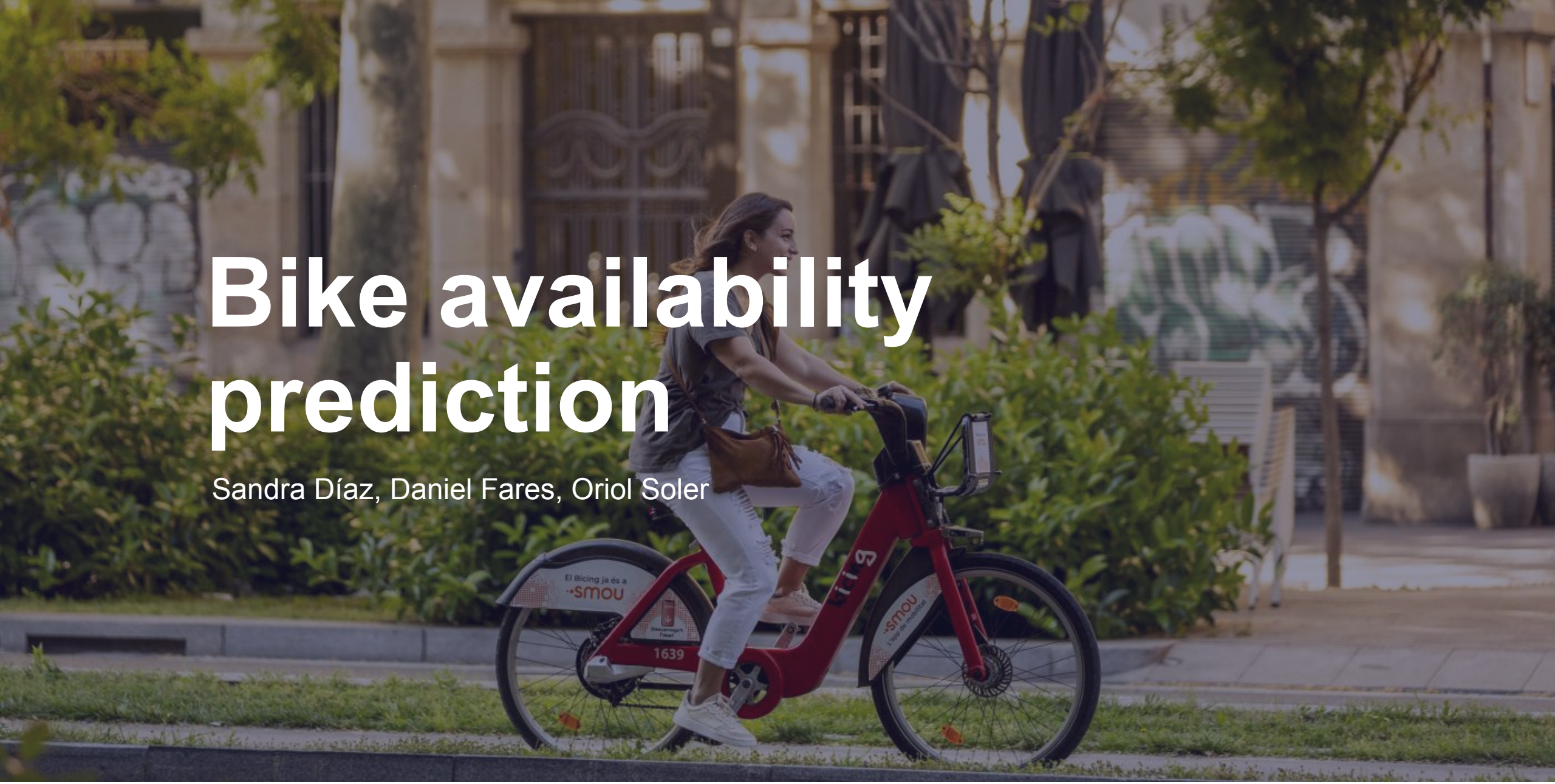


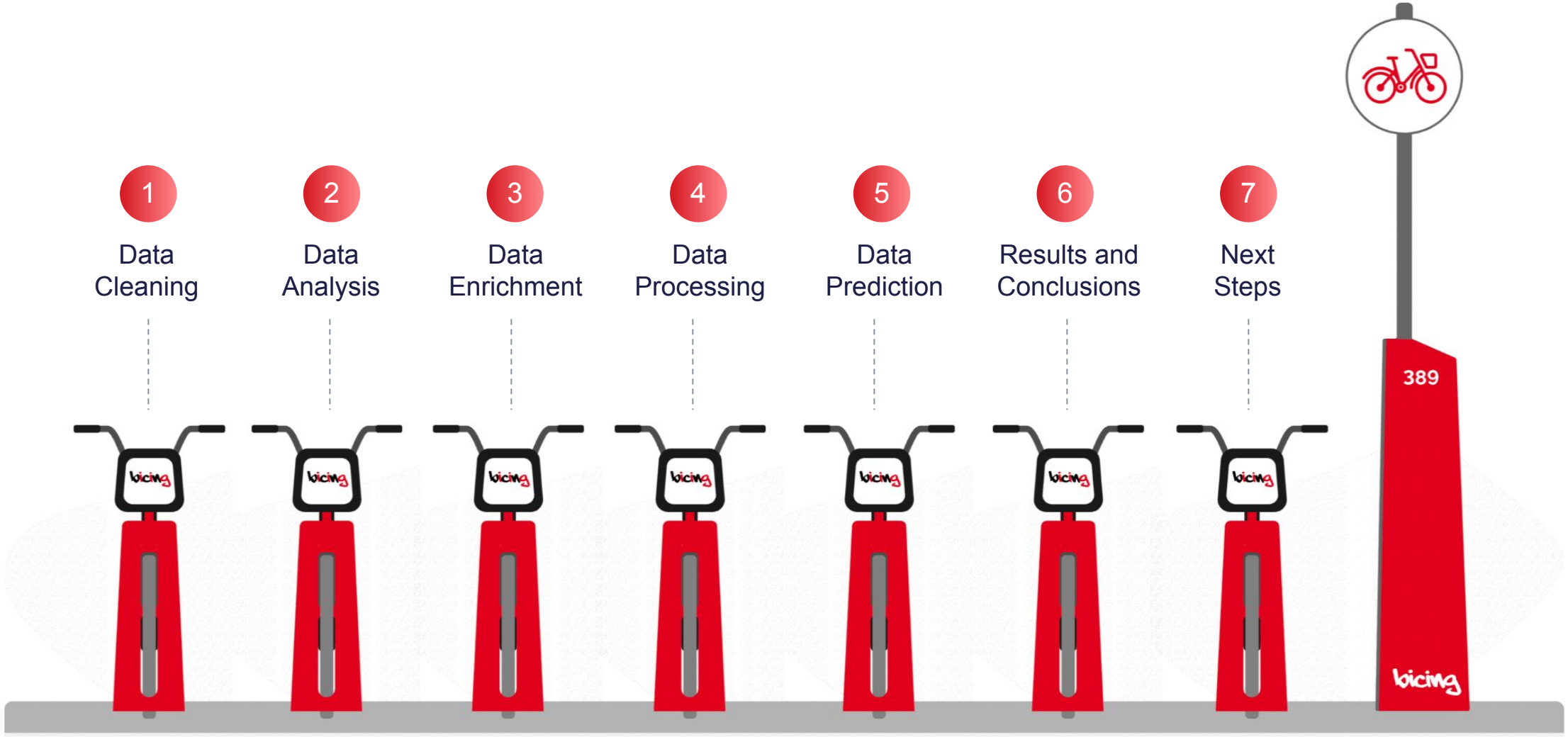
Bike availability prediction

Sandra Díaz, Daniel Fares, Oriol Soler

UNIVERSITAT DE BARCELONA X bicing

Capstone Path

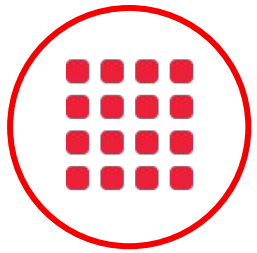
UNIVERSITAT DE BARCELONA x bicing



1. Data preparation

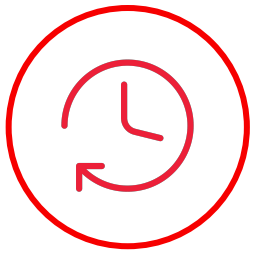
Goal: to explore and understand the data

How did the usage of the bicycles evolve during the analysed period?



Yearly
bike volume

Monthly
bike usage



Weekly
bike routine

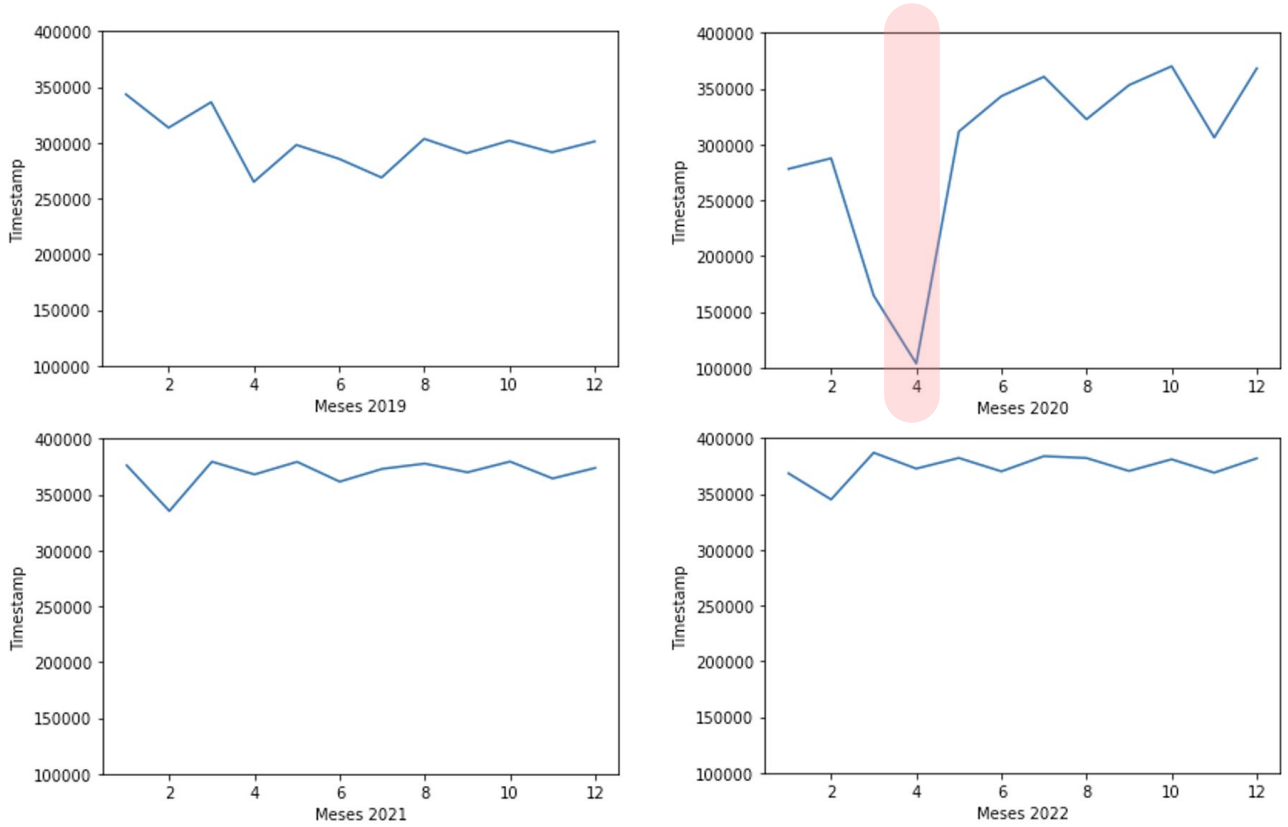
1. Data preparation

UNIVERSITAT DE BARCELONA x bicing

Goal: to explore and understand the data

Yearly
bike volume

Tendencia creciente en cuanto al uso de las bicicletas desde 2021. Previamente, hay una caída pronunciada en 2020 a causa del Covid.
Los datos del 2019 se han completado manualmente y ese puede ser el motivo de que la tendencia decreciente de principios de año del 2021 en adelante no aparezca.

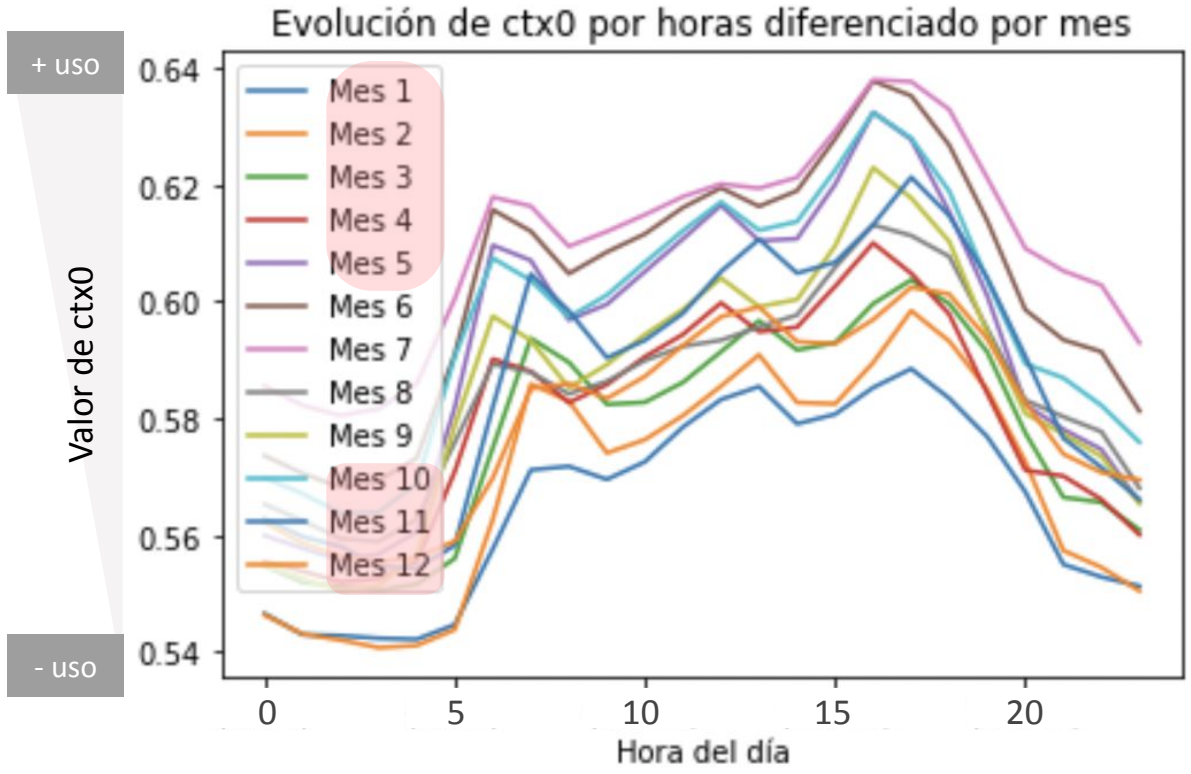


1. Data preparation

Goal: to explore and understand the data

Monthly bike usage

La disponibilidad de anclajes según su capacidad es mayor en los meses que van de diciembre a mayo. Esto lo asociamos a un mayor uso de las bicicletas en este periodo. Una posible hipótesis es el tiempo: en los meses más calurosos el comportamiento varía.

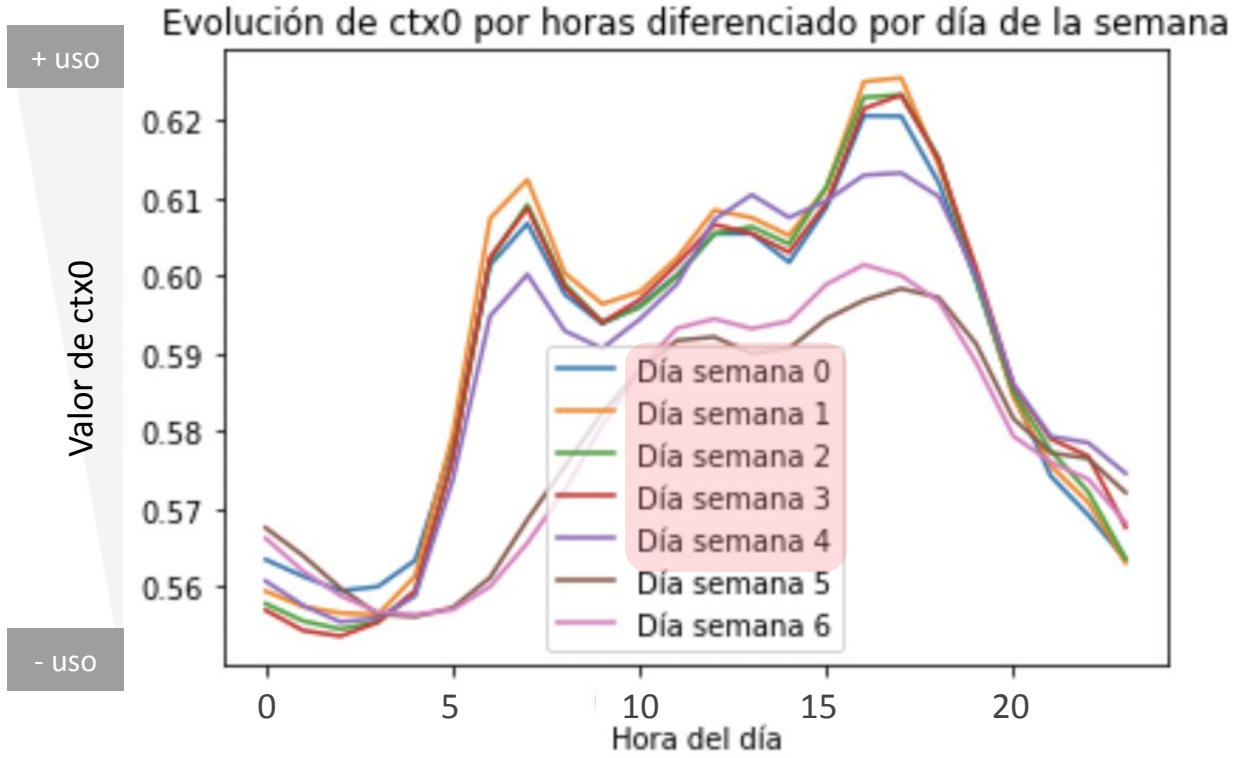


1. Data preparation

Goal: to explore and understand the data

Weekly
bike routine

El uso de las bicicletas es mayor los días laborales. Además, esto encaja con las horas: se localizan dos picos claros alrededor de las 8/9h de la mañana y a partir de las 18h.



1. Data preparation

Goal: to explore and understand the data

Did the mechanical bicycles and the electric ones behaved similarly?

Mechanical
bikes

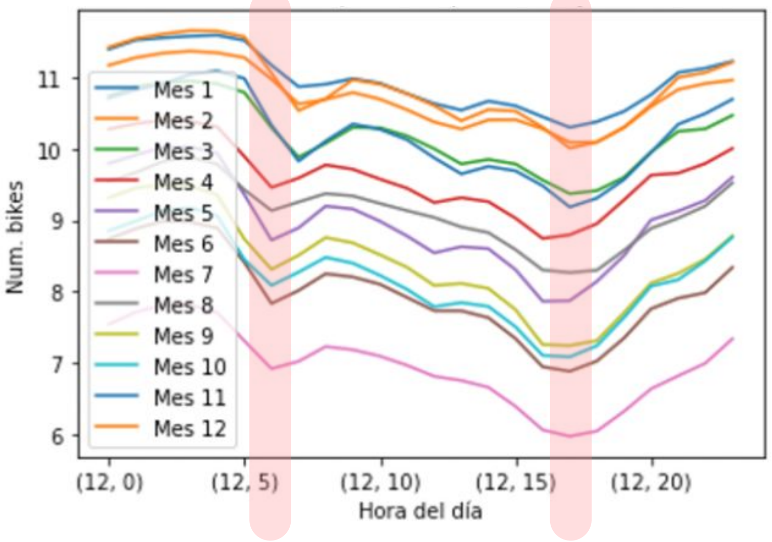
Electric
bikes

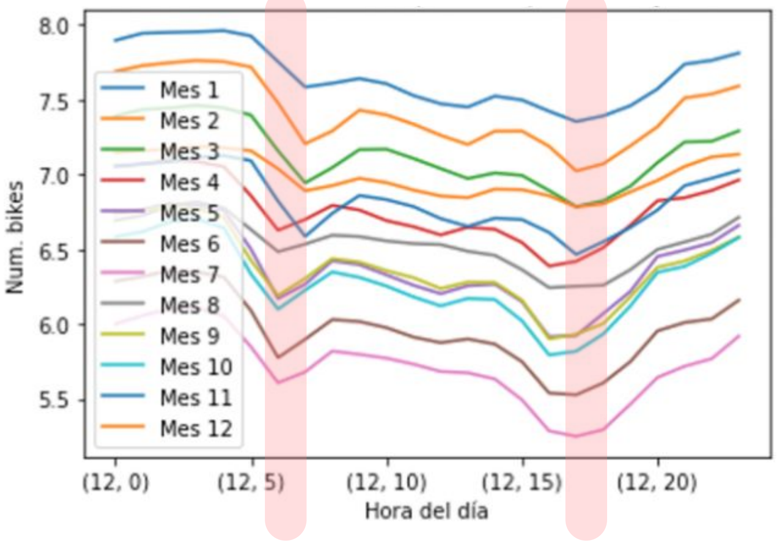
UNIVERSITAT DE BARCELONA x bicing

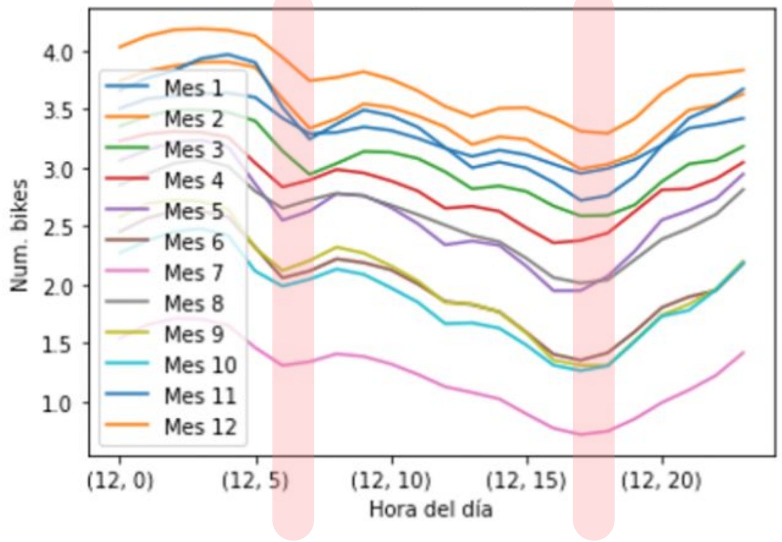
1. Data preparation

Goal: to explore and understand the data

Bikes availability per month and hour





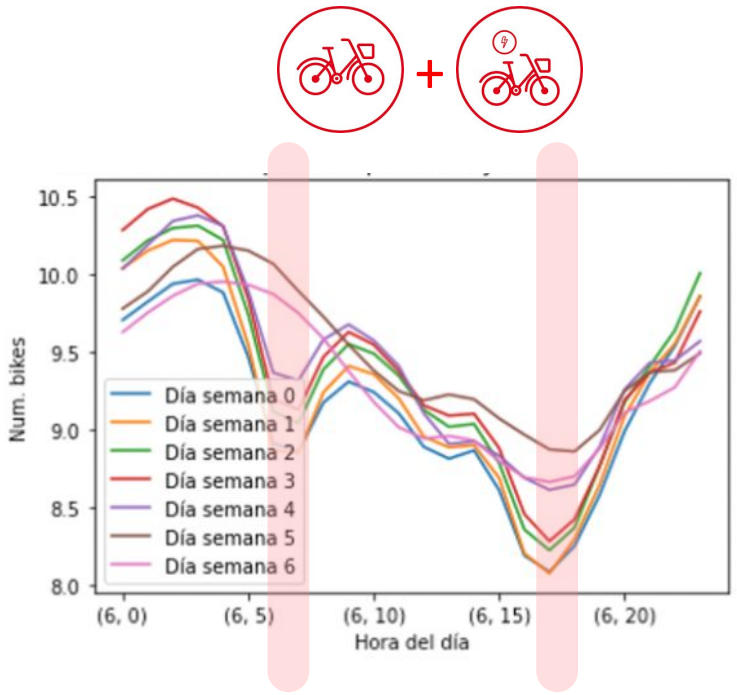


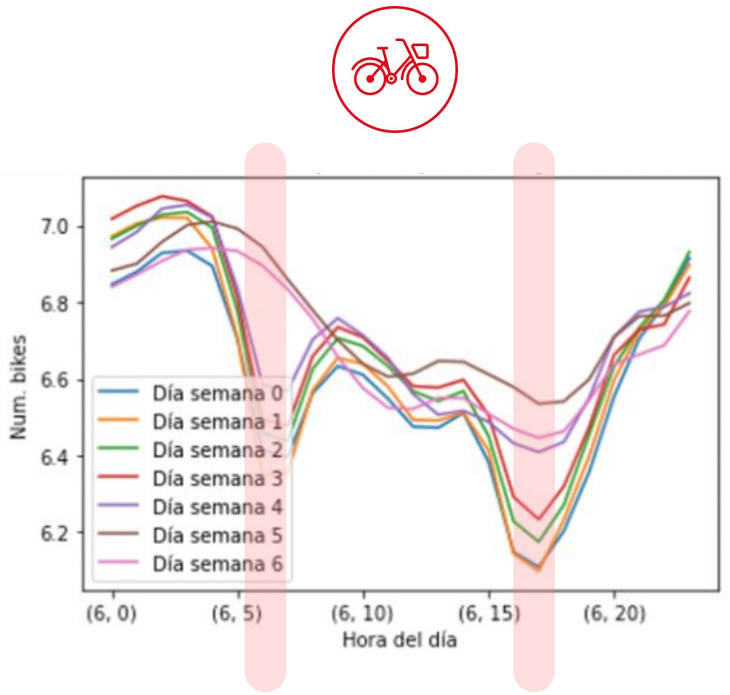
1. Data preparation

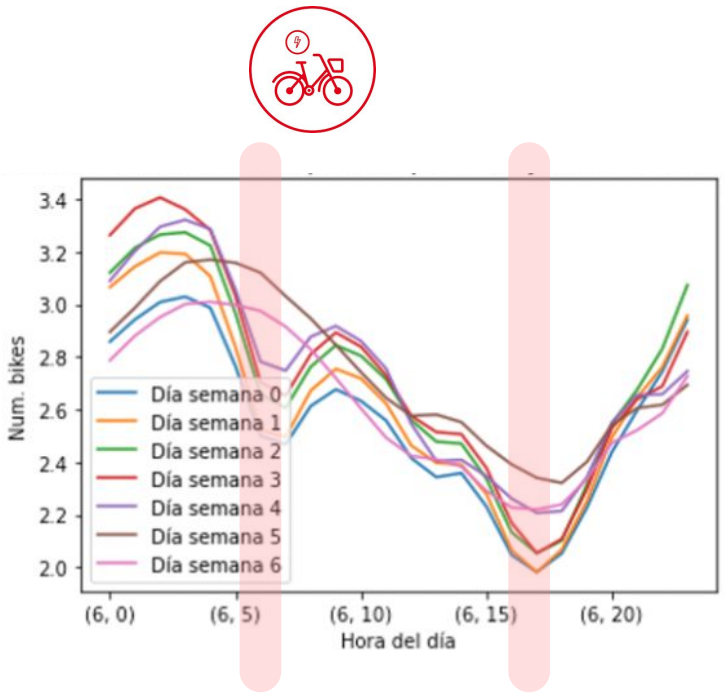
UNIVERSITAT DE BARCELONA x bicing

Goal: to explore and understand the data

Bikes availability per day of week and hour







1. Data preparation

UNIVERSITAT DE BARCELONA x bicing

Goal: to explore and understand the data

Key learnings

1

Se descarta la data de los **años** del 2019 y del 2020 por las anomalías detectadas en estos años.

2

Se seleccionan los **meses** de noviembre, diciembre, enero, febrero, marzo, abril y mayo porque presentan un comportamiento más similar en cuanto al uso de las bicis.

3

Se añade un **dataset con la información meteorológica** (temperatura y precipitación) para ajustar mejor la variación del uso mensual de las bicicletas.
También otro con información de **geolocalización** (longitud y latitud).

4

En los días laborables el uso de las bicicletas es mayor. En consecuencia, los días festivos se ven mermados. Es por eso que se añade un **dataset con la información de los días festivos** de cada año, para ajustar mejor la variación del uso diario de las bicicletas.

5

No existen diferencias muy relevantes entre el uso de las bicicletas eléctricas y mecánicas, por lo que se suprimen las variables que las distinguen.

2. Model design

Goal: to find the most accurate model

Train set

2021 and 2022.

Test set

March 2023.

Validation set

Bike stations which do not appear all the years (2019, 2022).

Pipeline

2. Model design

UNIVERSITAT DE BARCELONA x bicing

Goal: to find the most accurate model

Dataset 21-22	RMSE			
Model	**CV mean**	**Train**	**Val**	**Test**
Linear Regression	0.11361	0.11360	0.09456	0.11804
Linear Regression Lasso	0.11705	0.11704	0.09910	0.12136
Linear Regression Ridge	0.11361	0.11360	0.09456	0.11804
ElasticNet	0.11570	0.11569	0.09739	0.12007
Decision tree	0.10918	0.10752	0.09348	0.11381
Random forest	0.10759	0.10622	0.09016	0.11192
Gradient Boosting	0.10246	0.09484	0.09136	0.10755
eXtreme Gradient Boosting	0.10249	0.08899	0.09704	0.10877

Discarded models: KNN, Adaboost, Super vector machine.

2. Model design

Universitat de Barcelona x bicing

Goal: to find the most accurate model

Dataset 21-22	RMSE			
Model	**CV mean**	**Train**	**Val**	**Test**
eXtreme Gradient Boosting	0.10249	0.08899	0.09704	0.10877

n_estimators = 100

loss = 'squared_error'

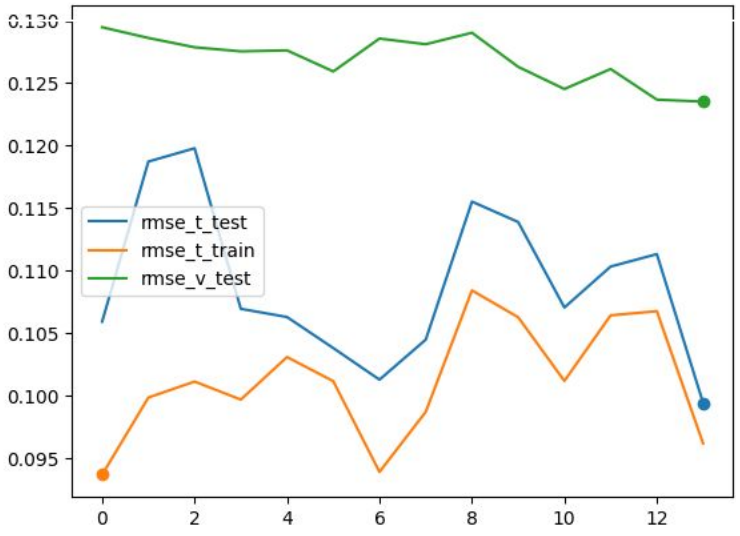
max_depth = 12

criterion = 'friedman_mse'

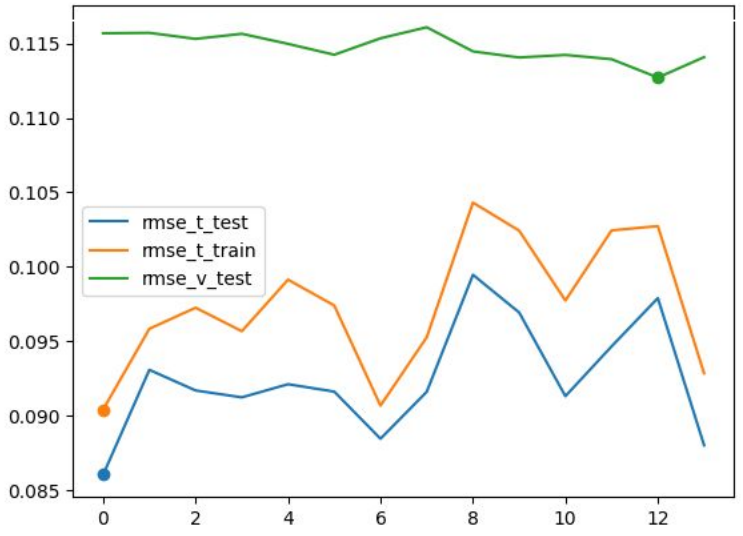
3. Results

UNIVERSITAT DE BARCELONA x bicing

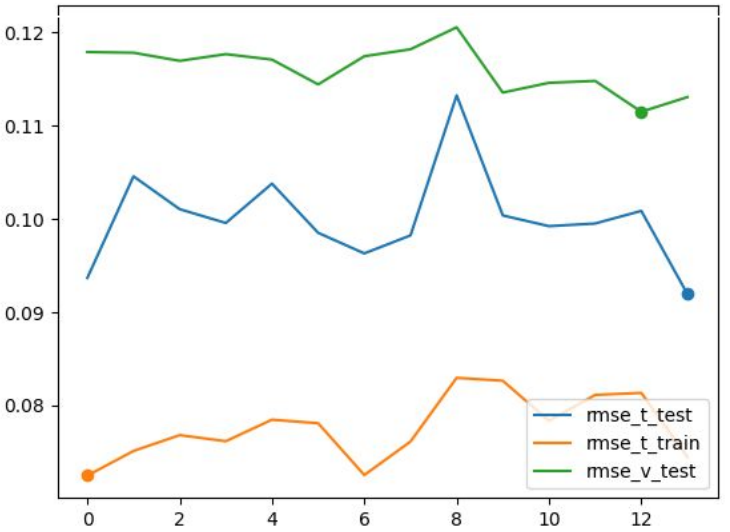
1 Decision Tree



2 Random Forest



3 eXtreme Gradient Boosting



4. Conclusions

UNIVERSITAT DE BARCELONA x bicing

eXtreme Gradient Boosting is the model with a higher performance when predicting bike availability in Barcelona, showing an error of 0.10281. Random Forest obtains the second position (0.10759 error) followed by Decision Tree on the top 3 (0.10918).

Ensembling models such as eXtreme Gradient Boosting and Random Forest **are more effective** when predicting bike availability in comparison to Decision Tree models due to the fact of combining multiple models which are weaker by default. Another positive point to resemble is that **overfitting does not affect that much** in those cases, showing more stability.

Enriching the data with meteorological and bank holidays datasets has been a game changer decision since the results were significatively improved compared to same models without this information.

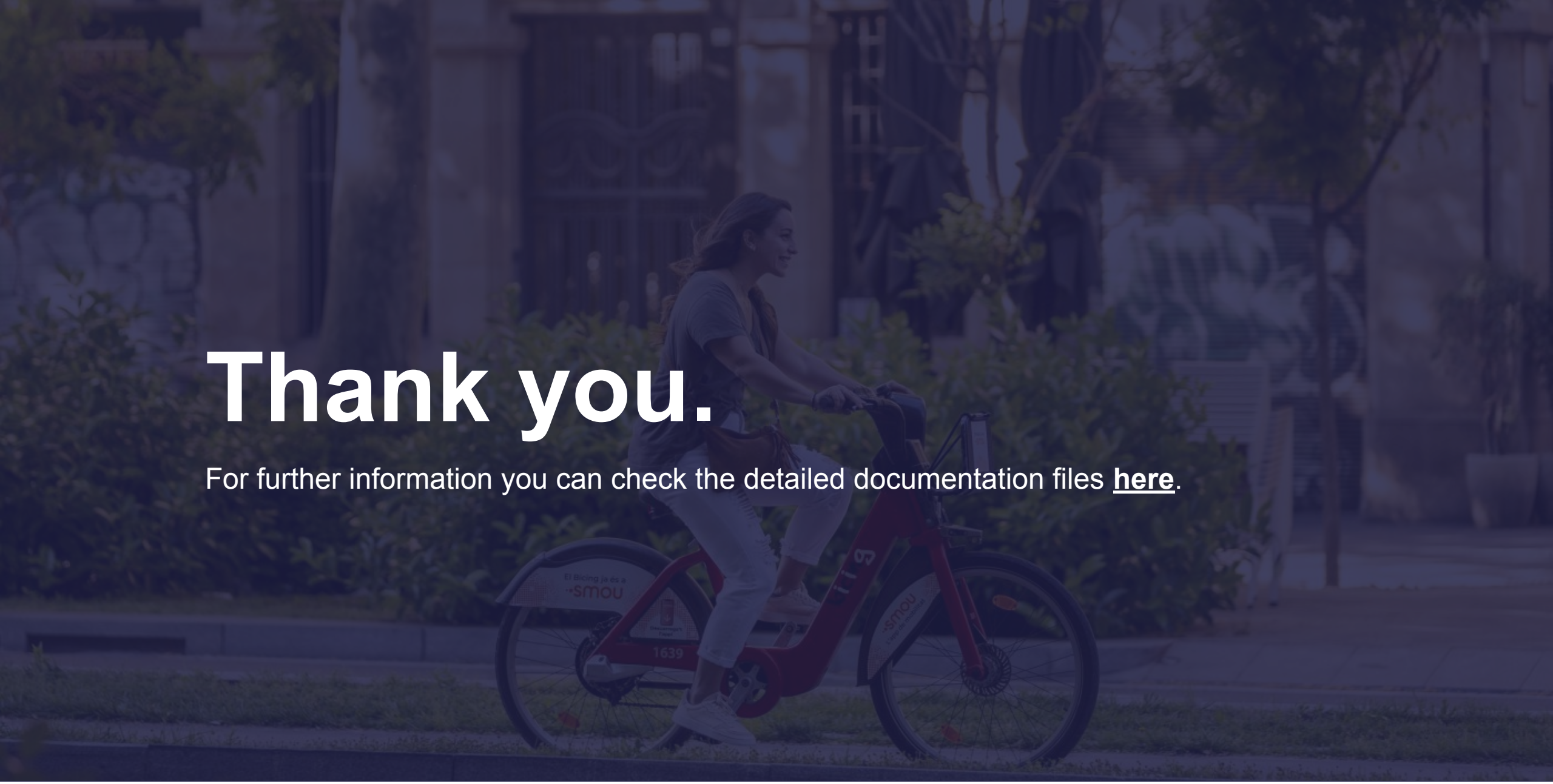
5. Next Steps & Proposals

UNIVERSITAT DE BARCELONA x bicing

Models with better performance require **higher computational capacity** and are **too sensitive to the parameters adjustments**. In order to obtain better results it would be recommended to **continue looking for the most optimal parameters**.

Elaborate **four different models** attending to the **seasonal weather variances**.

Additionally, it would be interesting to **study each station individually** as a way of understanding whether users could find no bikes nor docks available when needed. Those situations generate really bad users experiences and it would be highly recommended to avoid them.



Thank you.

For further information you can check the detailed documentation files **here**.

UNIVERSITAT DE BARCELONA X bicing