First steps of the project

Project E8, KAGGLE-BIKE-SHARING-DEMAND

Repository link: https://github.com/Scarch/ProjectE8

Title: Analysing and predicting bike rental demand based on weather, date and time
Team members: Sten-Egert Märtson, Kaur Kivilaan

Task 2. Business understanding

Identifying your business goals

Background

Bike rental has become more and more relevant throughout the last couple decades. Bike renting systems offer people an opportunity to travel with a bike from one destination to another but without the need to worry about taking care of and storing their own bike. It also saves some money for the people who only need to use bikes occasionally. Another advantage of bike renting is that it allows individuals to make one way trips, for example, when a person wants to go to work by bus because it's cold in the mornings but come back to home by bike to enjoy the fresh air.

Usually bike rental systems are automated, which means it's possible to gather data about the usage habits of the bike rental users by gathering data about rental times and locations. For example we could hypothesise that people tend to rent more bikes during the morning at around 7.30-9.00 a.m. when their work starts and in the evening at 4.30 - 6.00 p.m. when their work ends. Also we could examine how weather or holidays affect the habits of users. Because of the abundant amounts of data these systems create this field naturally becomes a target for analysis.

Business goals

A broader goal is to help those who offer and maintain bike rental services. We hope to provide insight into when bike demand is at its highest and lowest, from which one could

derive many things, such as the best time for maintenance or down-time. Therefore our hope is also that people can comfortably rent bikes and be assured that the entity they are renting from has taken the proper measures to ensure the bike is suitable for rental.

Another goal would be to help the cities in which bike rental services are offered. We can find in which times and locations people move, from which we can deduce where bike infrastructure needs to be the most robust. From the same information we can surmise when and where bike availability should be at its highest.

Business success criteria

Our success relies on whether we can make a model that could successfully predict the bike rental demand based on date, time and weather. We will consider our project a success under two conditions. The first condition is objective - attain a score of around 0.4 in the Kaggle competition Bike Sharing Demand (smaller score means better model, first place in the competition got a score of 0.33756). Our second condition is subjective - will we be able to understand and make conclusions on the usage habits of the bike rental users.

Assessing your situation

Inventory of resources

- Test and training data provided by Kaggle
    - Contains hourly rental data spanning two years (2011-2013) provided by Capital Bikeshare
    - Features like temperature, humidity and whether an hour fell on a holiday or workday.
- Public Capital Bikeshare data
    - Contains data for each ride from September of 2010 to October of 2024 (updated each month)
    - Features such as member status, start and end times/stations and ride duration.
- Project members (Sten-Egert Märtson, Kaur Kivilaan)
- Instructors of the Tartu University Introduction to Data Science course
- 2 laptops and 2 desktop PCs
- JetBrains educational licenses

Requirements, assumptions, and constraints

- Report needs to be submitted by Monday, Dec 2, at noon (12:00)
- Poster submission deadline is Monday, Dec 9, at noon (12:00)
- Overall deadline for the project is unclear (possibly 13th of December)

Risks and contingencies

- Exams and projects in other university courses may leave little time for this project
- As there is no concrete deadline for the project, we cannot be sure if we can do everything we want to do

Terminology

Bike rental / bicycle-sharing system - a system that allows users to rent a bike from one destination and return the bike to a rental system in another area. Usually those systems keep track of the date, time and count of users that are currently using the bikes. The

Costs and benefits

Cost: Time, electricity
Benefits: Could save money by maintaining less bikes during times of the year where bike rental isn't that much used. Could make more money by raising bike availability where demand is high.

Defining our data-mining goals

Data-mining goals

We hope to have a good model (or even multiple) that is capable of predicting the total count of bikes rented during each hour covered by the Kaggle competition test set. We also wish to gain non-trivial knowledge pertaining to bike rental and the habits of people who rent said bikes.

Data-mining success criteria

In quantitative terms, we will consider the project a success, if we reach a score of at most 0.4 by evaluation through root mean squared logarithmic error (RMSLE) with the Kaggle test set.

Evaluation formula and explanation is given here: www.kaggle.com/competitions/bike-sharing-demand/overview/evaluation.

Task 3. Data understanding

Gathering Data

Outlining data requirements

It's important to have data that shows us the amount of bikes rented during a certain time window. The time window should be precise enough so that we could make assumptions based on weather and current time. A good time window for that could be 1 hour. Also we need to either have data about the current weather of the time window or the location from where the bike was rented / ridden to (so we could at least manually scrape weather information from the web).

Verifying data availability

The Kaggle training and test data ranges from the first of January of 2011 to the first of January 2013. The data originates from Capital Bikeshare which operates in Washington D.C. This dataset has the amount of bike rented and the weather conditions divided into hourly instances. It also includes the dates, times, and whether the date is considered a working day, weekend or holiday.

We also have additional data from Capital Bikeshare data in the time span from 2010 to 2024 which has information about the starting locations and the destination locations of the bike rides in said time span, but it doesn't include data about the weather, which means we need to scrape the weather information ourselves (there are multiple free weather APIs on the internet).

Defining selection criteria

Our data sources:

- Kaggle's Bike Sharing demand competition
    - https://www.kaggle.com/competitions/bike-sharing-demand/data
    - Training will be done on train.csv
    - Testing will be done with test.csv
- Capital Bikeshare system data
    - https://capitalbikeshare.com/system-data

- ZIP files (e.g. 201810-capitalbikeshare-tripdata.zip) which contain csv files with the same names but with the relevant file extension

Describing data

Kaggle data fields i.e. necessary features for model creation

- **datetime** - hourly date + timestamp (*year-month-day hour*:*minute*:*second*)
- **season** - 1 = spring, 2 = summer, 3 = fall, 4 = winter
- **holiday** - whether the day is considered a holiday (1 if yes, 0 if not)
- **workingday** - whether the day is neither a weekend nor holiday (1 if neither, 0 if at least one of the conditions is true)
- **weather**
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog
- **temp** - temperature in Celsius
- **atemp** - "feels like" temperature in Celsius
- **humidity** - relative humidity from 0-100 in percentages
- **windspeed** - wind speed in kilometers per hour
- **count** - number of total rentals
- **casual** - number of non-registered user rentals initiated
- **registered** - number of registered user rentals initiated

Fields **count**, **casual** and **registered** are only given in the training dataset (train.csv) as they are the properties we are trying to predict. There are 10888 instances/rows in the training data and 6495 instances/rows in the test data (test.csv).

Capital Bikeshare system data fields

- **Duration** – duration of trip in seconds
- **Start date** – includes start date and time (*year-month-day hour*:*minute*:*second*)
- **End date** – includes end date and time (*year-month-day hour*:*minute*:*second*)
- **Start station number** – indicates starting station number (e.g. 31104)
- **Start station** - indicates starting starting station name (e.g. “Adams Mill & Columbia Rd NW”)
- **End station number** – indicates ending station number (e.g. 31108)
- **End station** - indicates ending station name (e.g. “4th & M St SW”)
- **Bike number** – includes ID number of bike used for the trip (e.g. W23481)
- **Member type** – indicates whether user was a "registered" member or a "casual"

- Member passes: Annual Member, 30-Day Member, Day Key Member
- Casual passes: Single Trip, 24-Hour Pass, 3-Day Pass, 5-Day Pass

At the time of creating this report (November 2024), there are 90 different zip files relating to the time span 2010-2024. Data for the years 2010-2017 are all given in one zip file, whereas the years after 2017 have their data split into monthly sections.

Exploring data

Kaggle training data set:

	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
count	10886.000000	10886.000000	10886.000000	10886.000000	10886.00000	10886.000000	10886.000000	10886.000000	10886.000000	10886.000000	10886.000000
mean	2.506614	0.028569	0.680875	1.418427	20.23086	23.655084	61.886460	12.799395	36.021955	155.552177	191.574132
std	1.116174	0.166599	0.466159	0.633839	7.79159	8.474601	19.245033	8.164537	49.960477	151.039033	181.144454
min	1.000000	0.000000	0.000000	1.000000	0.82000	0.760000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	2.000000	0.000000	0.000000	1.000000	13.94000	16.665000	47.000000	7.001500	4.000000	36.000000	42.000000
50%	3.000000	0.000000	1.000000	1.000000	20.50000	24.240000	62.000000	12.998000	17.000000	118.000000	145.000000
75%	4.000000	0.000000	1.000000	2.000000	26.24000	31.060000	77.000000	16.997900	49.000000	222.000000	284.000000
max	4.000000	1.000000	1.000000	4.000000	41.00000	45.455000	100.000000	56.996900	367.000000	886.000000	977.000000

Kaggle training and test datasets merged (“casual”, “registered” and “count” columns dropped):

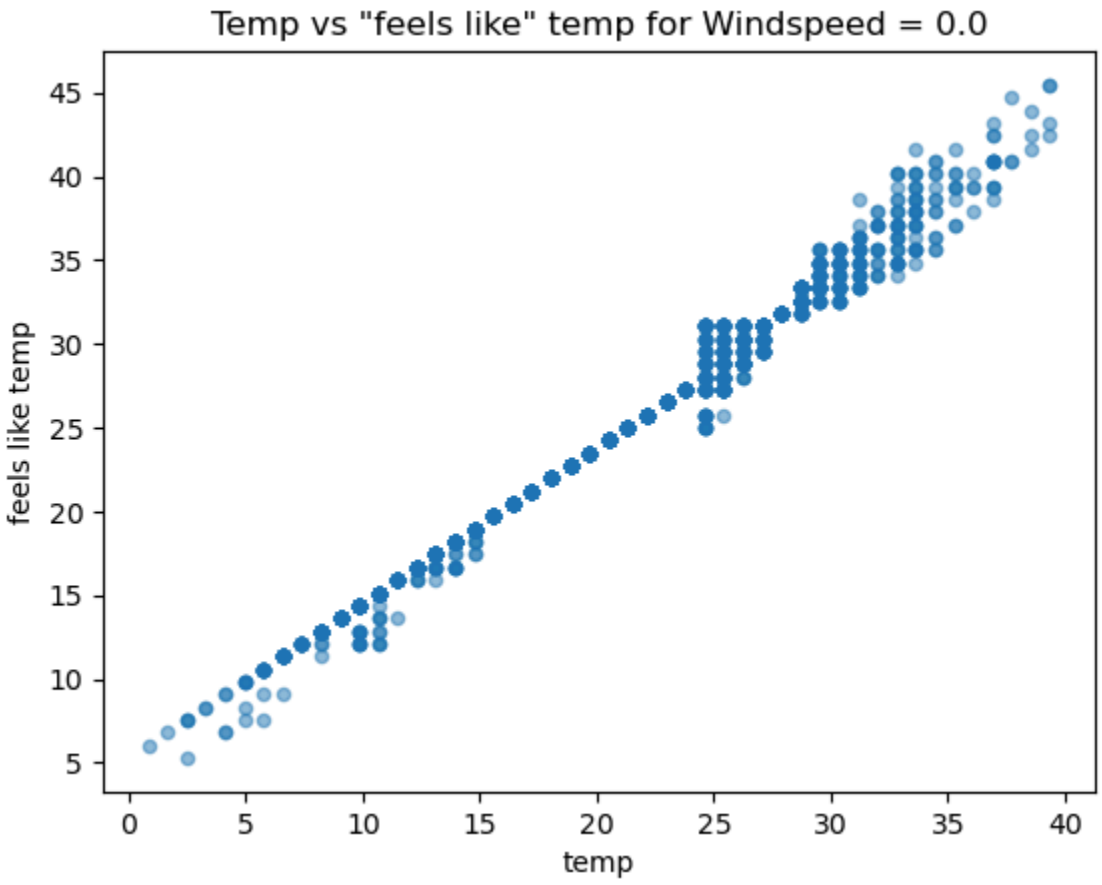
	season	holiday	workingday	weather	temp	atemp	humidity	windspeed
count	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	2.501640	0.028770	0.682721	1.425283	20.376474	23.788755	62.722884	12.736540
std	1.106918	0.167165	0.465431	0.639357	7.894801	8.592511	19.292983	8.196795
min	1.000000	0.000000	0.000000	1.000000	0.820000	0.000000	0.000000	0.000000
25%	2.000000	0.000000	0.000000	1.000000	13.940000	16.665000	48.000000	7.001500
50%	3.000000	0.000000	1.000000	1.000000	20.500000	24.240000	63.000000	12.998000
75%	3.000000	0.000000	1.000000	2.000000	27.060000	31.060000	78.000000	16.997900
max	4.000000	1.000000	1.000000	4.000000	41.000000	50.000000	100.000000	56.996900

Data in the Kaggle dataset seems to be mostly logical, except for a few anomalies in weather reports. The bike loan counts make sense, the overall amount of bike rentals in an hour varies from 1 to 977, where 25th percentile is 42, 50th is 245 and 75th is 284. The mean of bike rentals is around 192. The “casual” and “registered” user counts add up to the total “count” column in every row. There’s no anomalies in the “season”, “holiday”, “workingday” and “weather” columns - each of them has only the values specified in the documentation. The temperature (“temp”) column ranges from 0.82 to 41 degrees celsius. The “feels like” temperature (“atemp”) column ranges from 0 to 50 degrees celsius.

Although the temperature ranges seem alright, there are some anomalies in their differences, where in some extreme cases their difference is over 20 degrees, which doesn't seem logical.

The range of humidity values is from 0% to 100% - values of 100 make sense when compared to other weather conditions (these values mostly occur when it's a hotter day, when it's currently raining or when there's mist). Unlike values with a humidity of 100%, 0% humidity don't make sense, because they were all recorded in the same spring day, when it was mostly raining and the temperature varied between 13 to 19 degrees celsius. Those values should probably be dropped or replaced by correct values.

The range of wind speed ("windspeed") values varies between 0 to 57 km/h. Although the values where wind speed of 0 may seem a little bit atypical, when compared to other weather values, they made sense. For example, the wind speed was never 0 with "weather" column value being 4 (Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog) and every time the wind speed was 0, the temperature and "feels like" temperature values weren't far apart.



Verifying data quality

The data offered by Kaggle needs a little bit more examination in the columns describing weather, but overall, most of the data seems to be of high quality.

Capital Bikeshare states “The data [Capital Bikeshare system data] has been processed to remove trips that are taken by staff as they service and inspect the system, trips that are taken to/from any of our “test” stations at our warehouses and any trips lasting less than 60 seconds (potentially false starts or users trying to re-dock a bike to ensure it's secure).” Therefore we can be fairly sure that Capital Bikeshare has done their due diligence to maintain the quality of the data

Task 4. Planning our project

Tasks

1. **Data analysis (~12 hours, ~7 hours Kaur, ~5 hours Sten-Egert)**
   Should do further data analysis on data quality, whether some data needs to be deleted or replaced. Should also make initial conclusions about the data, such as finding out the times when the rental service seems to be the most occupied and when there are the least amount of clients. Additionally we should try to find some correlations between the weather and the amount of bikes rented. Involves creation of different plots/visualisations that describe the data.
2. **Data cleanup (~4 hours distributed evenly)**
   Data that was found to be of bad quality during the data analysis should be either deleted or replaced by better data. We have already found some issues with some features relating to weather. We will also try training the model on the data that isn't cleaned up as preemptive manipulation of data could possibly lose us some accuracy.
3. **Adding/scraping additional data (weather) (time depends on API limits, ~6 hours, ~4 hours Sten-Egert, ~2 hours Kaur)**
   a. Involves using weather APIs
      For the data that doesn't have weather features (Capital Bikeshare system data) we should add the data by using either an API or web scraper. Of course this data should also be cleaned up.
4. **Testing different models (~16 hours, distributed evenly)**
   a. Random forest
   b. Lasso and ridge regression
   c. Neural network
      After we have cleaned up and analysed the data, we should start testing and assessing different machine learning models. This also involves hyperparameter tuning.
5. **Analyse the (working) machine learning model(s) (~4 hours, distributed evenly)**
   After we have found a suitable machine learning model (or multiple), we should analyse it (if possible). Key point of analysis is to find out how different features affect how many people rent a bike.
6. **Create a poster (~8 hours distributed evenly)**
   After we have found and analysed a machine learning model that works, we should report our findings on a poster.

Methods and tools

- scikit-learn library
- cuML library suite
- Tensorflow
- Plotting libraries
    - seaborn
    - Matplotlib
- JetBrains PyCharm (educational license)
- Jupyter Notebook
- Python 3.x
- Weather API
    - https://open-meteo.com/
    - https://www.weather.gov/documentation/services-web-api