitoring infrastructure
Data Science project flow
Data Science project flow

Scope & feasibility
data collection & labeling
codebase setup

Project Planning
Data Science project flow

- **Project Planning**
  - Scope & feasibility
  - Data collection & labeling
  - Codebase setup

- **Data Exploration**
  - Data pipelines
  - Literature review and EDA
  - Product complexity
Data Science project flow

Project Planning
- Scope & feasibility
- Data collection & labeling
- Codebase setup

Data Exploration
- Data pipelines
- Literature review and EDA
- Product complexity

Model Development
- Establish baselines
- Optimizations

Model Evaluation
- Data (all features are useful)
- Model (hyperparameters tuned, bias, fairness, interpretability)
- Infrastructure (training is reproducible)
Data Science project flow

Project Planning
- Scope & feasibility
- Data collection & labeling
- Codebase setup

Data Exploration
- Data pipelines
- Literature review and EDA
- Product complexity

Model Development
- Establish baselines
- Optimizations

Model Evaluation
- Data (all features are useful)
- Model (hyperparameters tuned, bias, fairness, interpretability)
- Infrastructure (training is reproducible)

Model Deployment
- Expose via REST API
- Version system
Data Science project flow

Scope & feasibility
- data collection & labeling
- codebase setup

Data Exploration
- data pipelines
- literature review and EDA
- product complexity

Model Development
- establish baselines
- optimizations
- data (all features are useful)
- model (hyperparameters tuned, bias, fairness, interpretability)
- infrastructure (training is reproducible)

Model Evaluation
- expose via REST API
- version system

Model Deployment
- performance checks
- data distribution drifts
- feature drifts

Model Monitoring
-
Data Scientists team structure
Data Scientists team structure

**CENTRALIZED**

data science
Data Scientists team structure

CENTRALIZED

data science

EMBEDDED

product

marketing
Struggles....
My personal struggles

Difficulty

Communication

Time
Not adjusting messaging based on audience
Not adjusting messaging based on audience

Average Quarterly Loan Default Risk

Since the p-value < 0.05, we reject the null hypothesis. Hence we can conclude that there is statistically significant negative correlation (-0.82) between an applicant’s FICO score and the interest rate charged.
Not adjusting messaging based on audience

Average Quarterly Loan Default Risk

Since the p-value < 0.05, we reject the null hypothesis. Hence, we can conclude that there is statistically significant negative correlation (-0.82) between an applicant’s FICO score and the interest rate charged.

Monthly Attrition Rates

Decision Tree misclassification rate is 7%, while Logistic Regression is 18%. So with an ensemble approach, we can capitalize on the strengths of both. Logistic Regression identifies predictors having strong influence on the target variable whereas decision tree identifies the high-risk churners more accurately.
Not adjusting messaging based on audience

Average Quarterly Loan Default Risk

Since the p-value < 0.05, we reject the null hypothesis. Hence we can conclude that there is statistically significant negative correlation (-0.82) between an applicant’s FICO score and the interest rate charged.

Monthly Attrition Rates

Decision Tree misclassification rate is 7%, while Logistic Regression is 18%. So with an ensemble approach, we can capitalize on the strengths of both. Logistic Regression identifies predictors having strong influence on the target variable whereas decision tree identifies the high-risk churners more accurately.
My personal struggles
Going straight to complex techniques
Going straight to complex techniques

Simple vs fancy models

Me

Deep learning

Simple problem
Going straight to complex techniques

Simple vs fancy models

Problems
Scaling Machine Learning solutions
Interpretability
Never give-up is the worst advice
Never give-up is the worst advice

90% of all machine learning models never make it to production
Never give-up is the worst advice

90% of all machine learning models never make it to production

Mistakes
Not embracing failures
Emotional decisions
My personal struggles
I depended a lot on data engineers, but I rarely admitted it.
I depended a lot on data engineers, but I rarely admitted it.
Underestimating the value of domain experts
Underestimating the value of domain experts

Analysis Paralysis (research, research,...)
Underestimating the value of domain experts

Analysis Paralysis (research, research, ...)

Setting too high a bar for success

(striving for 99.9% accuracy when 96% is fine with business folks)
Underestimating the value of domain experts

Analysis Paralysis (research, research, ...)

Setting too high a bar for success
(striving for 99.9% accuracy when 96% is fine with business folks)

Underestimating domain rules
(helps enhance model performance)
My personal struggles

- Media
- Relationships
- Ego
- Communication

Difficulty

Time
Are data scientists going to be useless in 10 years?
Are data scientists going to be useless in 10 years?

2012..
Are data scientists going to be useless in 10 years?

2012..

Data Scientist: The Sexiest Job of the 21st century

- Harvard Business Review
Are data scientists going to be useless in 10 years?

2012..
Data Scientist: The Sexiest Job of the 21st century
- Harvard Business Review

2018..
Are data scientists going to be useless in 10 years?

2012..

Data Scientist: The Sexiest Job of the 21st century
- Harvard Business Review

2018..

We need more data engineers, NOT data scientists
Are data scientists going to be useless in 10 years?

2012..
Data Scientist: The Sexiest Job of the 21st century
- Harvard Business Review

2018..
We need more data engineers, NOT data scientists
Will machine learning as a service make data scientists obsolete?
Automated Machine Learning (AutoML)
The killer of data scientists?
Automated Machine Learning (AutoML)
The killer of data scientists?

AutoML Workflow

Data Collection → Data Preparation → Model Selection → Model Training → Hyper Parameter Tuning → Model Validation
Automated Machine Learning (AutoML)
The killer of data scientists?

AutoML Workflow

Data Collection → Data Preparation → Model Selection → Model Training → Hyper Parameter Tuning → Model Validation

- Datarobot: $6.3B valuation
- H2O.ai: $400M valuation
Automated Machine Learning (AutoML)
The killer of data scientists?

AutoML Workflow

Data Collection → Data Preparation → Model Selection → Model Training → Hyper Parameter Tuning → Model Validation

Datarobot: $6.3B valuation

H2O.ai: $400M valuation

Preferred terminology
Data Mining → Data Science / AI → Automation
My personal struggles

- Media
- Relationships
- Ego

Difficulty

Communication

Time
Shield and Charge
Understand when to provide context and how deep
Understand when to provide context and how deep

**Monthly Attrition Rates**

Decision Tree misclassification rate is 7%, while Logistic Regression is 18%. So with an ensemble approach, we can capitalize on the strengths of both. Logistic Regression identifies predictors having strong influence on the target variable whereas decision tree identifies the high-risk churners more accurately.

I recommend launching this feature because it led to a 24% increase in predicting which customers will leave
Understand when to provide context and how deep

**Monthly Attrition Rates**

Decision Tree misclassification rate is 7%, while Logistic Regression is 18%. So with an ensemble approach, we can capitalize on the strengths of both. Logistic Regression identifies predictors having strong influence on the target variable whereas decision tree identifies the high-risk churners more accurately.

I recommend launching this feature because it led to a 24% increase in predicting which customers will leave.

**Predicting Business metrics**

- **Attrition Rates** → Which customers will leave?
- **Weekly maintenance costs** → Which parts are going to break?
- **Daily downtimes** → Which machine will cause an outage?
Start simple, release, and then iterate
Start simple, release, and then iterate

Process Monitoring (Anomaly Detection)

Average

Time
Start simple, release, and then iterate

Process Monitoring (Anomaly Detection)

Clustering instead of Variational Autoencoders
Start simple, release, and then iterate

Process Monitoring (Anomaly Detection)

Clustering instead of Variational Autoencoders

...or just Statistical Process Control (SPC)
Develop an understanding of data lineage and sourcing
Develop an understanding of data lineage and sourcing

Me in Grad School: models.. models.. models.. data
Develop an understanding of data lineage and sourcing

Me in Grad School: models.. models.. models.. data
Me in Industry: data.. data.. dataaa.. models
Develop an understanding of data lineage and sourcing

Me in Grad School: models.. models.. modelss.. data
Me in Industry: data.. data.. dataaa.. models

Data Planning

Where is the data going to come from?
When is it going to be available in reality?
How quickly can we get it?
How much data will there be?
Personal brand is on display for everyone to see
Personal brand is on display for everyone to see

Communicate frequently
Personal brand is on display for everyone to see

Communicate frequently

Always demand a deadline
Personal brand is on display for everyone to see

Communicate frequently
Always demand a deadline
Set the right expectations
Personal brand is on display for everyone to see

Communicate frequently

Always demand a deadline

Set the right expectations

Articulate project's output
Personal brand is on display for everyone to see

Communicate frequently
Always demand a deadline
Set the right expectations
Articulate project’s output

Produce a Minimal Viable Product (MVP)
Build a reputation for being dependable
Build a reputation for being dependable
Build a reputation for being dependable

Projects

76%  
Fail

24%  
Success
Build a reputation for being dependable

Projects

76% Fail
24% Success

Marketing 101: Skills don’t sell
Solutions sell, Results sell, Value sells
AutoML solves a real problem but it isn't perfect (yet)..
AutoML solves a real problem but it isn’t perfect (yet)....

Need faster insights - can I do it myself?
AutoML solves a real problem but it isn't perfect (yet)...

Need faster insights - can I do it myself?

We have data. Can we predict what will happen next?
AutoML solves a real problem but it isn’t perfect (yet).

Need faster insights - can I do it myself?

We have data. Can we predict what will happen next?

Stakeholders + Data Scientist: 3 - 4 months
AutoML solves a real problem but it isn't perfect (yet)...

Need faster insights - can I do it myself?

We have data. Can we predict what will happen next?

Stakeholders + Data Scientist: 3 - 4 months
Stakeholders + AutoML: 3 - 4 hours
AutoML is just Automated Machine Learning
and NOT Automate My Learning
AutoML is just Automated Machine Learning
and NOT Automate My Learning

Data

collection, measurement errors,
selection effects, identification, censoring
AutoML is just Automated Machine Learning and NOT Automate My Learning

**Data**
collection, measurement errors, selection effects, identification, censoring

**Science**
simpson’s paradox, figuring out sample size requirements, knowing how to communicate findings
AutoML is just Automated Machine Learning and NOT Automate My Learning

**Data**
collection, measurement errors, selection effects, identification, censoring

**Science**
simpson’s paradox, figuring out sample size requirements, knowing how to communicate findings

More valuable: defining what exactly is the problem to be solved
AutoML is just Automated Machine Learning and NOT Automate My Learning

Data
- collection, measurement errors,
- selection effects, identification, censoring

Science
- Simpson’s paradox,
- figuring out sample size requirements,
- knowing how to communicate findings

More valuable: defining what exactly is the problem to be solved

Less valuable: unhealthy obsession to tools and packages
Learnings
Communicate ideas clearly = advantage
Communicate ideas clearly = advantage

Unafraid to ask questions
Communicate ideas clearly = advantage
Unafraid to ask questions
Rejection is just redirection
Communicate ideas clearly = advantage
Unafraid to ask questions
Rejection is just redirection
Give credit to data engineers
Communicate ideas clearly = advantage

Unafraid to ask questions

Rejection is just redirection

Give credit to data engineers

Sales > Model building
Communicate ideas clearly = advantage

Unafraid to ask questions

Rejection is just redirection

Give credit to data engineers

Ride the bull
Adapt to change

Sales > Model building
Communicate ideas clearly = advantage

Unafraid to ask questions

Rejection is just redirection

Give credit to data engineers

Under-promise Over-deliver

Ride the bull Adapt to change

Sales > Model building
Communicate ideas clearly = advantage
Unafraid to ask questions
Rejection is just redirection
Give credit to data engineers
Sales > Model building

Keep process as lean as possible
Under-promise, Over-deliver
Ride the bull, Adapt to change
I still don't know very much...