Data Science Bootcamp

Bike purchase prediction

Jacqueline Van Grellier

14th February 2022

The bike, the new trend in town

- **Context:** In 2020, an increase of bike purchase has been observed in France, and also abroad.

- **Objective:** To predict if a person will purchase a bike or not, depending on different features related to the person.

- **Data source:** from Kaggle (September 2020)
https://www.kaggle.com/heeraldedhia/bike-buyers?select=bike_buyers_clean.csv

Our journey

- Data Collection
- Data Exploring
- Data Cleaning
- Models
- Results
- What’s next?

Data Collection

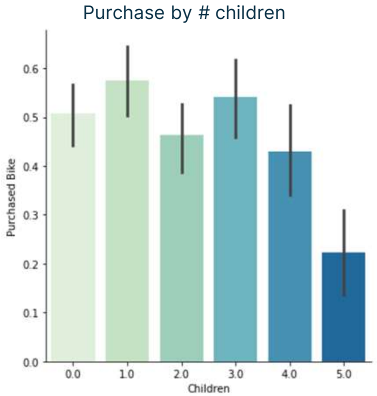
	ID	Marital Status	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age	Purchased Bike
0	12496	Married	Female	40000.0	1.0	Bachelors	Skilled Manual	Yes	0.0	0-1 Miles	Europe	42.0	No
1	24107	Married	Male	30000.0	3.0	Partial College	Clerical	Yes	1.0	0-1 Miles	Europe	43.0	No
2	14177	Married	Male	80000.0	5.0	Partial College	Professional	No	2.0	2-5 Miles	Europe	60.0	No
3	24381	Single	NaN	70000.0	0.0	Bachelors	Professional	Yes	1.0	5-10 Miles	Pacific	41.0	Yes
4	25597	Single	Male	30000.0	0.0	Bachelors	Clerical	No	0.0	0-1 Miles	Europe	36.0	Yes

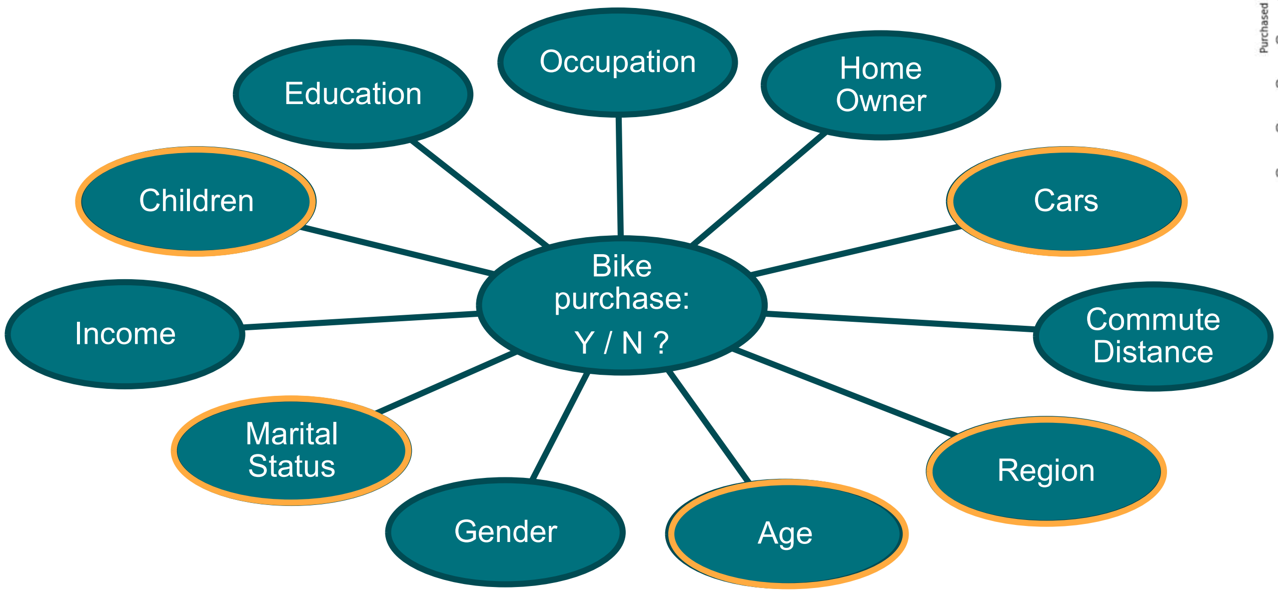
(1000, 13)

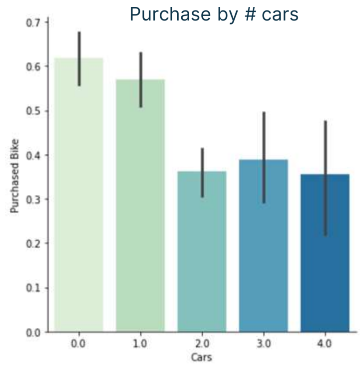
How to predict the
bike purchase?
Feature variable (x)

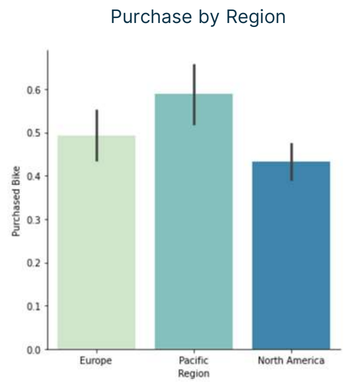
What are we
predicting?
Target variable (y)

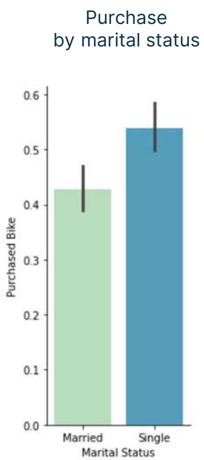
Data Exploring

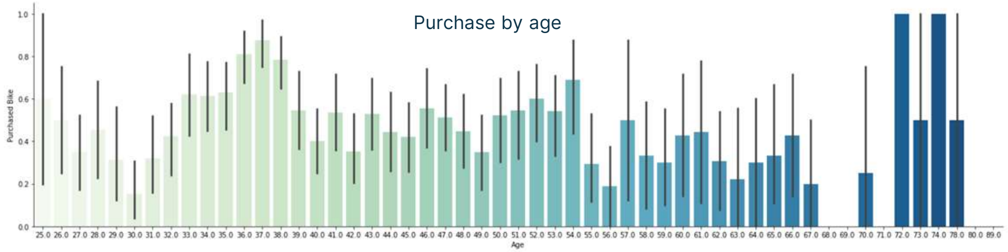












Data Cleaning

Missing data

➔ For feature variables

Using median for numerical variables

Using the most frequent value for categorical variables

Data Update

➔ For target variable:

0 means “No purchase”

1 means “Purchase”

Data removal

➔ Person ID removed as uncessary

Models

- Use of classification models
  1. Logistic regression
  2. Decision tree
  3. Random Forests

- Optimize the models

  Testing of several parameters on decision tree and random forests models

- Objective

  Best prediction rate, with test performance as closest as possible to train performance

Results: best model prediction

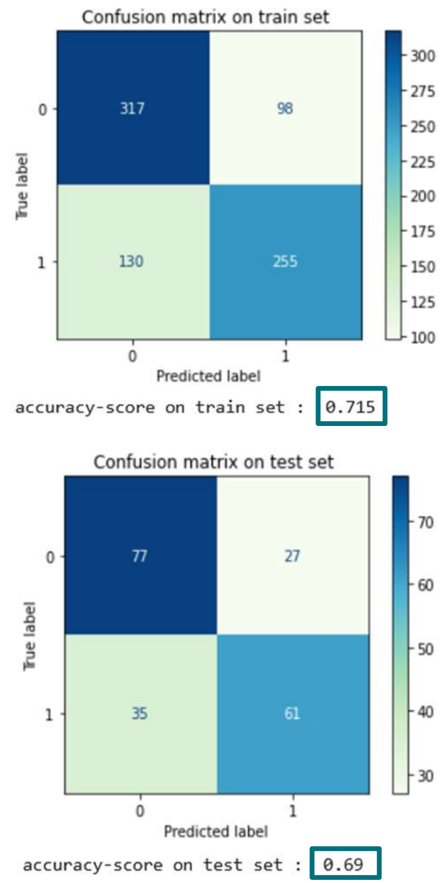
Initial and best performance results

Accuracy-score	Logistic Regression	Decision Tree		Random Forest		Avg (on 20 tests)
# ID	**LG1**	**DT1**	**DT2**	**RF1**	**RF2**	**AVG**
On train set	0,66500	0,99500	0,71125	0,99250	0,71500	0,66500
On test set	0,61500	0,65000	0,64500	0,70500	0,69000	0,61500
Difference (train-test)	0,05000	**0,34500**	0,06625	**0,28750**	**0,02500**	0,05000

Performance of **trained** model too high
vs performance of **tested** model

Best performance
(diff.RF2 < diff.RF1 < diff.DT1)

Confusion matrix of best model

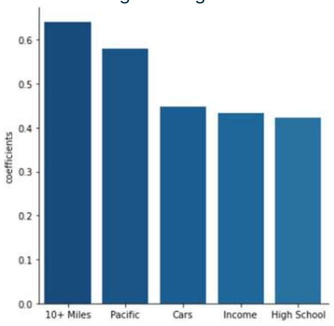


Results: key feature variables

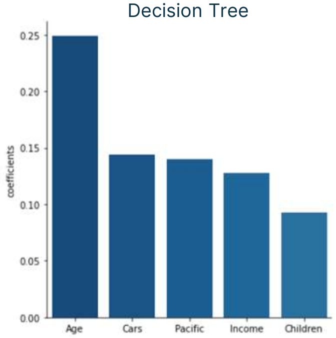
Top 5 of features importance (for best model in each model type)

Features weight		
Logistic Regression LG1	**Decision Tree DT1**	**Random Forest RF2**
1. 10+ Miles	1. Age	1. Age
2. Pacific	2. Cars	2. Children
3. Cars	3. Pacific	3. Cars
4. Income	4. Income	4. Income
5. High School	5. Children	5. Married

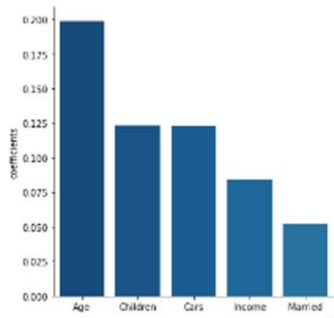
Top 5 of features importance
Logistic Regression



Top 5 of features importance
Decision Tree



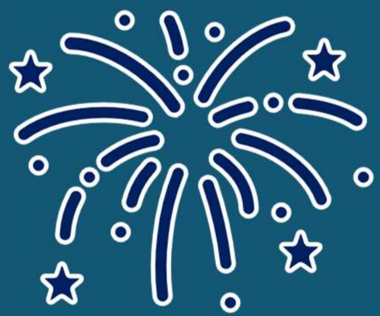
Top 5 of features importance
Random Forest



What’s next?

- To collect more data (only 1 000 entries)
- New feature variables to improve the model accuracy
  - Home location: in town or in countryside
  - Public transport availability: yes or no
  - Bike infrastructure: yes or no

Pensez à l’antivol !



Jacqueline Van Grellier

Jedha

Merci,

à bientôt !