PREDICTING AIRFARE PRICES WITH REGRESSION

JONATHAN FONG, JORDAN MOSAKOWSKI, CONNER YIN

INTRODUCTION

Air Travel is *expensive*

Objective: Predict flight prices using a supervised regression model

Some Questions:

How do connections and plane changes affect airfares?
Which airlines offer the lowest fares?
What is the optimal time from departure time to purchase tickets?

DATASET INFORMATION

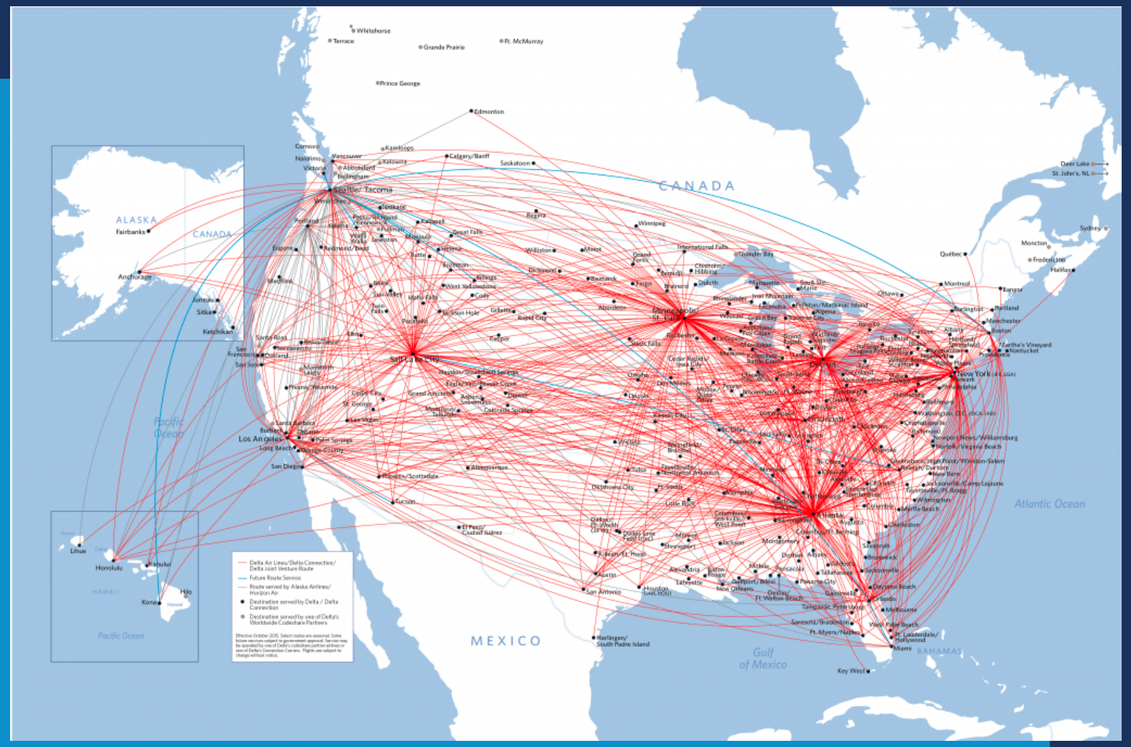
Data taken from Expedia (GitHub dataset)
Target feature is total fare/ticket price

TOTAL SAMPLES
82.1 M

NUMBER OF FEATURES
27

DATA TAKEN FROM
6M FLIGHTS

DATA TAKEN OVER
6 MONTHS



DATASET NOTE

Not every flight is nonstop

Connections offer flexibility, especially for businesses that need to be somewhere at a specific time.

Data Preprocessing

DATA CLEANING

How do we handle **missing values**?

- Extrapolate from other features
  - Calculate travel distance from starting and ending airport
- Populate with arbitrary values e.g. 0
  - Useful for nominal values where only presence matters

DATA TRANSFORMATION

- 12 of the 27 features provide information about a segment/leg of a trip
- Values use || delimiter between segment data
- Transformation process for each segmented feature:
  - Find value with most ||
  - Create new features based on number of ||
  - divide string and assign each string to correct feature

DATA TRANSFORMATION

BOOLEAN VALUES

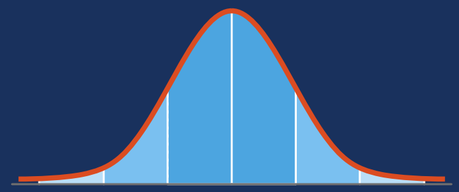
Convert booleans to numerical values (0 for False, 1 for True)

ONE HOT ENCODING

Nominal Features need to be encoded in a 2D binary array

STRING PROCESSING

Ensure single numerical representation e.g. dates to epoch time



STANDARDIZATION

Standardization/Normalization ensure features have same impact on model

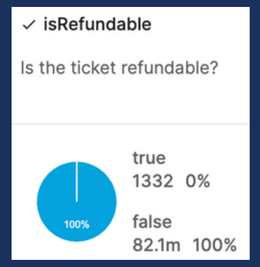
DATA REDUCTION

WHY?
Improve computational efficiency and avoid high dimensionality problems and noise

FEATURE SELECTION

Remove features with little to no effect on the analysis such as those that contain the same information

segmentsDeparture...	segmentsDeparture...
Departure time for each leg (Unix time).	Departure time for each leg (human-legible date format).



FEATURE EXTRACTION

Dimensionality Reduction using PCA
PCA retains intrinsic variability in data and ensures that the most important patterns and relationships are present

Models & Optimization

REGRESSION ALGORITHMS

RIDGE REGRESSION → Addresses multicollinearity by adding L2 penalty term to cost function

ELASTIC NET REGRESSION → Combination of Ridge and LASSO, use L1 and L2 penalties to control sparsity and magnitude in training data

NEURAL NETWORKS → Use Stochastic Gradient Descent (SGD) to fit models in small batches of data

OPTIMIZATION

Utilizing **RMSE** as measure of accuracy

- Meaningful measurement of error in same units as target

Model considered optimized enough when RMSE < **$10**

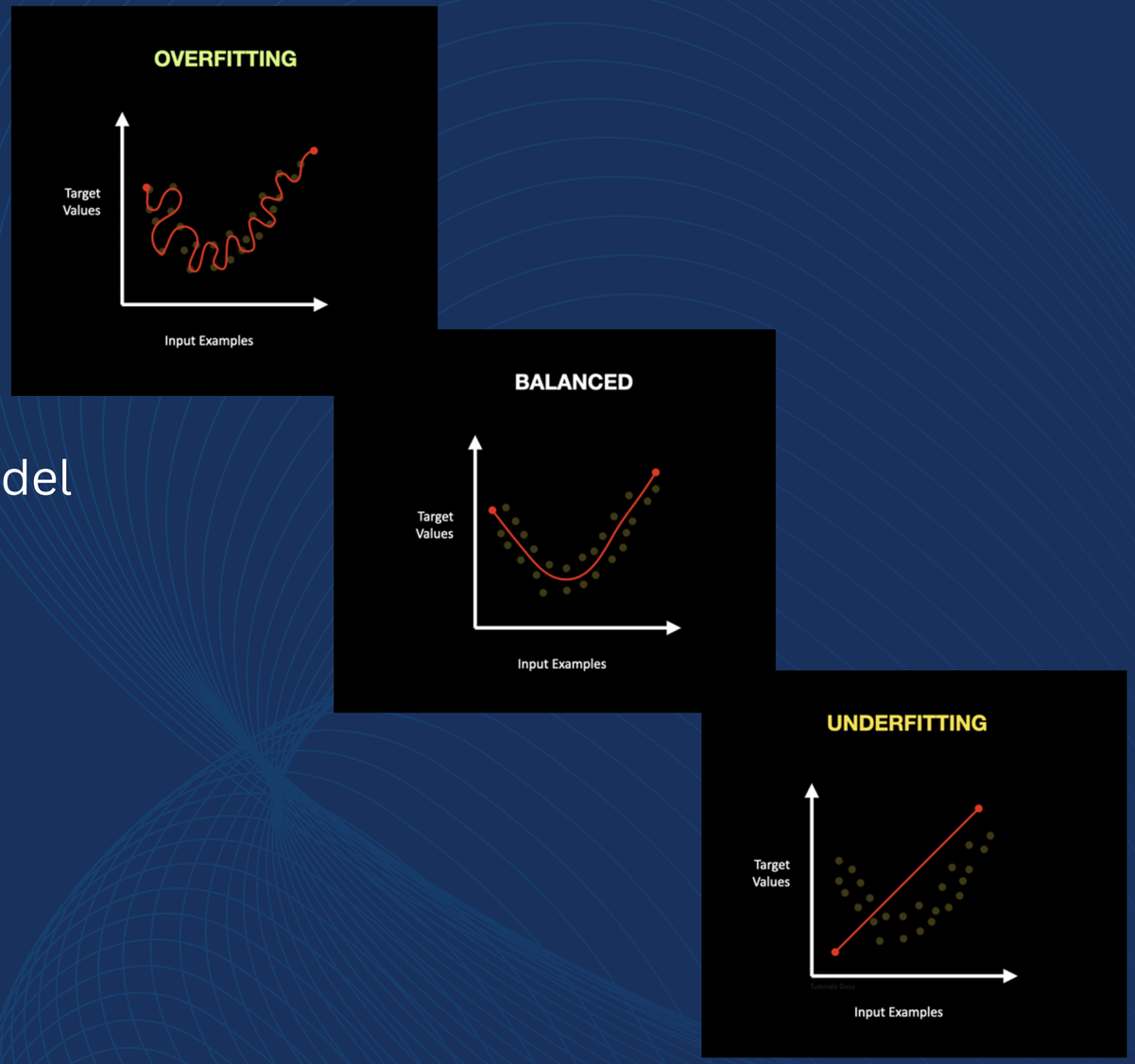
- Low relative to average flight price
- Good enough for extracting trends from the data

OVERFITTING

- Decrease dimensionality using PCA
- Identify features in original dataset to remove
- Increase regularization terms in each model

UNDERFITTING

- Increase dimensionality using PCA
- Reduce or remove regularization terms in each model



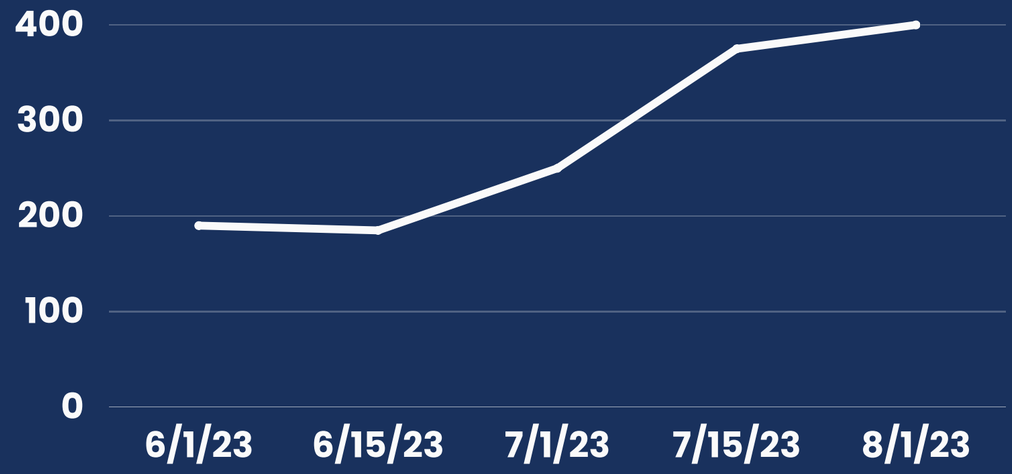
Making Predictions

MAKING PREDICTIONS

Just predicting price on it's own isn't very useful

Examples of data we can extract:

- Price information over time
- Cheaper connecting flights



SUMMARY

- Preprocessing
  - Binary values, string representation, one-hot encoding
- Models
  - Ridge Regression
  - Elastic Net
  - Neural Network
- Optimization
  - PCA
  - L1 and L2 error terms

THANKS

QUESTIONS?