20/11/2020 – Midterm project

Housing prices – Linear regression

Jamil AKILI

Introduction

- Overview of the overall experience of the project with a focus on:
  - Main difficulties and challenges
  - How they were handled
  - Key takeaways and next steps
- Addressing Ironhackers (not a client presentation)
- Focus: Python
- Data overview:

	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	with_basement	yr_built	zipcode	price
0	3	1.00	1340	5650	1.0	0	0	3	7	0	1955	98178	221900
1	3	2.25	1690	7639	2.0	0	0	3	7	1	1951	98125	538000
2	2	1.00	2720	8062	1.0	0	0	3	6	0	1933	98028	180000
3	4	3.00	1360	5000	1.0	0	0	5	7	1	1965	98136	604000
4	3	2.00	1800	7503	1.0	0	0	3	8	0	1987	98074	510000

Main challenge: achieving and maintaining a structured approach

- Different ways to manipulate the data
  - Categories vs. continuous numerical vs. discrete numerical
- Different possible approaches to pre-processing
- Different models to be tested (LinReg, Lasso, Ridge, RFR, KNN,…)
- Different indicators for model accuracy (R2, R2 adjusted, RMSE, MAE…)

How to test all approaches and run multiple iterations on multiple models

WHILE

documenting each step of the process and measuring impact on model accuracy?

The solution: a python function

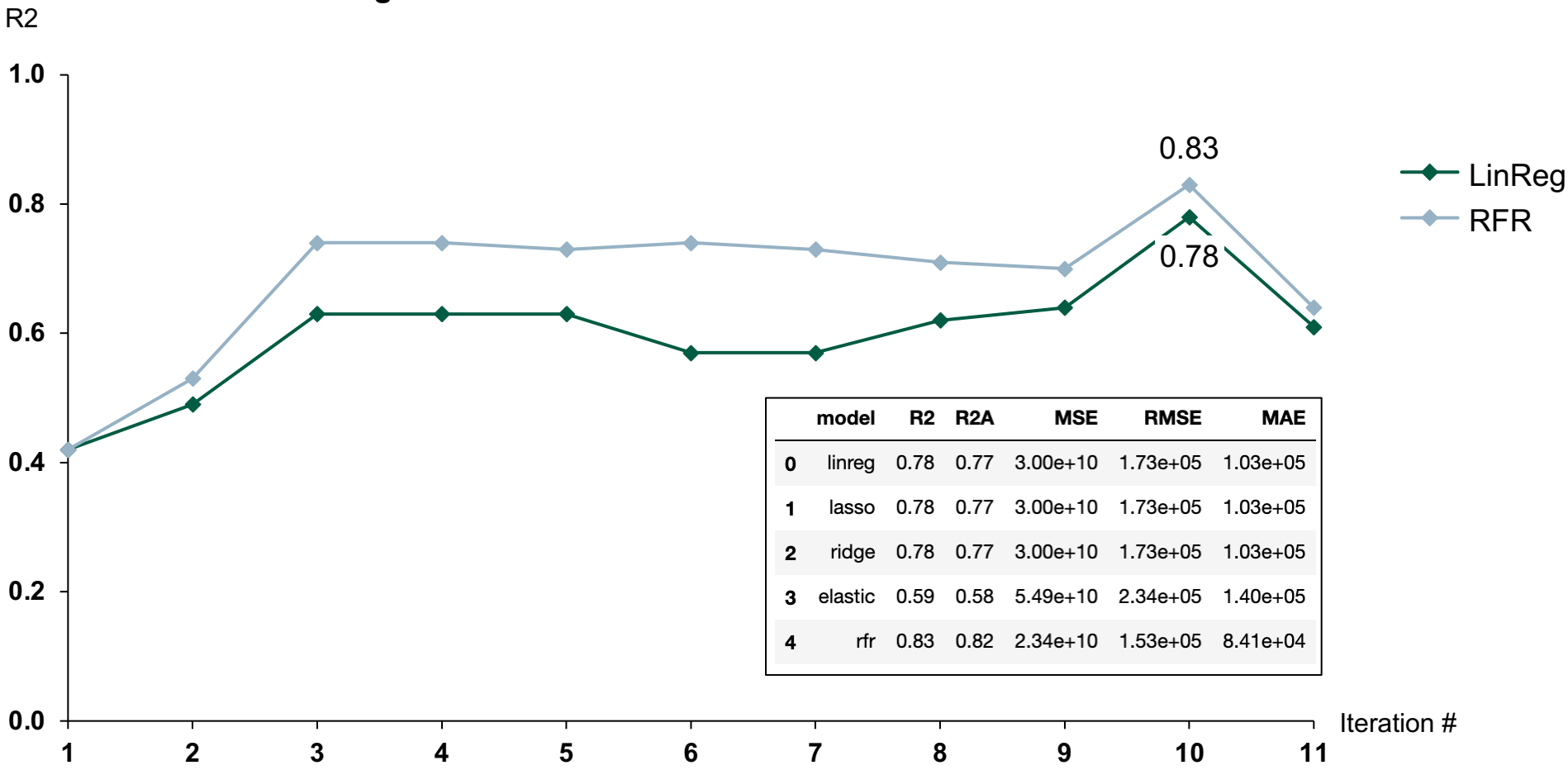
```python
def run_models(a,b,c,d):
```

	model	R2	R2A	MSE	RMSE	MAE
0	linreg	0.42	0.42	7.77e+10	2.79e+05	1.80e+05
1	lasso	0.42	0.42	7.77e+10	2.79e+05	1.80e+05
2	ridge	0.42	0.42	7.77e+10	2.79e+05	1.80e+05
3	elastic	0.41	0.40	8.00e+10	2.83e+05	1.78e+05
4	rfr	0.42	0.42	7.81e+10	2.79e+05	1.80e+05

- (+) My very first self-written, reusable « block » of code
- (+) Introduces a solid structure in the notebook and helps keep track
- (+) Makes the rest of the modelling work more enjoyable and allows for more iterations
- (-) Risky waste of time at the beginning of the project
- (-) The code could be wrong
- (-) Could/should have included KNN in the function

Maximum R2 value reached: 0.83 for RFR

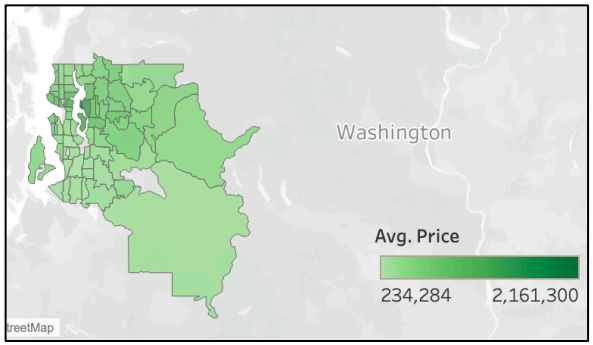
Evolution of R2 for LinReg and RFR

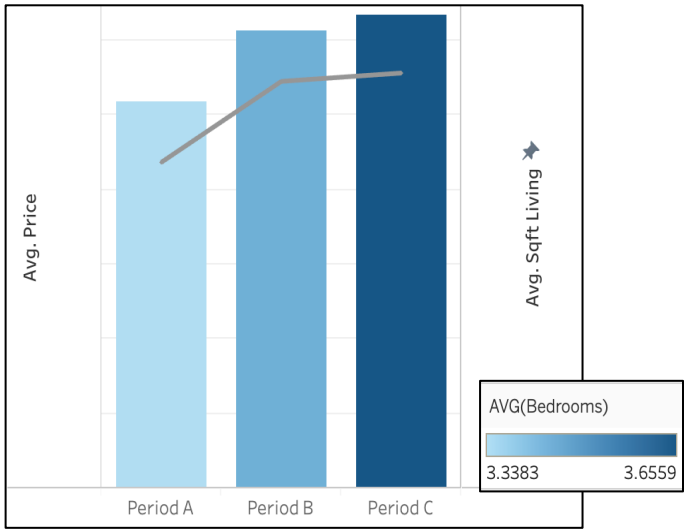
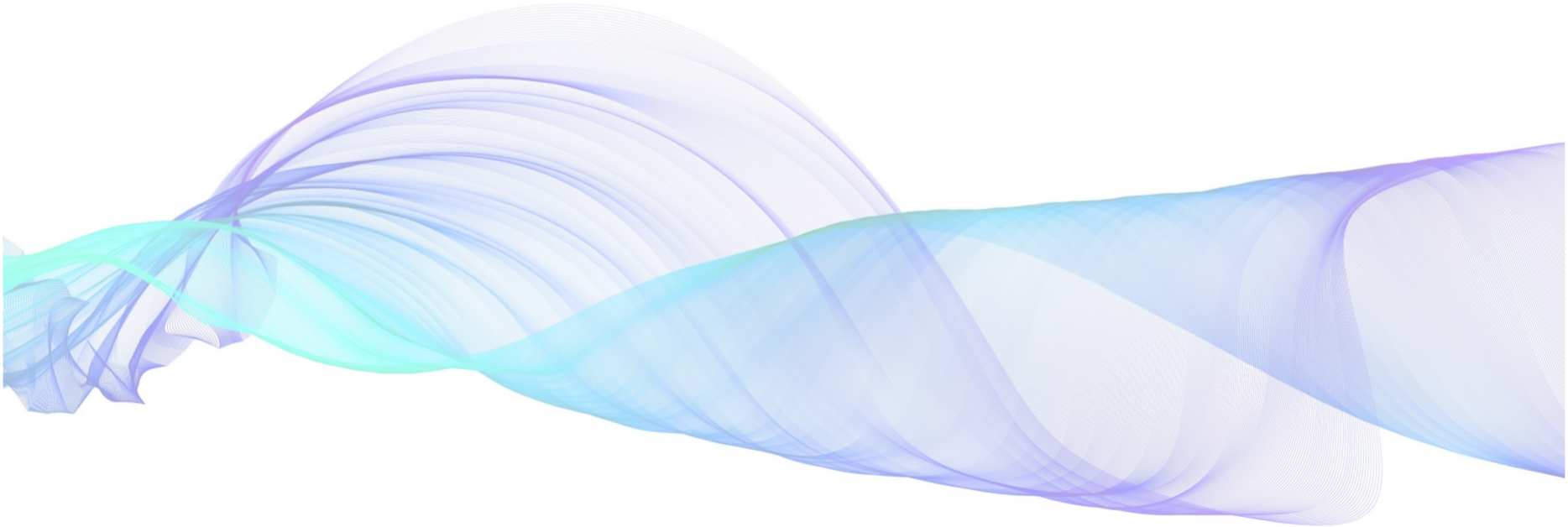


	model	R2	R2A	MSE	RMSE	MAE
0	linreg	0.78	0.77	3.00e+10	1.73e+05	1.03e+05
1	lasso	0.78	0.77	3.00e+10	1.73e+05	1.03e+05
2	ridge	0.78	0.77	3.00e+10	1.73e+05	1.03e+05
3	elastic	0.59	0.58	5.49e+10	2.34e+05	1.40e+05
4	rfr	0.83	0.82	2.34e+10	1.53e+05	8.41e+04

Findings / next steps

- EDA based on thorough understanding of the dataset has the highest impact on model accuracy
- Scaling and normalizing have little impact on the linear regression model
- Some discrete data should be binarized, even if it makes the data volume bigger (e.g. zipcode)…
- …but not all discrete data (e.g. “year_built”)
- Use Tableau to explore other, possibly more relevant, KPIs (e.g. **price/sqft_living** as a proxy for **price/sqft**)





Thank you!