Final Project

Spring 2024

Hongchang Gao

Topic

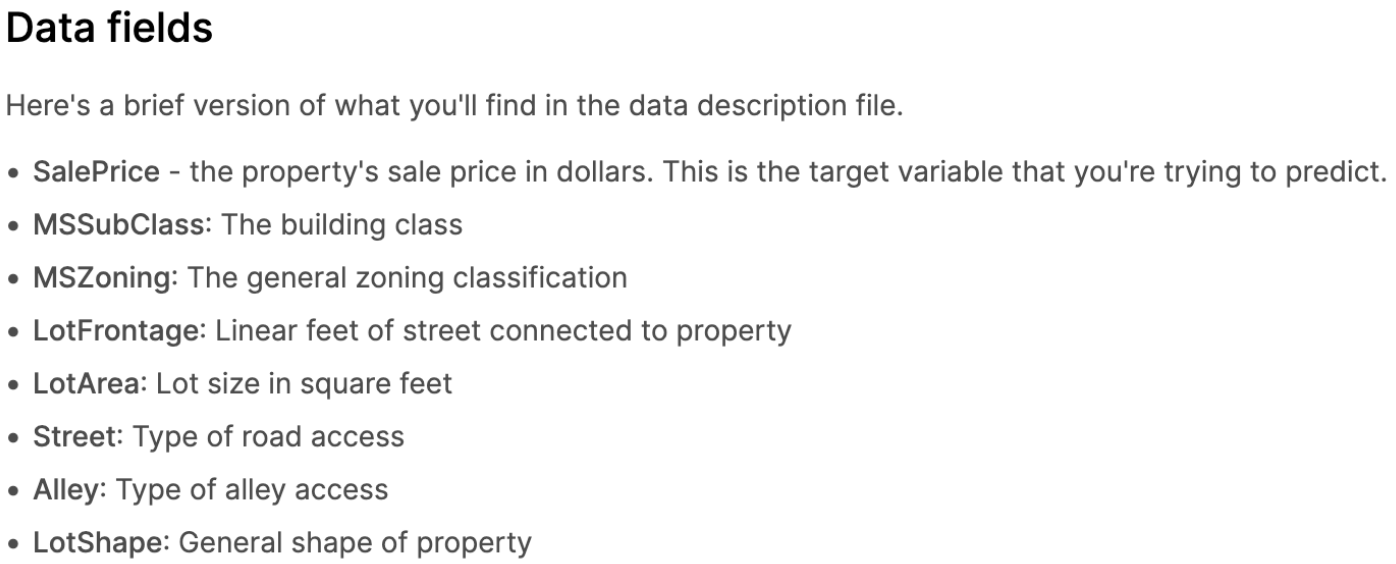
- Free to choose your topic
  - Supervised Learning
    - Regression
    - Classification
  - Unsupervised Learning
    - Clustering
- Examples:
  - Fraud detection
  - Image classification
  - Community detection
  - Review classification
  - Recommender system
  - ….

Topic (continue)

- Resources
  - Kaggle https://www.kaggle.com/datasets
  - AWS https://registry.opendata.aws/
  - https://www.opendataphilly.org/dataset
  - https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b

Topic (continue)

- Example 1: House prices:
  - https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview
  - With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.



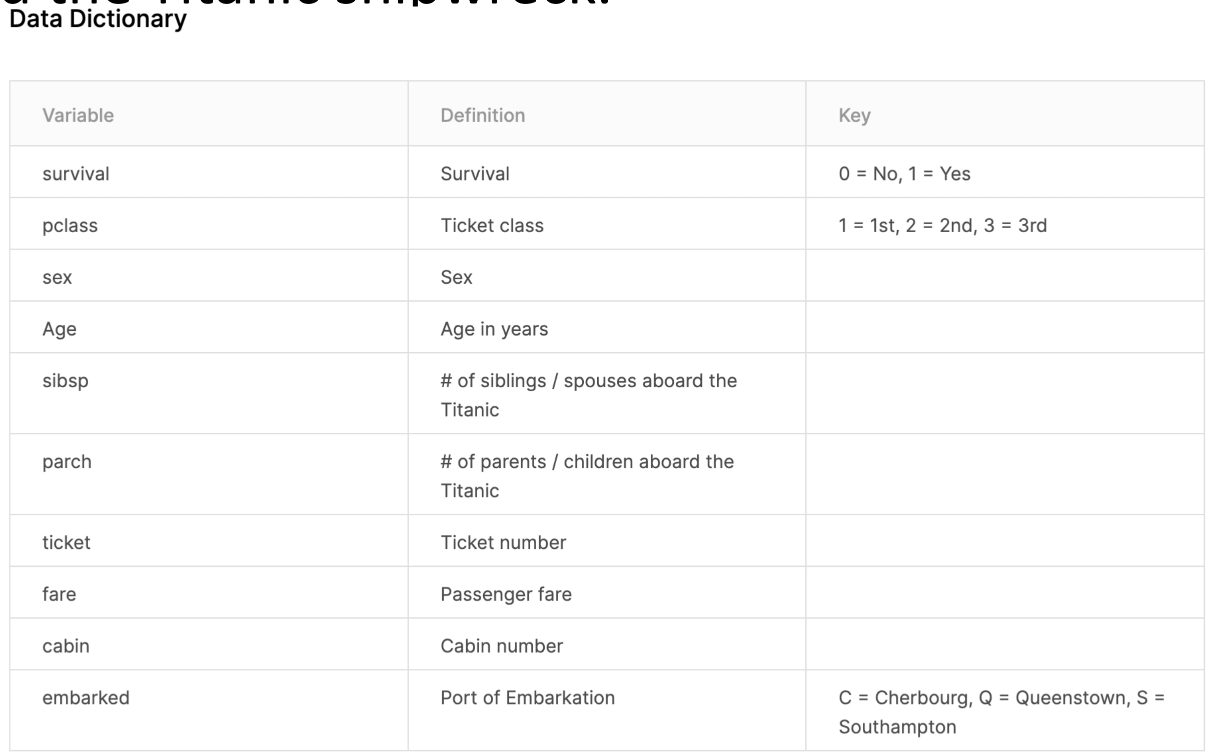
Data fields

Here's a brief version of what you'll find in the data description file.

- **SalePrice** - the property's sale price in dollars. This is the target variable that you're trying to predict.
- **MSSubClass**: The building class
- **MSZoning**: The general zoning classification
- **LotFrontage**: Linear feet of street connected to property
- **LotArea**: Lot size in square feet
- **Street**: Type of road access
- **Alley**: Type of alley access
- **LotShape**: General shape of property

Topic (continue)

- Example 2: Titanic - Machine Learning from Disaster
  - https://www.kaggle.com/c/titanic/overview
  - Use machine learning to create a model that predicts which passengers survived the Titanic shipwreck.

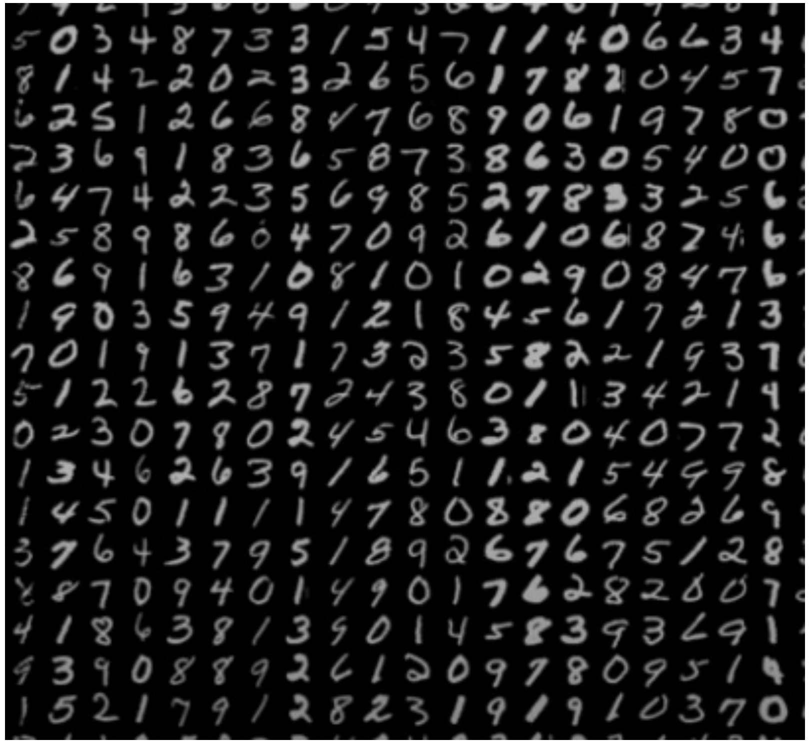


Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

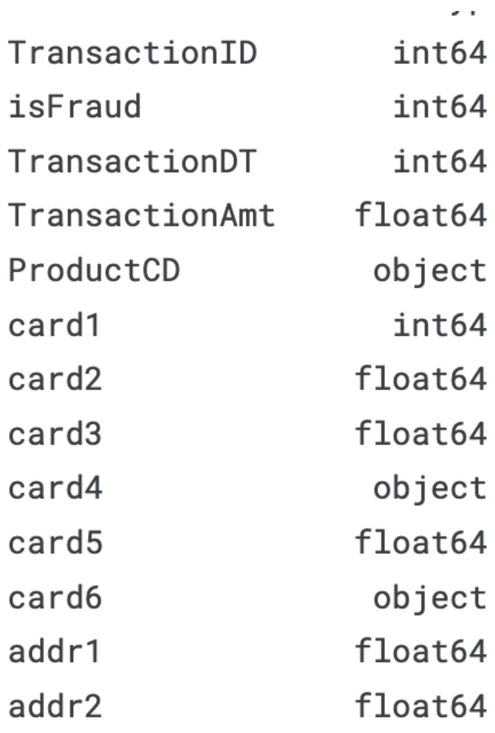
Topic (continue)

- Example 3: Digit Recognizer
  - https://www.kaggle.com/c/digit-recognizer/overview
  - Image classification



Topic (continue)

- Example 4: IEEE-CIS Fraud Detection
  - https://www.kaggle.com/c/ieee-fraud-detection
  - Detect fraud transaction



```
TransactionID       int64
isFraud             int64
TransactionDT       int64
TransactionAmt    float64
ProductCD          object
card1               int64
card2             float64
card3             float64
card4              object
card5             float64
card6              object
addr1             float64
addr2             float64
```

Topic (continue)

- Example 5: Santander Customer Transaction Prediction
  - https://www.kaggle.com/c/santander-customer-transaction-prediction/overview
  - Identify who will make a transaction
  - An anonymized dataset containing numeric feature variables, the binary target column, and a string ID_code column.

Topic (continue)

- Why is this task important?
- How to obtain/preprocess the data?
- How to explore the data?
- Which model to use?
- How to analyze the result?
- …

Project Team

- Each team can have
  - One or two students
- Amount of the work
  - The expected amount of work for the whole project per student is 30 hours.
  - The team with 2-person should double the work, ~60 hours

Project Format

- Project format
  - Proposal
  - Progress Report I
  - Progress Report II
  - Final Report

1. Proposal

- Proposal should include
  - Project title and student name(s)
  - Introduction section
    - Give motivation; describe the problem; review related works
  - Proposed work section
    - Explain the idea; explain proposed approaches
  - Timeline
    - State the timeline of the project
  - References
    - provide at least 2 related references (could be a web link)
- 1.5-2 page report
  - 11pt font size
  - Single space

2. Progress Report

- Summarize the progress
  - What has been done
  - What has not been done
  - What will be done during the following week
- 2-page report
  - 11pt font size
  - Single space

3. Final Report

- Final report should include:
  - Project title and student name
  - Introduction Section: give motivation; describe the problem; summarize your contribution
  - Approach: explain the idea; explain proposed approaches
  - Results: show results in form of tables and figures, discuss results
  - Conclusion: summarize the whole project and its outcome
  - Acknowledgements: clearly acknowledge people that helped you finish the project and the web resources you used (please exclude the professor and the TAs)
- 5-page report
  - 11pt font size
  - Single space

Project Timeline

Item	Grade	Due date
Project proposal	15%	March 30, 23:59 (week 1)
Project approval	-	March 31
Start working on the project	-	April 1
Progress report I	10%	April 7, 23:59 (week 2)
Progress report II	15%	April 14, 23:59 (week 3)
Lightning talk	20%	April 23&25 (week 4)
Final report	40%	April 28 23:59 (week 5)