AI Club's Project Workshop

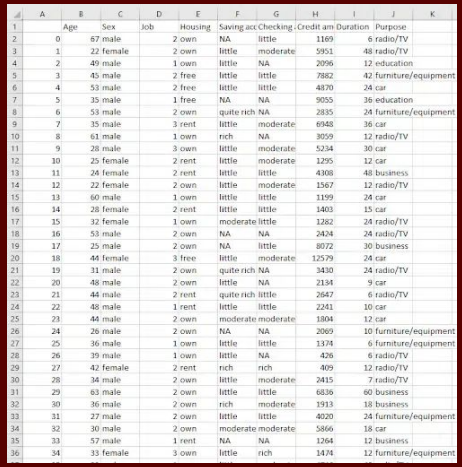
- Build your own ML project from scratch
- Add to your portfolio
- Compete for prizes

Week 2: Problem & Dataset Selection

- Definition of Done
  - Problem identified and defined
  - Dataset selected and accessible
- Extra Credit
  - Explore your data
    - What ranges are your features in?
    - How many labels do you have to predict?
    - Is what you're trying to predict equally represented in the data?
      - Do you expect it to be?
    - Are there any features you don't think you need?
  - Preliminary set of models identified
  - Train a model!

What Makes a Good Problem?

- Clean, available datasets
  - There is no model without data to learn from
- Interesting to you
- Some level of interactivity
- Classification problem?
  - Best suited to beginners, strongest support from us



	A	B	C	D	E	F	G	H	I	J	K
1		Age	Sex	Job	Housing	Saving acc	Checking a	Credit am	Duration	Purpose	
2	0	67	male	2	own	NA	little	1169	6	radio/TV	
3	1	22	female	2	own	little	moderate	5951	48	radio/TV	
4	2	49	male	1	own	little	NA	2096	12	education	
5	3	45	male	2	free	little	little	7882	42	furniture/equipment	
6	4	53	male	2	free	little	little	4870	24	car	
7	5	35	male	1	free	NA	NA	9055	36	education	
8	6	53	male	2	own	quite rich	NA	2835	24	furniture/equipment	
9	7	35	male	3	rent	little	moderate	6948	36	car	
10	8	61	male	1	own	rich	NA	3059	12	radio/TV	
11	9	28	male	3	own	little	moderate	5234	30	car	
12	10	25	female	2	rent	little	moderate	1295	12	car	
13	11	24	female	2	rent	little	little	4308	48	business	
14	12	22	female	2	own	little	moderate	1567	12	radio/TV	
15	13	60	male	1	own	little	little	1199	24	car	
16	14	28	female	2	rent	little	little	1403	15	car	
17	15	32	female	1	own	moderate	little	1282	24	radio/TV	
18	16	53	male	2	own	NA	NA	2424	24	radio/TV	
19	17	25	male	2	own	NA	little	8072	30	business	
20	18	44	female	3	free	little	moderate	12579	24	car	
21	19	31	male	2	own	quite rich	NA	3430	24	radio/TV	
22	20	48	male	2	own	little	NA	2134	9	car	
23	21	44	male	2	rent	quite rich	little	2647	6	radio/TV	
24	22	48	male	1	rent	little	little	2241	10	car	
25	23	44	male	2	own	moderate	moderate	1804	12	car	
26	24	26	male	2	own	NA	NA	2069	10	furniture/equipment	
27	25	36	male	1	own	little	little	1374	6	furniture/equipment	
28	26	39	male	1	own	little	NA	426	6	radio/TV	
29	27	42	female	2	rent	rich	rich	409	12	radio/TV	
30	28	34	male	2	own	little	moderate	2415	7	radio/TV	
31	29	63	male	2	own	little	little	6836	60	business	
32	30	36	male	2	own	rich	moderate	1913	18	business	
33	31	27	male	2	own	little	little	4020	24	furniture/equipment	
34	32	30	male	2	own	moderate	moderate	5866	18	car	
35	33	57	male	1	rent	NA	NA	1264	12	business	
36	34	33	female	3	own	little	rich	1474	12	furniture/equipment	

Limitations of Machine Learning

- Supervised ML Problem Types
  - **Classification**: Is this email spam or not spam?
  - **Regression**: What will the house price be?
- How to Frame a Problem for ML
  - Predict [specific outcome] based on [available data]
  - **Bad**: Buy and sell stocks/crypto
  - **Good**: Predict if Apple will gain or lose value based on yesterday's change
- Guiding Questions
  - Is your prediction target specific?
  - Do you have data to learn from?
  - Can you quantitatively measure success?

Project Ideas & Examples 1

- Pet Breed Classification
  - **Problem:** Identify cat and dog breeds from photos
  - **Dataset:** Oxford-IIT Pet Dataset
  - **Models:** CNN, ResNet
- Heart Disease Detection
  - **Problem:** Predict heart disease risk from medical data
  - **Dataset:** Cleveland Heart Disease Dataset
  - **Models:** Random Forest, Logistic Regression, SVM
- Network Security
  - **Problem:** Detect malicious network connections
  - **Dataset:** KDD Cup 99, NSL-KDD
  - **Models:** Decision Trees, Neural Networks, Ensemble

Project Ideas & Examples 2

- Music Genre Classification
  - **Problem:** Classify songs by genre from audio
  - **Dataset:** GTZAN
  - **Models:** SVM, Random Forest, Deep Learning
- Stock Price Prediction
  - **Problem:** Predict stock movements from historical data
  - **Dataset:** Yahoo Finance, Alpha Vantage
  - **Models:** LSTM, Linear Regression
- House Price Prediction
  - **Problem:** Predict home prices from property features
  - **Dataset:** Boston Housing, Ames Housing Dataset
  - **Models:** Linear Regression, XGBoost, Random Forest

Dataset Checklist

- Sufficient size (1000+ samples preferred)
- Accessible and downloadable
- Reasonably clean (some messiness is okay!)
  - Missing values
  - Bad headers for columns
- Legal to use

Where to Find Great Datasets

- General Recommendations
  - Kaggle: kaggle.com/datasets
  - UCI ML Repository: archive.ics.uci.edu/ml
  - Google Dataset Search: datasetsearch.research.google.com
  - Papers with Code: paperswithcode.com/datasets
- Domain-Specific Sources
  - Government: data.gov, census.gov
  - Finance: Yahoo Finance, Alpha Vantage
  - Images: ImageNet, COCO, OpenImages
  - Text: Common Crawl, Project Gutenberg

Example Project

- Kellen will now show off Week 2’s example project!

Jupyter Notebook Demo

- There may have been some confusion about the tutorials from last meeting
- Let’s go over what a Python Notebook is and how to use them

Let’s Build Something Amazing!

- **TODOs**
  - ✅ Form teams
  - ✅ Complete tutorials on unfamiliar tools
  - ✅ Choose your problem
  - ✅ Find, download, and verify your dataset
  - ✅ Tutorials Finished!
- **Resources**
  - Project idea slides
  - Dataset recommendations
  - Tutorials on our website; scan the QR code!
- **Questions? Stuck?**
  - Raise your hand! We’re here to help