Review of My Prior Data Science Projects

Zhehan (Tiffany) Zhu
University of Pittsburgh

Overview of Completed Data Science Projects

- **Number of Completion**
  - 30+ projects
- **Project Source**
  - 2 Ph.D. and 4 Master courses
  - 3 Independent studies
- **Team Size**
  - Independent
  - 3-9 members
- **Cross-functional Collaboration**
  - Information Science, Public Administration, Finance, Marketing

- **My contributions**
  - Data extraction
  - Data cleaning
  - Data modelling
  - Writing reports
  - Presentations
- **Programming Language**
  - R, Python, Stata

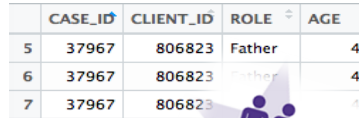
RStudio

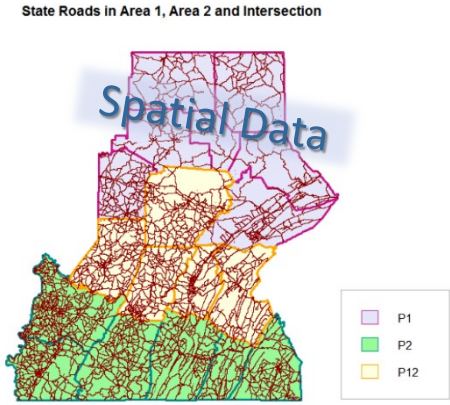
Notebook

Overview of Completed Data Science Projects

- **Data Source**
  - Public
    - Company websites, social media, Kaggle, etc.
  - Local Government
  - Research paper
  - Synthetic data
- **Data Type**
  - Structured data – quantitative, categorical
  - Unstructured data - image, text
- **Data Dimensionality**
  - Up to 73,000+ records
  - Up to 2,600+ attributes

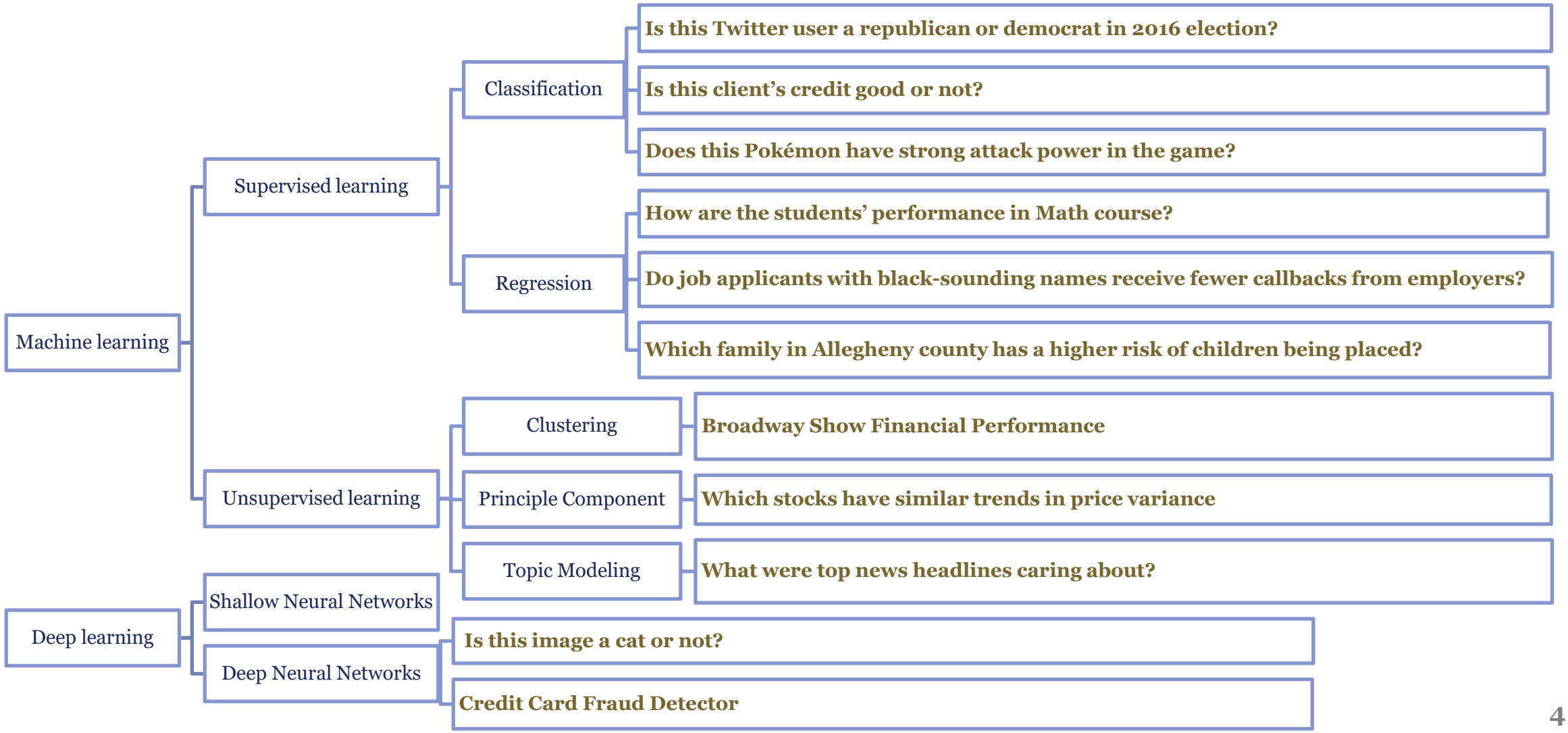
	CASE_ID	CLIENT_ID	ROLE	AGE	GENDER	ACPT_DT	CLOSE_DT
5	37967	806823	Father	48	Male	JAN-2016	APR-2013
6	37967	806823	Father	48	Male	JAN-2016	AUG-2006
7	37967	806823	Father	48	Male	JAN-2016	OCT-2014
8	37967	806823	Father	48	Male	JAN-2016	NA
9	37967	807660	Mother	38	Female	JAN-2016	APR-2013
10	37967	807660	Mother	38	Female	JAN-2016	AUG-2006
11	37967	807660	Mother	38	Female	JAN-2016	OCT-2014
12	37967	807660	Mother	38	Female	JAN-2016	NA
13	37967	907364	Child	5	Female	JAN-2016	APR-2013
14	37967	907364	Child	5	Female	JAN-2016	AUG-2006





	user_ID	text	created_at	favorite_count
44487	1763	How about you finish 4th grade before your next debate @realDonaldTrump You sound like a confuse...	10/09/2016 21:23:07 EDT	6
44488	7717	@EdwardSharpe6 third row. Dead center tv side.	09/26/2016 21:46:48 EDT	8527
44489	4112	Spoketh the spoken word master. #ImWit debates https://t.co/6klzXnpUb9	09/26/2016 21:51:46 EDT	23869
44490	4112	Fun fact: if police were actually afra ything, they wouldn't shoot so many unarmed black ...	09/26/2016 21:50:56 EDT	23869
44491	4112	"Charlotte, a city I love. A city where I h stments. Name after a very, very hot chick. Who would...	09/26/2016 21:48:37 EDT	23869
44492	4112	The race relations section of the debate. # debates	09/26/2016 21:45:08 EDT	23869
44493	4112	RT @slackmistress: Donald r to ore. #Debates2016	09/26/2016 21:43:49 EDT	23869
44494	4112	RT @HarvardBiz: Why Hillary Gets Interrupted More than Donald Trump https://t.co/uT82T9FPv...	09/26/2016 21:24:22 EDT	23869
44495	4112	Fun fact: asking Hillary to answer a policy question is as thrilling for her as it is for Trump to answer wh...	09/26/2016 21:21:40 EDT	23869
44496	4112	Trump: everything's bad, all the jobs are leaving. Holt: ANSWER MY ACTUAL QUESTION, CHEETO JESUS...	09/26/2016 21:18:02 EDT	23869

Predictive Analytics Algorithms Applied



Predictive Analytics Algorithms Applied

Type	Problem	Data Dimension	Language	Methodology	Team
Classification	**Is this Twitter user a republican or democrat in 2016 election?**	(60000+, 9)	Python	Classification: Decision Trees Topic modeling: Latent Dirichlet Allocation	3 members
	Is this client's credit good or not?	(1000, 21)	R	Logistic Regression Cross-Validation	Independent
	Does this Pokémon have strong attack power in the game?	(800, 13)	R	Logistic Regression K-Nearest Neighbors (K-NN) Naive Bayesian Decision Tree Support Vector Machine (SVM) Cross-Validation	Independent
Regression	**How are the students' performance in Math course?**	(600+, 33)	R	Linear Regression Non-Linear Regression Cross-Validation	Independent
	Do job applicants with black-sounding names receive less call backs from employers?	(4800+, 38)	R	Linear Regression T-Test F-Test	Independent
	Which family in Allegheny county has a higher risk of children being placed?	(8800+, ?)	R	Exploratory Data Analysis (EDA) Linear Regression	9 members

Predictive Analytics Algorithms Applied

Type	Problem	Data Dimension	Language	Methodology	Team
Clustering	**Broadway Show Financial Performance**	(1700+, 17)	R, Python	Web Scraping Factor Analysis K-Means Clustering Principal Component Analysis (PCA)	Independent
Principal Component	**Which stocks have similar trends in price variance?**	(127, 30)	R	K-Means Clustering Hierarchical Clustering Principal Component Analysis (PCA)	Independent
Topic Modeling	**What were top news headlines caring about?**	(73000+, 2)	Python	Topic modeling: Latent Dirichlet Allocation	Independent
Deep Neural Networks	**Is this image a cat or not?**	250 images	Python	2-layer neural network 4-layer neural network	Independent
	Credit Card Fraud Detector	(1000, 31)	Python	3-layer neural network Gradient Descent Momentum Adam	Independent