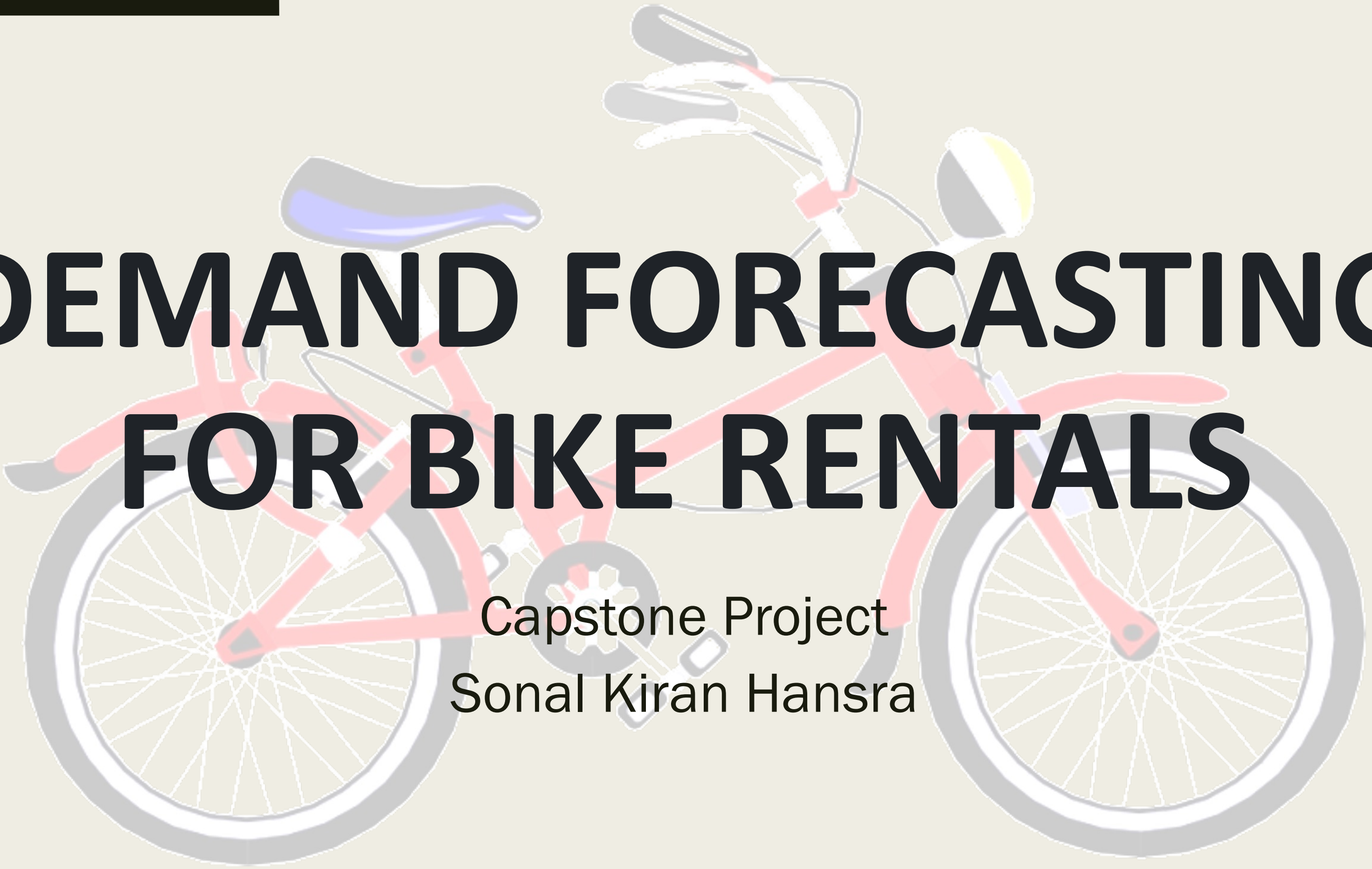

DEMAND FORECASTING FOR BIKE RENTALS

Capstone Project

Sonal Kiran Hansra

Project Overview

The Challenge –

- Growing **awareness** of **the importance of exercise** alongside the increasing recognition of the **impacts of global warming** has prompted a **notable surge in individuals opting for bicycle commutes** whenever feasible
- However, **not everyone** willing to partake in bicycle commuting **either owns or desires to own a bicycle**. This has led to the **creation of several bike-sharing enterprises**, facilitating convenient and affordable bike rentals for users.
- Since the bicycle commuting culture is still evolving, **accurately estimating the demand for bikes** on any given day presents a **formidable challenge** for bike-sharing companies.

Project Overview

The Solution –

- The idea is to **leverage machine learning** to build **demand forecast models** that are able to predict bike rental demands at a day level with a reasonable accuracy, while also doing a **comparative analysis of different models**.

Project Overview

The Impact –

- By **accurately predicting the demand** for bikes on any given day, bike-sharing companies can **optimize** their **operations**, ensuring an adequate supply of bicycles at high-demand locations and periods
- This will help eliminate the frustration of potential riders facing unavailability, thus **enhancing overall user experience and satisfaction**.
- Ultimately, **increased bike availability** will serve as a **catalyst for the growing trend of bicycle commuting**, advancing the global movement towards **sustainability**.

Data Introduction

For this project we have collected data from **3 different sources** –

- **Bike rental data** from Capital Bike Share
- **Weather data** from NOAA's National Climatic Data Center
- **Holidays data** from DC Department of Human Resources

The data spans over 11.5 years, from Jan-2012 till Jun-2023, and is split across multiple files. Let's look at these files to ascertain the best method to combine these files and create a single dataset.

Data Introduction

This is what each variable in our dataset means:

■ **Capital Bikeshare Data**

- **ride_id -** includes ID number of the ride
- **rideable_type -** indicates whether the type of bike was "classic", "docked" or "electric"
- **started_at -** includes start date and time
- **ended_at -** includes end date and time
- **start_station_name -** includes starting station name
- **start_station_id -** includes starting station number
- **end_station_name -** includes ending station name
- **end_station_id -** includes ending station number
- **start_lat -** includes starting station latitude
- **start_lng -** includes starting station longitude
- **end_lat -** includes ending station latitude
- **end_lng -** includes ending station longitude
- **member_casual -** indicates whether user was a "registered" member or a "casual" rider

Data Introduction

This is what each variable in our dataset means:

- **Weather Data**
  - **station** - station ID
  - **name** - name of weather station
  - **date** - date of obseravtion
  - **avg_wind_speed** - average wind speed
  - **num_days_multiday_prcp** - number of days included in the multiday precipitation total
  - **multiday_prcp** - multiday precipitation total
  - **peak_gust_time** - peak gust time
  - **prcp** - precipitation
  - **snowfall** - snowfall
  - **snowdepth** - snow depth
  - **temp_avg** - average temperature
  - **temp_max** - maximum temperature
  - **temp_min** - minimum temperature

Data Introduction

This is what each variable in our dataset means:

- **Weather Data**
- **temp_obs -** temperature at the time of observation
- **dir_fastest_2min_wind -** direction of fastest 2-minute wind
- **dir_fastest_5min_wind -** direction of fastest 5-minute wind
- **speed_fastest_2min_wind -** fastest 2-minute wind speed
- **speed_fastest_5min_wind -** fastest 5-minute wind speed
- **wt: weather type**
  - we have 17 columns for different weather types

Data Introduction

This is what each variable in our dataset means:

- **Holiday Data**
  - **date** - date
  - **weekend** - indicates whether it was a weekend
  - **holiday** - indicates whether it was a holiday
  - **weekend_holiday** - indicates whether it was either a weekend or a holiday

Findings From Data Preparation

- This **bike-sharing data** contains **~35M rows** and is takes about **5.3GB of space**! It is time and memory intensive to work with such huge files.
- Since the **demand forecast model** we aim to build is at the **day level**, we can **roll-up our data** and **reduce** the number of **rows**. Additionally, there are **several irrelevant columns**. We will only require the start date and member type columns.
  - *There were some unknowns in this data but they make only 0.0001% of the data*
  - *There are 4 missing dates in our dataset. We will have to add these dates and impute these new rows appropriately when we move on with further analysis.*

- The **weather data** has observations from **3 stations** informing the weather data for any given day. We have to figure out the best way to merge data from the 3 different stations. We also need to change the data type of the \`date\` column from 'object' to 'datetime
- There is a lot of missing data! 20 columns have almost 99% data missing.
- The **holiday data** is clean and has no duplicates.

Next Steps

- Implement all action items identified in the data preparation
- Merge datasets to create a single data source for modelling
- Perform intensive EDA, impute missing values and de-duplicate the data as required
- Create dummy variables
- Calculate baseline accuracy based on naïve forecasting method

Thank You!