BREAK THROUGH TECH

Machine Learning Foundations

Lab 2

Week of June 2

Today's Agenda

(15 minutes)	Icebreaker
(25 minutes)	Week 2 Concept Overview + Q&A
(20 minutes)	Breakout Groups: Big Picture Questions
(10 minutes)	Class Discussion
(10 minutes)	Break
(35 minutes)	Breakout Groups: Lab Assignment Working
Session 1	
(10 minutes)	Working Session 1 Debrief
(35 minutes)	Breakout Groups: Lab Assignment Working
Session 2	
(10 minutes)	Working Session 2 Debrief
(10 minutes)	Concluding Remarks & Survey

BREAK THROUGH TECH

Icebreaker: “Get Pumped!”

Icebreaker: Get Pumped!

Objectives:

- Share a “pump up song” that inspires, builds your confidence, or otherwise brings you energy

vevo

Icebreaker: Get Pumped! Instructions

- Think about your own “pump up song”. Remember, this song might:
  - Inspire you, give you energy, or confidence
  - Be related to a personal value that is important to you
  - Remind you of a person, place, or thing

- Once you’ve picked your Pump Up Song, add it to the Google doc linked in the chat. The Break Through Tech team will make a playlist with your Pump Up Songs for upcoming Maker Day and other events! https://tinyurl.com/pumpup2024

- When you’re finished, share your Pump Up Song in the chat and why it gives you energy.

Icebreaker: Get Pumped! Share out

- Who wants to share a clip from their “Pump Up” song?
- Why is this song meaningful to you?

BREAK
THROUGH
TECH

Week 2 Concept Overview + Q&A

WEEK 2 CONCEPT OVERVIEW

This week you explored a number of topics. To refresh your memory, your goals were to:

- Build a dataset suitable for ML applications
- Create an appropriate label for supervised learning
- Create features that are suitable for ML applications
- Use exploratory analysis to understand your data
- Clean your data by identifying and fixing common data issues

Build Your Data Matrix

Pandas DataFrame: Common Methods :

- DataFrames
  - df.head(10)
  - Properties:
    - df.shape
    - df.index
    - df.loc[5]
    - df.dtypes – returns type of each column

Sampling:

- Taking a Sample of Data
  - indices = np.random.choice(df.index, size=100, replace=False)
    df_subset = df.loc[indices]
  - df_subset = df.sample(100)

Build Your Data Matrix

- Filtering data
  - condition1 = df['workclass'] == 'Private’ - returns series of True/False values
  - condition2 = df_subset[‘sex_selfID’].isnull() – returns True/False values
  - df_filter = df[condition1 & ~condition2]

- Groups within a column
  - df_subset[‘sex_selfID’] unique() – returns unique values in column
  - counts = df_subset['sex_selfID'].value_counts()
  - df_subset.groupby(['sex_selfID', 'label']).size()

- Modifying/Merging labels
  - condition = columns_not_self_employed & columns_not_null

    df['workclass'] = np.where(condition, 'Not-self-emp', df['workclass'])

    .

Introduction to Feature Engineering

- Cast column to type int
  - df[‘col’] = df[‘col’].astype(int)

- Creating new sample that doesn’t include original sample
  - df_never_sampled = df.drop(labels=df_subset.index, axis=0, inplace=False)
  - df = df.drop([‘col1’, col2’], axis=1)   - drop col1 and col2 from df

- Ordering categorical data

  edu = ['Preschool', '1st-4th', '5th-6th', '7th-8th', '9th', '10th', '1th', '12th', 'HS-grad', 'Prof-school', 'Assoc-acdm', 'Assoc-voc', 'Some-college', 'Bachelors', 'Masters', 'Doctorate']

  df['education'] = pd.Categorical(df['education'], ordered=True, categories=edu)

- Converting categorical data to binary (Feature engineering: one-hot encoding)
  - df_binary = pd.get_dummies(df)

Explore Your Data:
Common Data Exploration Functions

- df.describe() – returns per column statistics about df
- Finding column with highest variance
  - df_summ = df.describe()
  - df_summ.loc['std'].idxmax(axis=1)
  - df_summ.idxmax(axis = 1)['std’]

- Does any column have negative values
  - np.any(df_summ.loc['min'] < 0)
  - np.any(condition)

Explore Your Data:
Visualize Data using Seaborn and Matplotlib

- Histogram, Pairplot, and Barplot
  - sns.histplot(data=df, x="age")
  - sns.pairplot(data=df, hue='label') – pairwise scatterplots of each pair of columns, and color by 'label'
  - sns.pairplot(data=df, kind='kde', corner=True)   - use kernel density estimator type plot
  - sns.barplot(data = df_sub, x='education', y='label')   - shows average of labels for each value of x

- Editing figure
  - plt.figure(figsize=(13,7)) – set width, height in inches
  - plt.ylim(0, 600) – can use this to zoom in on a smaller region of the y axis
  - plt.xticks(rotation=45)

Explore Your Data: Correlation

- df.corr()['label'] – returns correlation of each feature with the label
  - exclude = ['label', 'non_winsorized_label']

    corrs = df.corr()['label'].drop(exclude, axis=0)

- Sort correlations in descending order

  corrs_sorted = corrs.sort_values(ascending=False)

  col_names = corrs_sorted.index - returns column names in descending order of correlation with label.

Clean Your Data: Replacing Outliers

- Value corresponding to x percentile.
  - val = np.percentile(df[‘col_name’], x)

- Scipy to winsorize data (remove outliers)
  - scipy.stats as stats
  - df[‘col-win’] = stats.mstats.winsorize(df[‘col’], limits=[0.01, 0.01]) – replace lower and top 1% with values at 1% and 99% respectively.

- zscore = (value – mean)/std - measures how far away each point is from the mean – zscore of 1 means point is 1 std away from mean.
  - zscores = stats.zscore(df[‘col’])

Clean Your Data: Replacing Missing Values

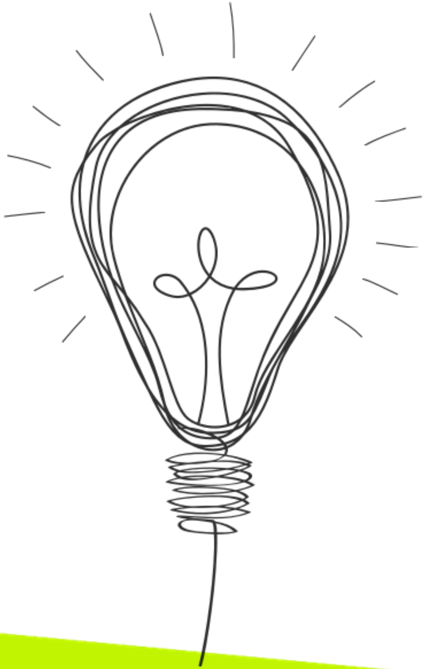
- Find and count missing values
  - df.isnull()
  - nan_count = np.sum(df.isnull())

- Replace missing values with mean

```
mean_ages=df['age'].mean()
df['age'].fillna(value=mean_ages, inplace=True)
```

Questions & Answers

What questions do you have about the online content this week?



BREAK
THROUGH
TECH

Breakout Groups: Big Picture Questions

Big Picture Questions

You have 20 minutes to discuss the following questions within your breakout groups:

1. Why is data preparation so important to the machine learning development process?
2. Considering that data preparation often takes the majority of model development time, how would you communicate to stakeholders (bosses, product managers, leadership, etc.) why you need to budget time for data preparation?
3. What does it mean to have a “modeling dataset”?
4. What is the difference between nominal data and ordinal data? Explain with an example.
5. Why is data visualization an important part of the data preparation process?
6. Name a few libraries used for data analysis and visualization and explain when you would use each library.

Class Discussion

Break

BREAK THROUGH TECH

BREAK
THROUGH
TECH

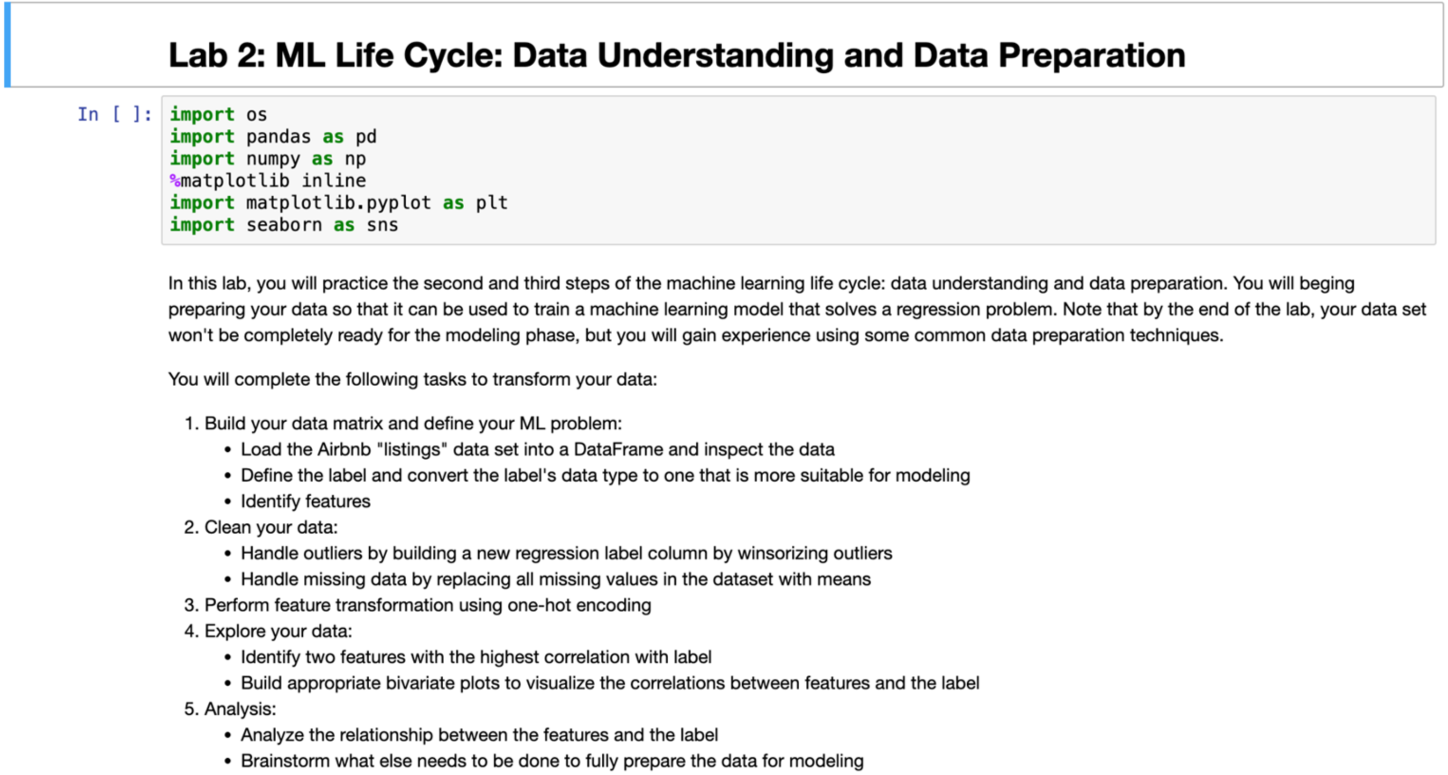
Breakout Groups: Lab Assignment

Lab 2

In this lab, you will:

- Build your data matrix
  - Identify the features and the label in your dataset
- Clean your data by winsorizing outliers and replacing missing values with means
- Perform feature engineering:
  - Perform feature transformations using one-hot encoding
  - Perform exploratory data analysis to find relationships between the features and the label in preparation for feature selection

Lab 2



Lab 2: ML Life Cycle: Data Understanding and Data Preparation

In [ ]:

```
import os
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
```

In this lab, you will practice the second and third steps of the machine learning life cycle: data understanding and data preparation. You will beging preparing your data so that it can be used to train a machine learning model that solves a regression problem. Note that by the end of the lab, your data set won't be completely ready for the modeling phase, but you will gain experience using some common data preparation techniques.

You will complete the following tasks to transform your data:

1. Build your data matrix and define your ML problem:
   - Load the Airbnb "listings" data set into a DataFrame and inspect the data
   - Define the label and convert the label's data type to one that is more suitable for modeling
   - Identify features
2. Clean your data:
   - Handle outliers by building a new regression label column by winsorizing outliers
   - Handle missing data by replacing all missing values in the dataset with means
3. Perform feature transformation using one-hot encoding
4. Explore your data:
   - Identify two features with the highest correlation with label
   - Build appropriate bivariate plots to visualize the correlations between features and the label
5. Analysis:
   - Analyze the relationship between the features and the label
   - Brainstorm what else needs to be done to fully prepare the data for modeling

BREAK
THROUGH
TECH

Working Session 1 Debrief

Lab Debrief

So far,

- What did you enjoy about this lab?
- What did you find difficult about this lab?

BREAK
THROUGH
TECH

Working Session 2 Debrief

Lab Debrief

- What did you enjoy about this lab?
- What did you find hard about this lab?
- What questions do you still have about this lab?
- How did you approach problem-solving during the exercise?
- What would you do differently if you were to repeat the exercise?

BREAK
THROUGH
TECH

Concluding Remarks

Concluding Remarks

- Key takeaways
- Additional resources

Next week

In the following week, you will:

- Define the core foundational elements of model training and evaluation
- Develop intuition for different classes of algorithms
- Analyze the mechanics of two popular supervised learning algorithms: decision trees and k-nearest neighbors
- Develop intuition on tradeoffs between different algorithmic choices

And in the lab, you will:

- Define your ML problem and build your DataFrame
- Prepare your data:
  - Perform feature engineering by converting categorical features to one-hot encoded values
- Train multiple decision trees and evaluate their performances:
  - Train decision tree classifiers with various hyperparameter values.
  - Visualize and evaluate the accuracy of the models' predictions
- Train multiple KNN classifiers and evaluate their performances:
  - Train KNN classifiers with various hyperparameter values.
  - Visualize and evaluate the accuracy of the models' predictions
- Determine the best performing model for your predictive problem

BREAK
THROUGH
TECH

Content + Lab Feedback Survey

Content + Lab Feedback Survey

To complete your lab, please answer the following questions about BOTH your online modules and your lab experience. Your input will help pay it forward to the Break Through Tech student community by enabling us to continuously improve the learning experience that we provide to our community.

Thank you for your thoughtful feedback!

https://forms.gle/eUQQZgS6BPRpqgZ7A

Important concepts for next week

- Generalization
- Underfitting and overfitting
- Model complexity
- Entropy / information gain