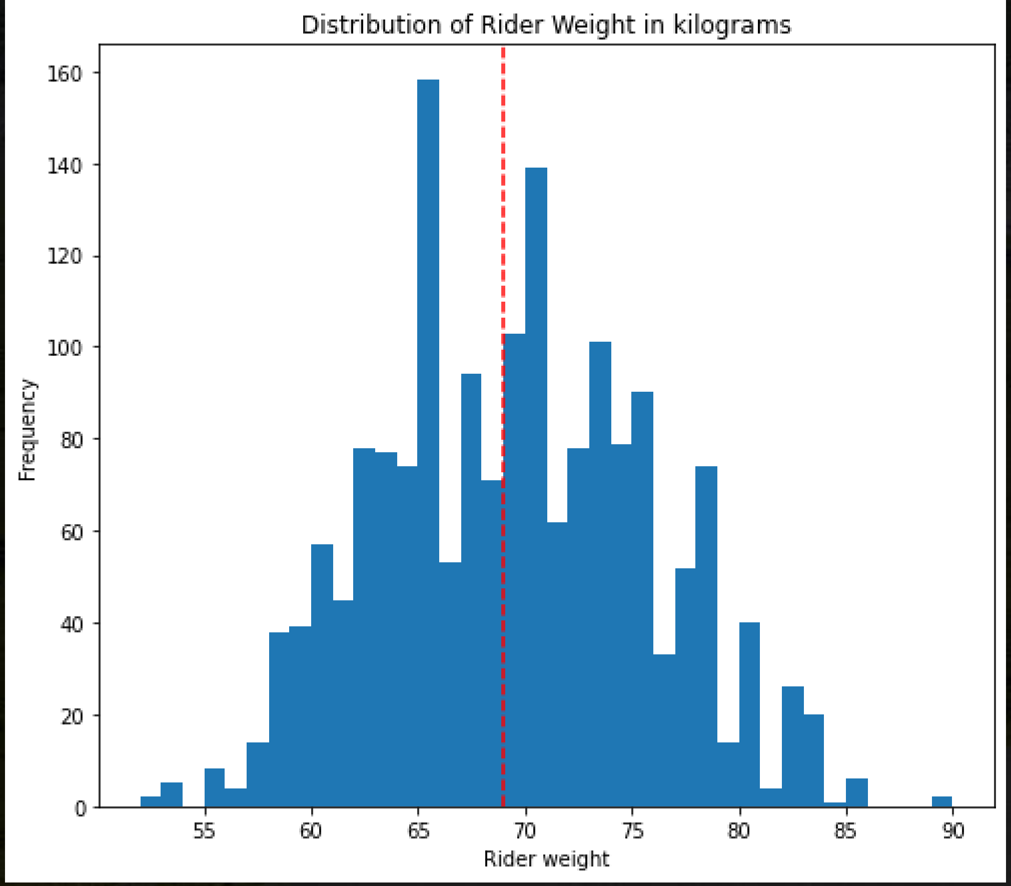
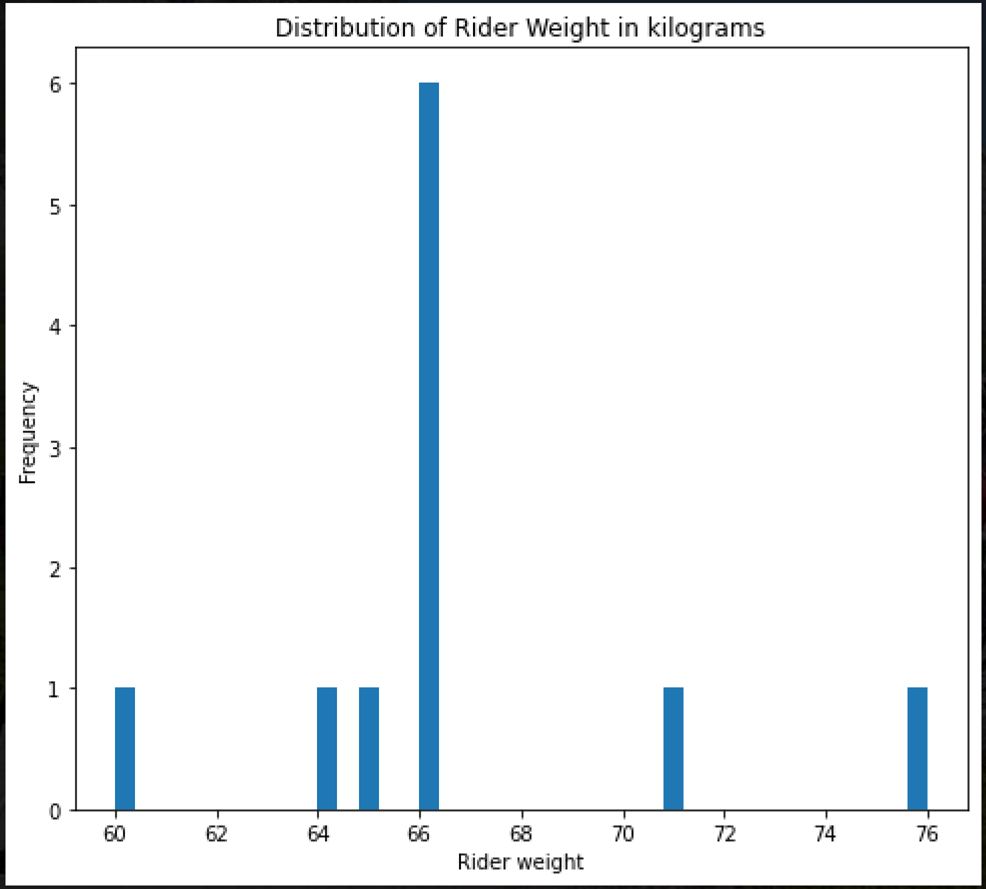
Do you have what it takes to win Le Tour De France?

Looking at historical data of past riders, and trying to predict the outcomes for the next Tour De France using Machine Learning

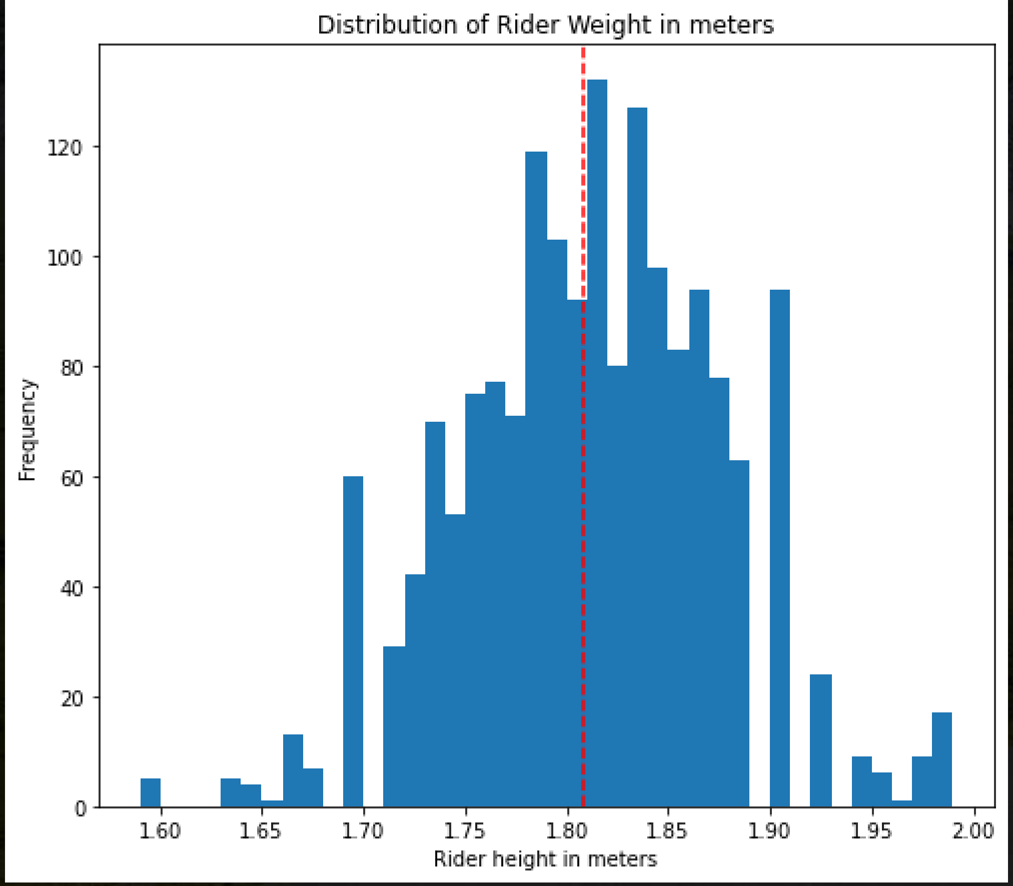
By: Raminderpreet Singh Khaira

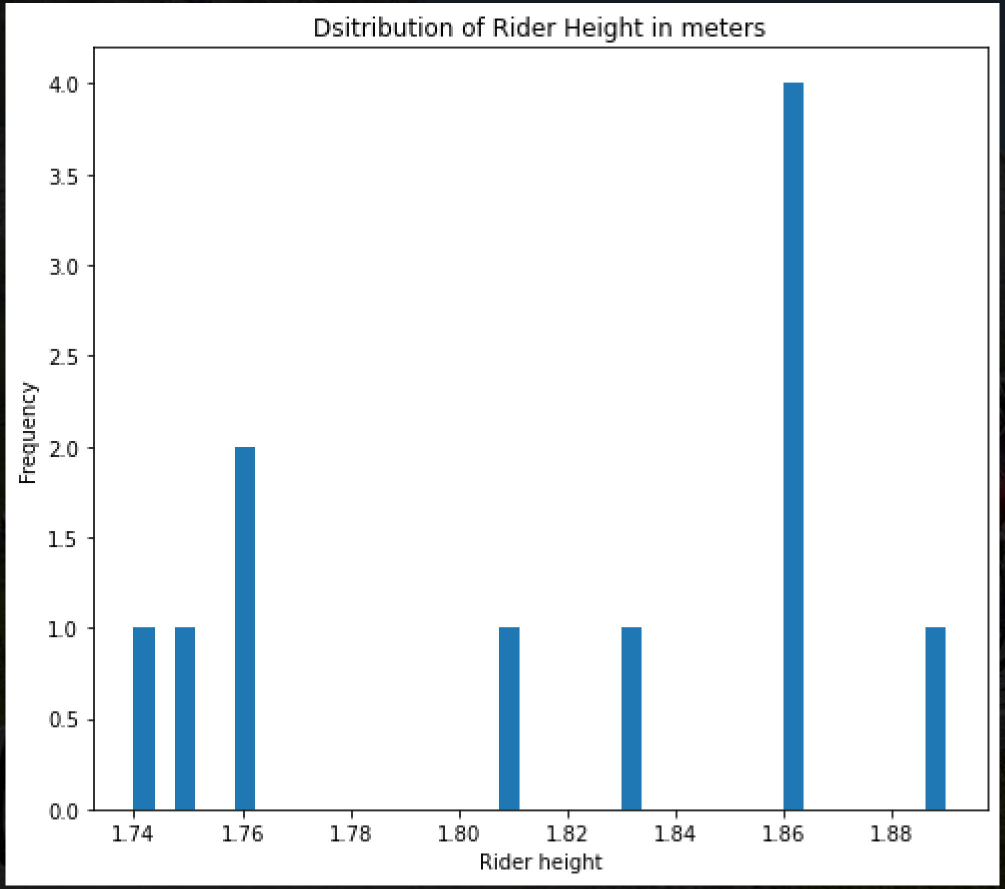
Clean-Up and EDA



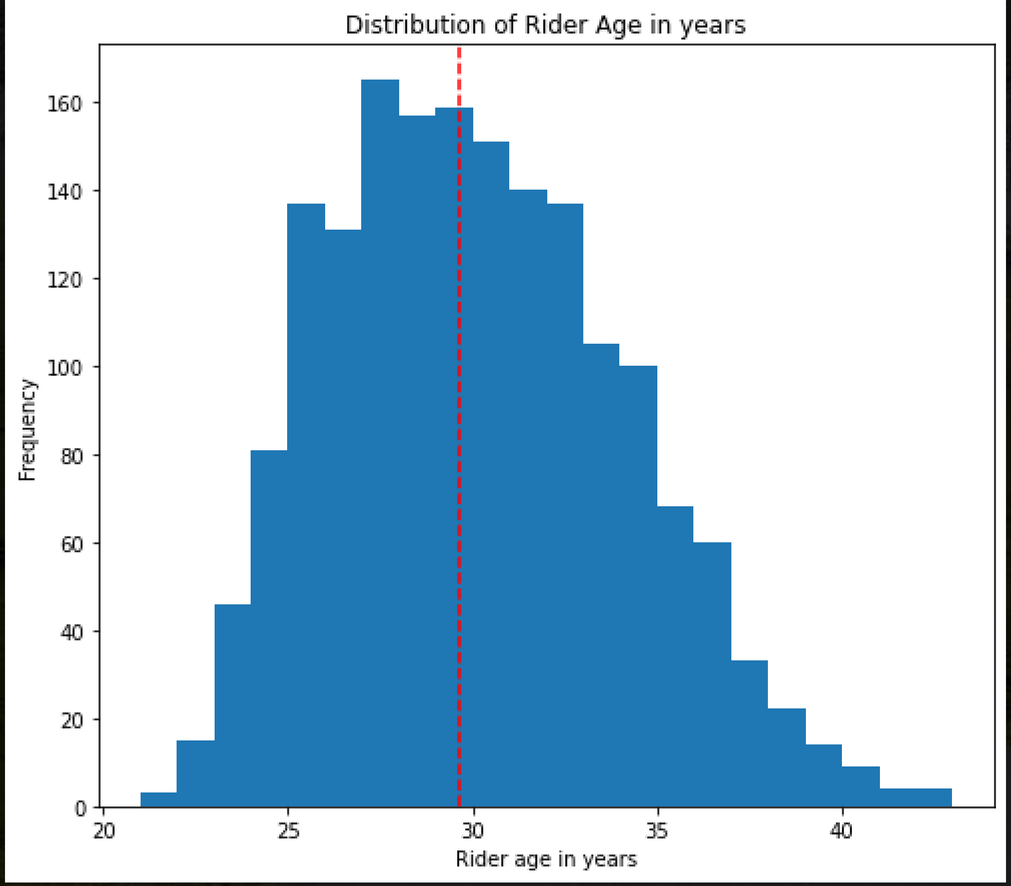


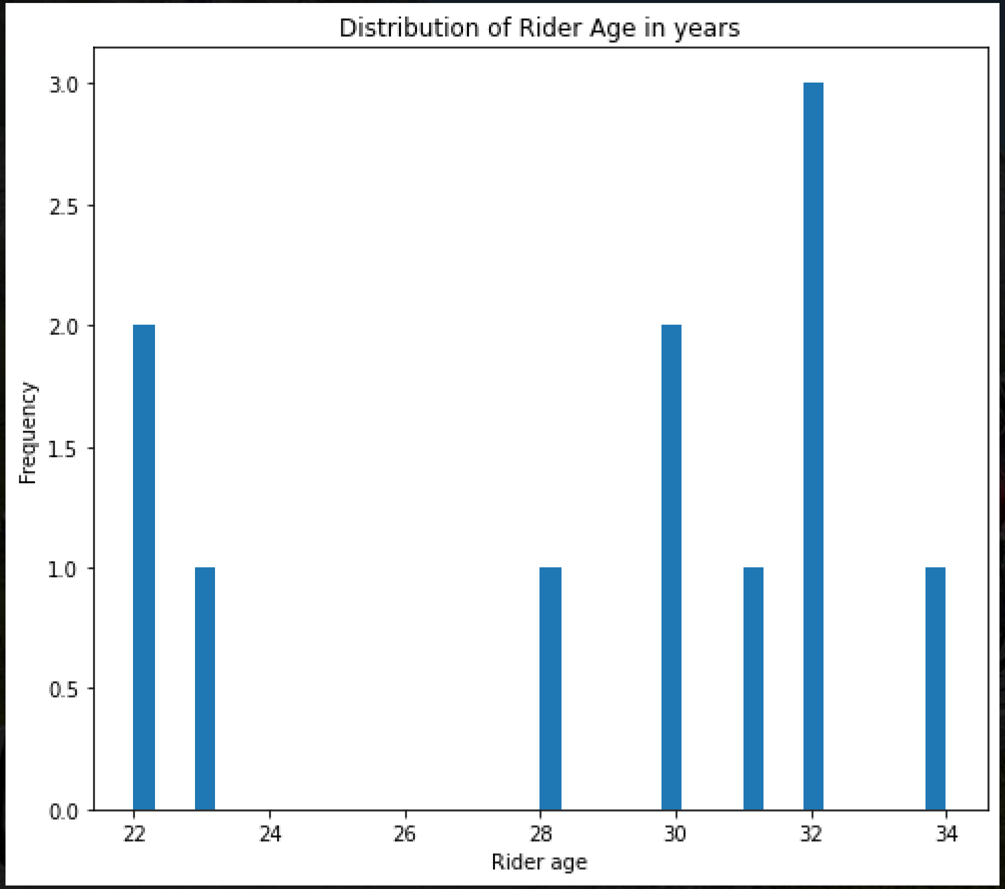
Clean-Up and EDA



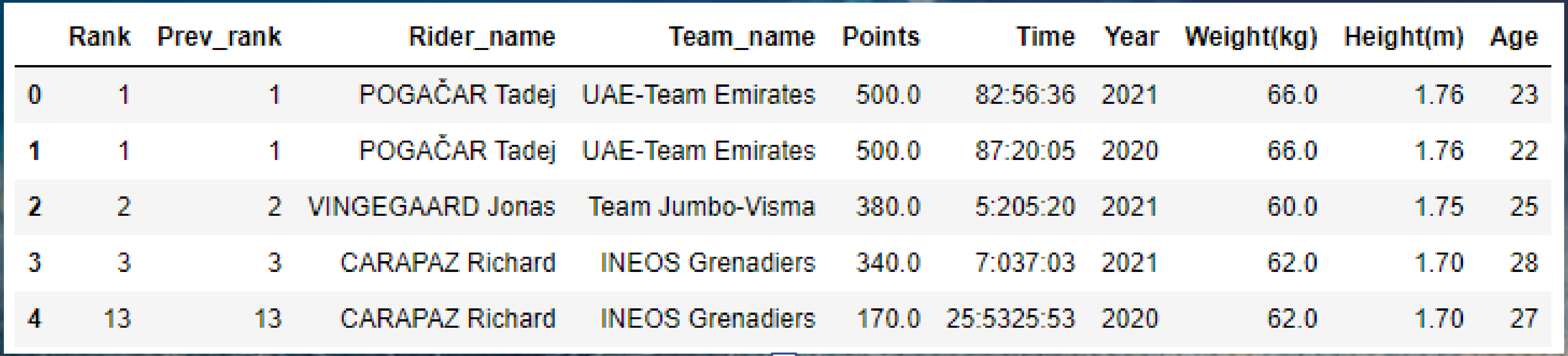


Clean-Up and EDA

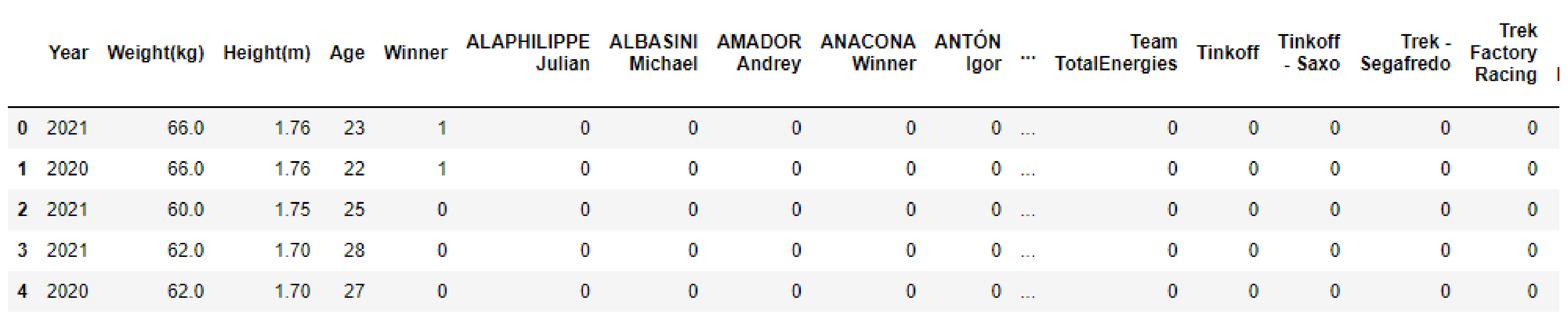




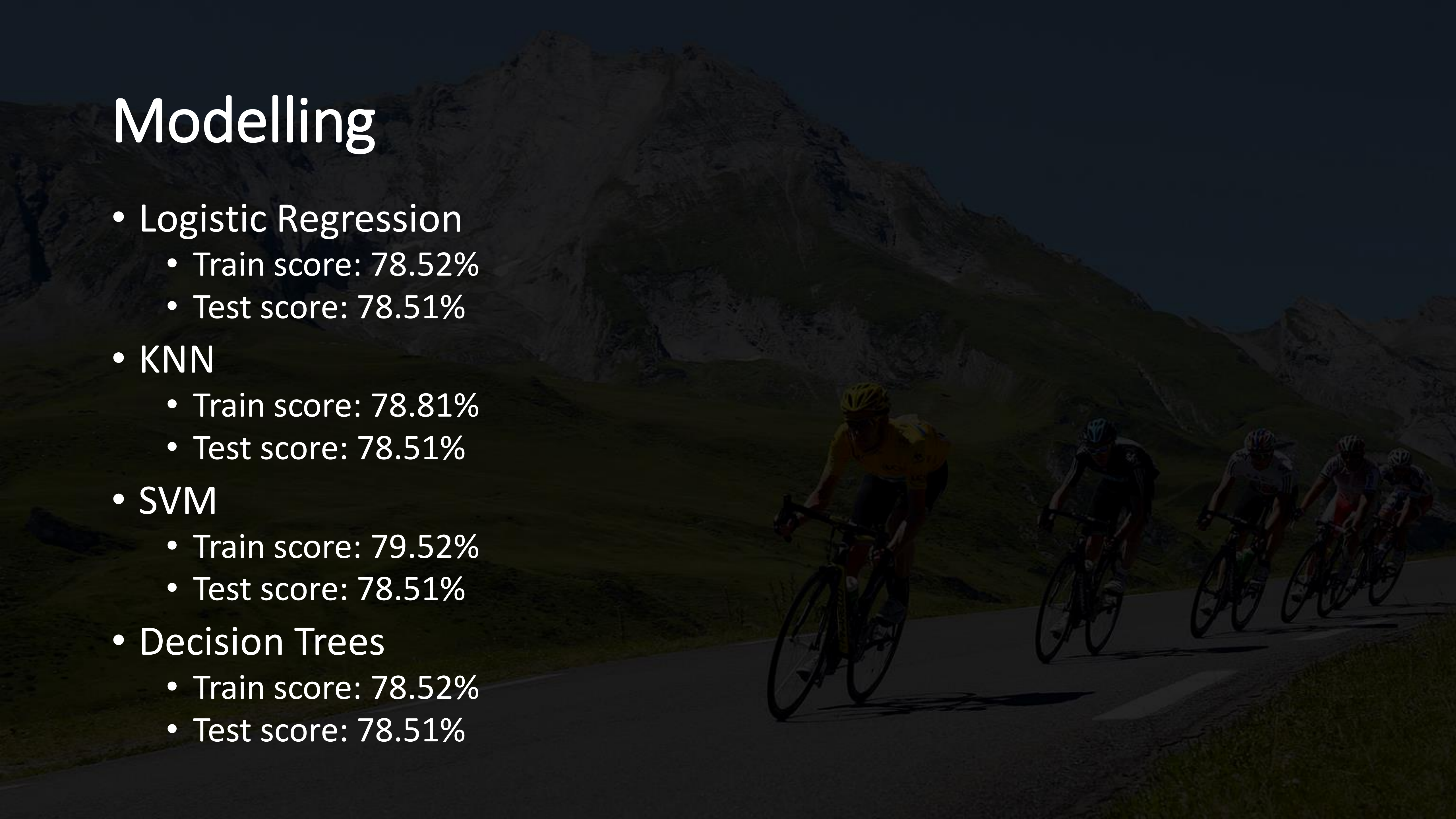
Data Description (categorical)



	Rank	Prev_rank	Rider_name	Team_name	Points	Time	Year	Weight(kg)	Height(m)	Age
0	1	1	POGAČAR Tadej	UAE-Team Emirates	500.0	82:56:36	2021	66.0	1.76	23
1	1	1	POGAČAR Tadej	UAE-Team Emirates	500.0	87:20:05	2020	66.0	1.76	22
2	2	2	VINGEGAARD Jonas	Team Jumbo-Visma	380.0	5:205:20	2021	60.0	1.75	25
3	3	3	CARAPAZ Richard	INEOS Grenadiers	340.0	7:037:03	2021	62.0	1.70	28
4	13	13	CARAPAZ Richard	INEOS Grenadiers	170.0	25:5325:53	2020	62.0	1.70	27



	Year	Weight(kg)	Height(m)	Age	Winner	ALAPHILIPPE Julian	ALBASINI Michael	AMADOR Andrey	ANACONA Winner	ANTÓN Igor	...	Team TotalEnergies	Tinkoff	Tinkoff - Saxo	Trek - Segafredo	Trek Factory Racing
0	2021	66.0	1.76	23	1	0	0	0	0	0	...	0	0	0	0	0
1	2020	66.0	1.76	22	1	0	0	0	0	0	...	0	0	0	0	0
2	2021	60.0	1.75	25	0	0	0	0	0	0	...	0	0	0	0	0
3	2021	62.0	1.70	28	0	0	0	0	0	0	...	0	0	0	0	0
4	2020	62.0	1.70	27	0	0	0	0	0	0	...	0	0	0	0	0

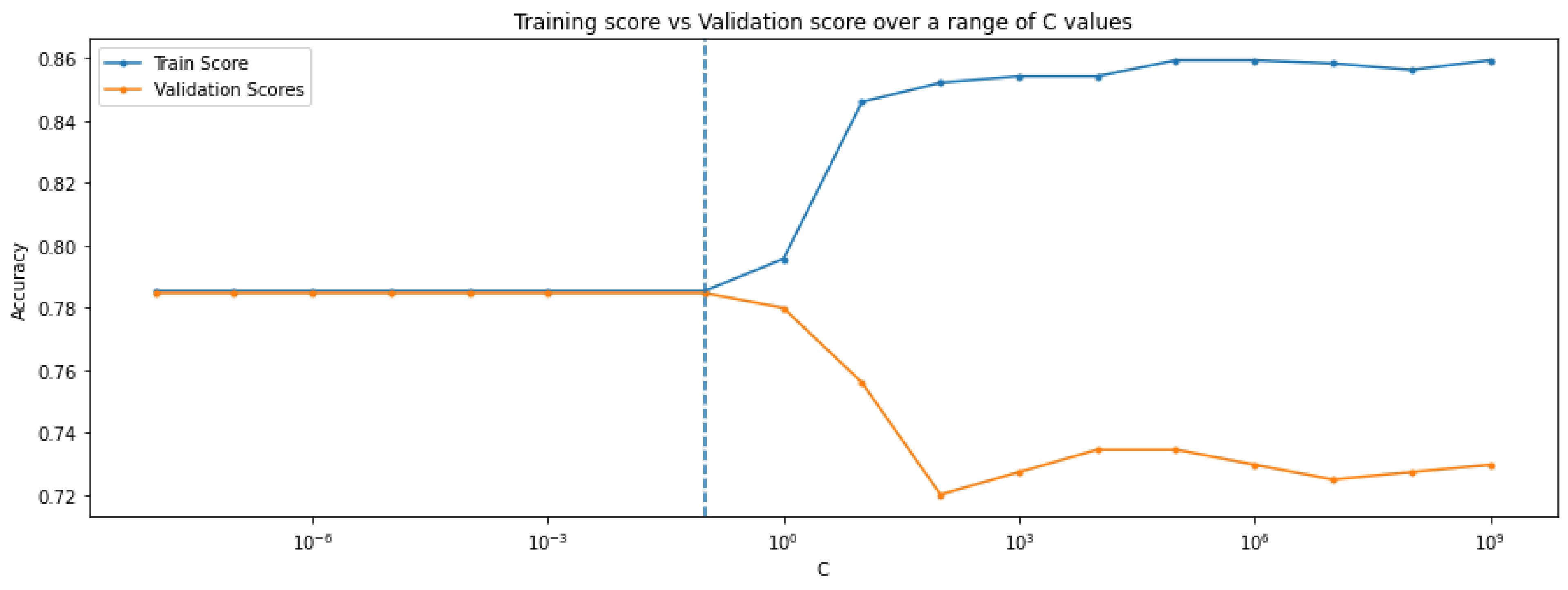


Modelling

- Logistic Regression
  - Train score: 78.52%
  - Test score: 78.51%
- KNN
  - Train score: 78.81%
  - Test score: 78.51%
- SVM
  - Train score: 79.52%
  - Test score: 78.51%
- Decision Trees
  - Train score: 78.52%
  - Test score: 78.51%

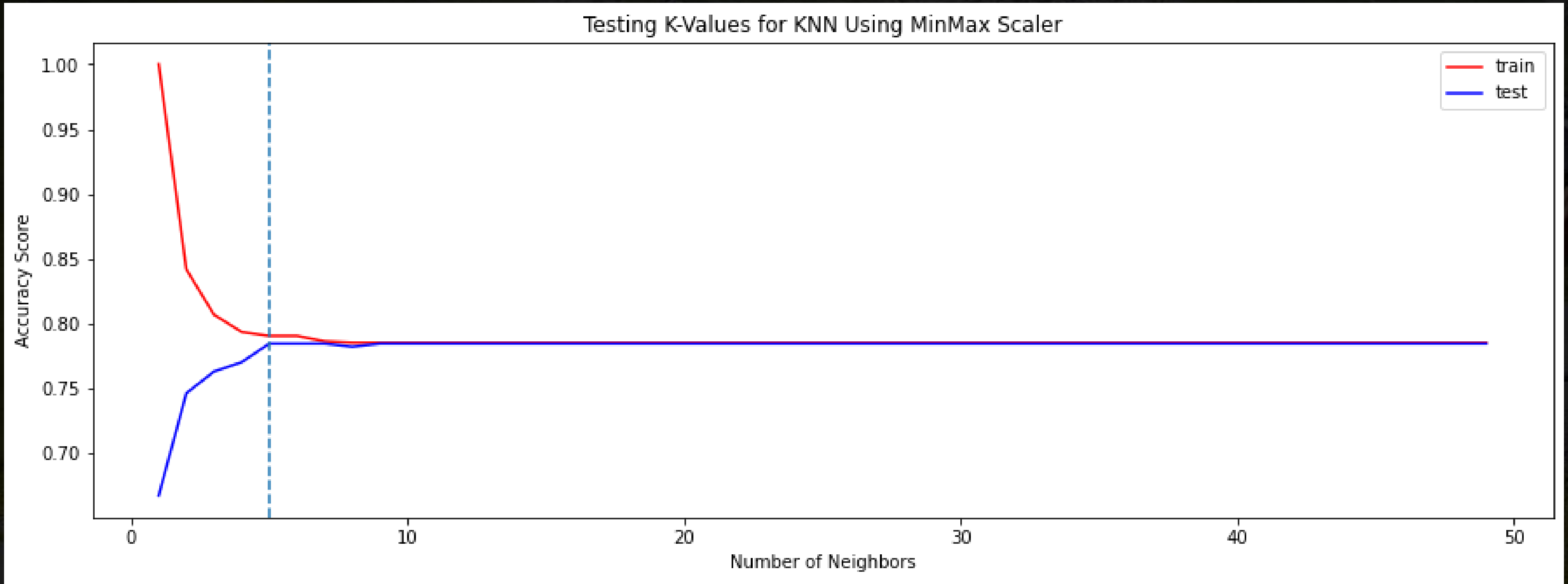
Modelling

- Logistic Regression
  - Train score: 78.52%
  - Test score: 78.51%



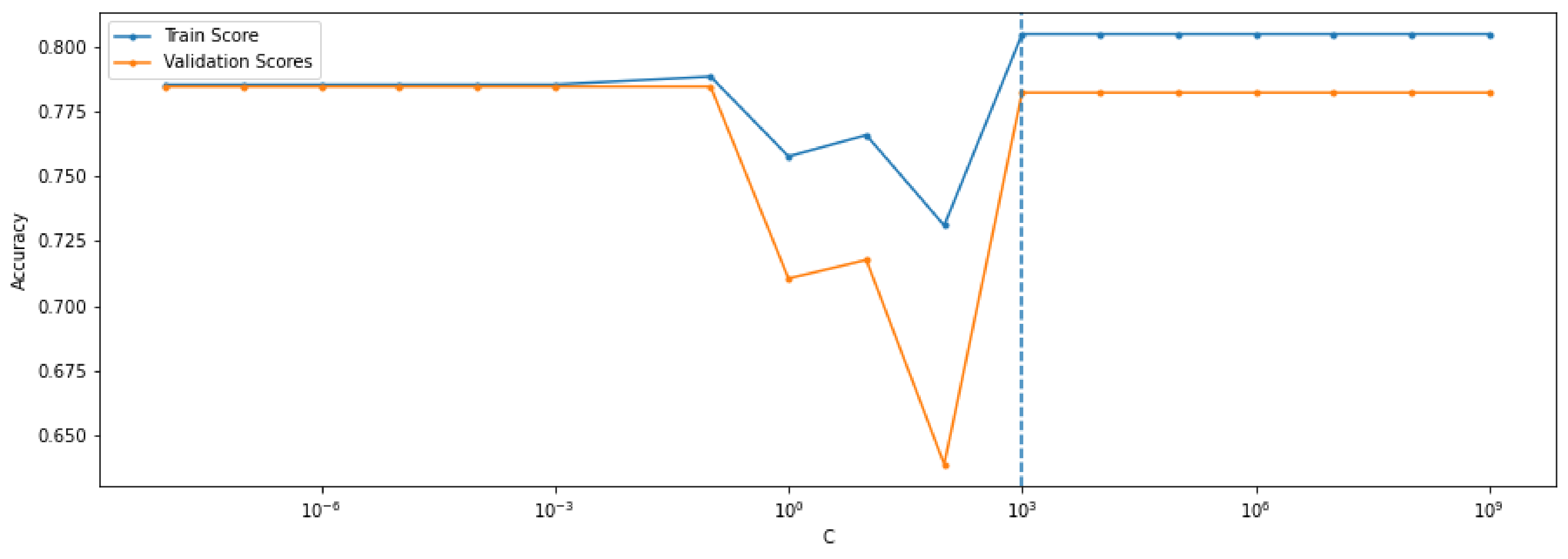
Modelling

- KNN
  - Train score: 78.81%
  - Test score: 78.51%



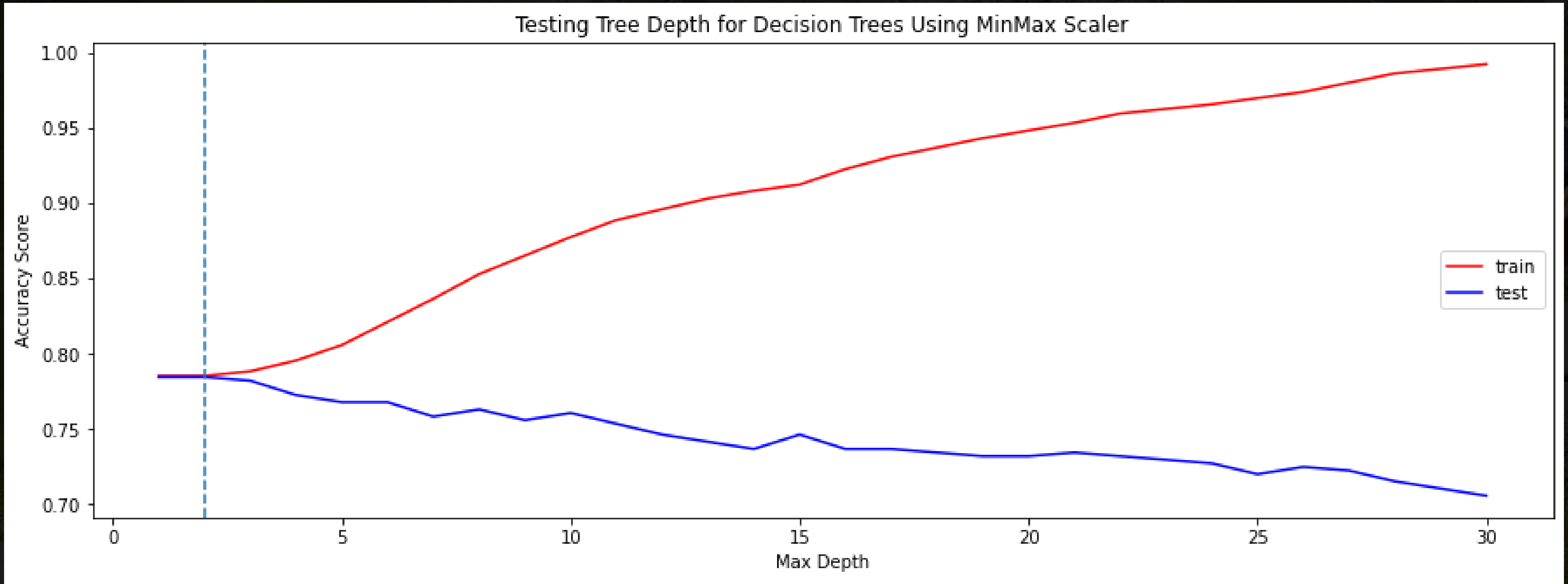
Modelling

- SVM
  - Train score: 79.52%
  - Test score: 78.51%

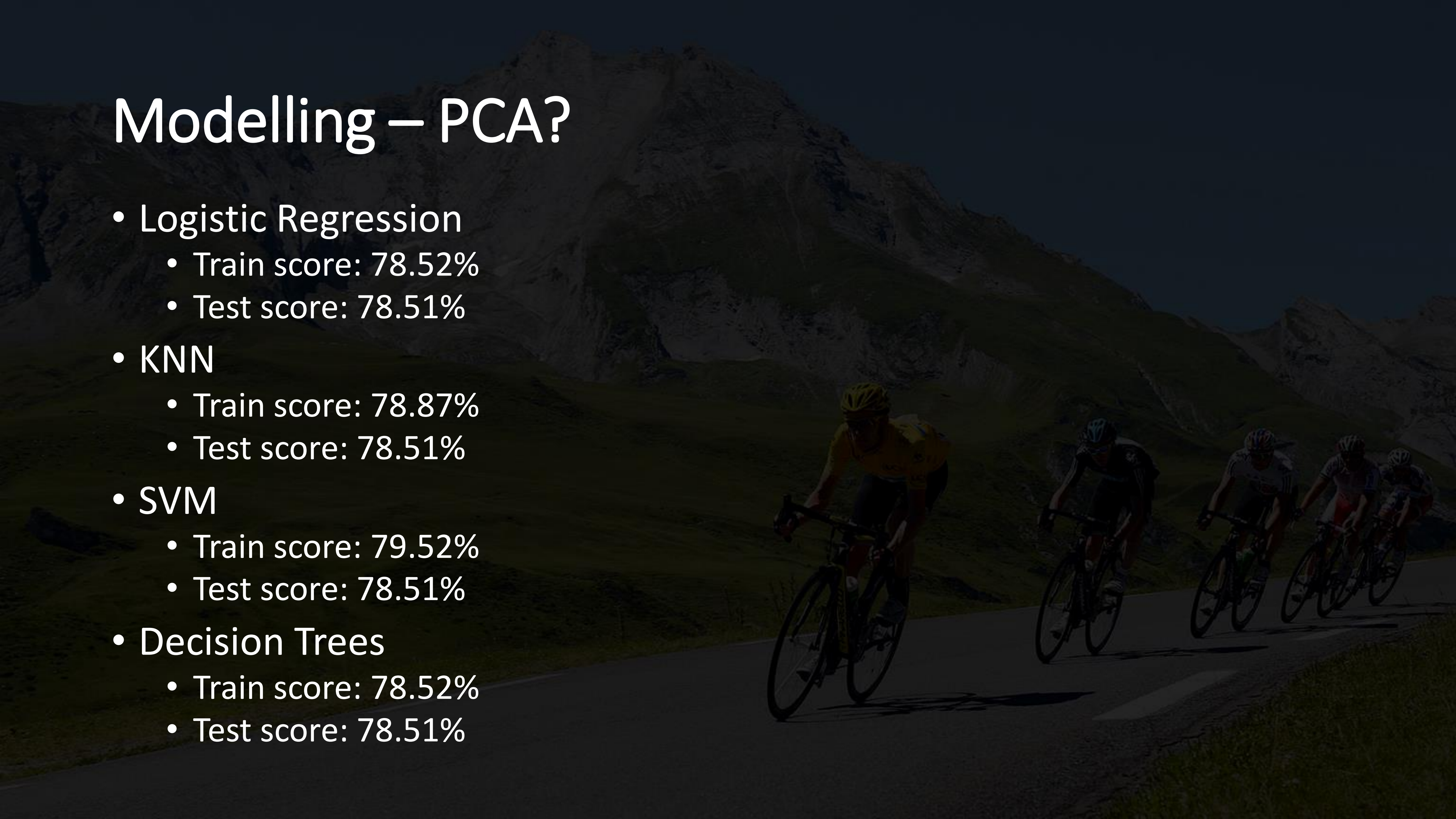


Modelling

- Decision Trees
  - Train score: 78.52%
  - Test score: 78.51%



Modelling – PCA?



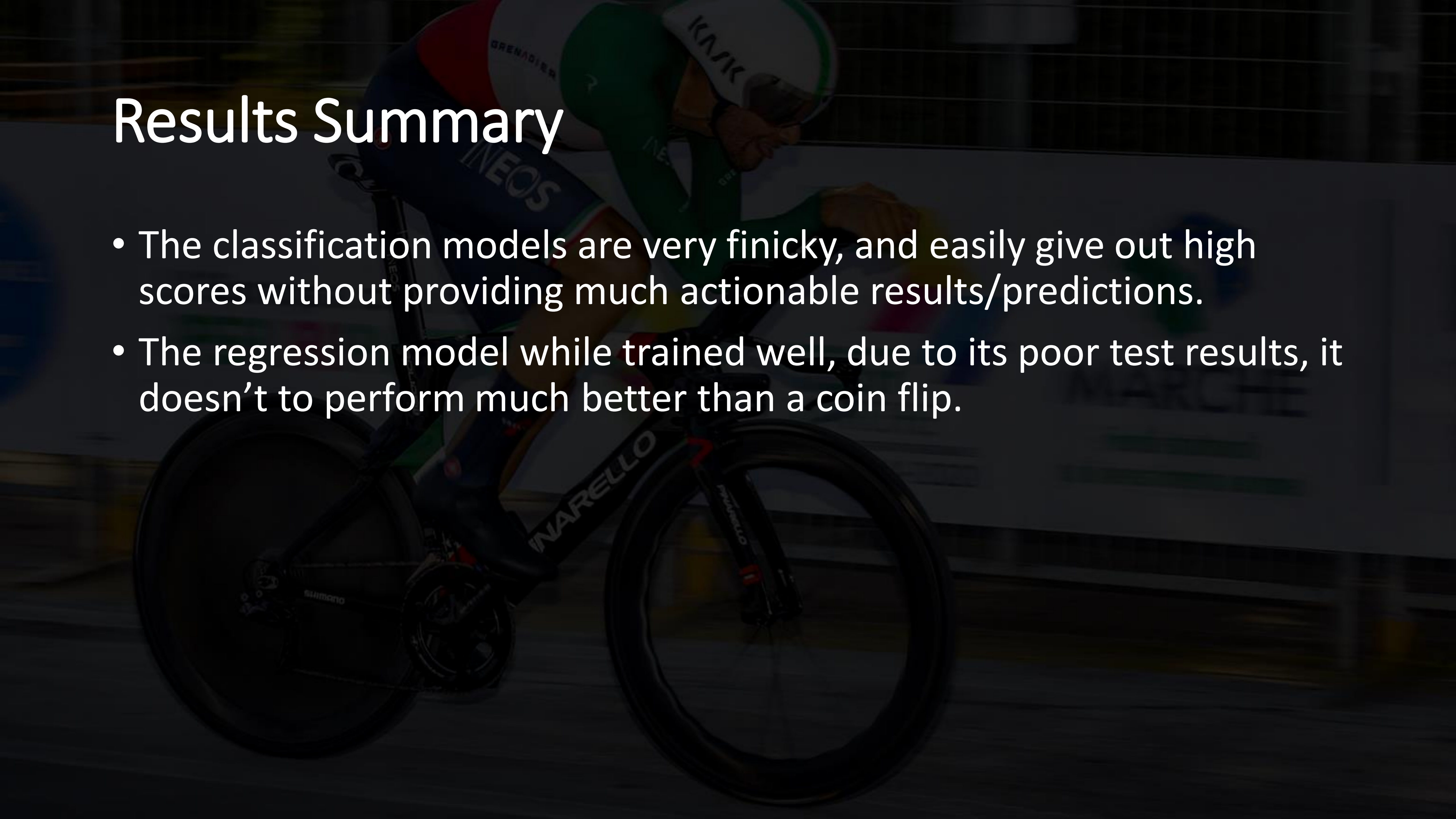
- Logistic Regression
  - Train score: 78.52%
  - Test score: 78.51%
- KNN
  - Train score: 78.87%
  - Test score: 78.51%
- SVM
  - Train score: 79.52%
  - Test score: 78.51%
- Decision Trees
  - Train score: 78.52%
  - Test score: 78.51%

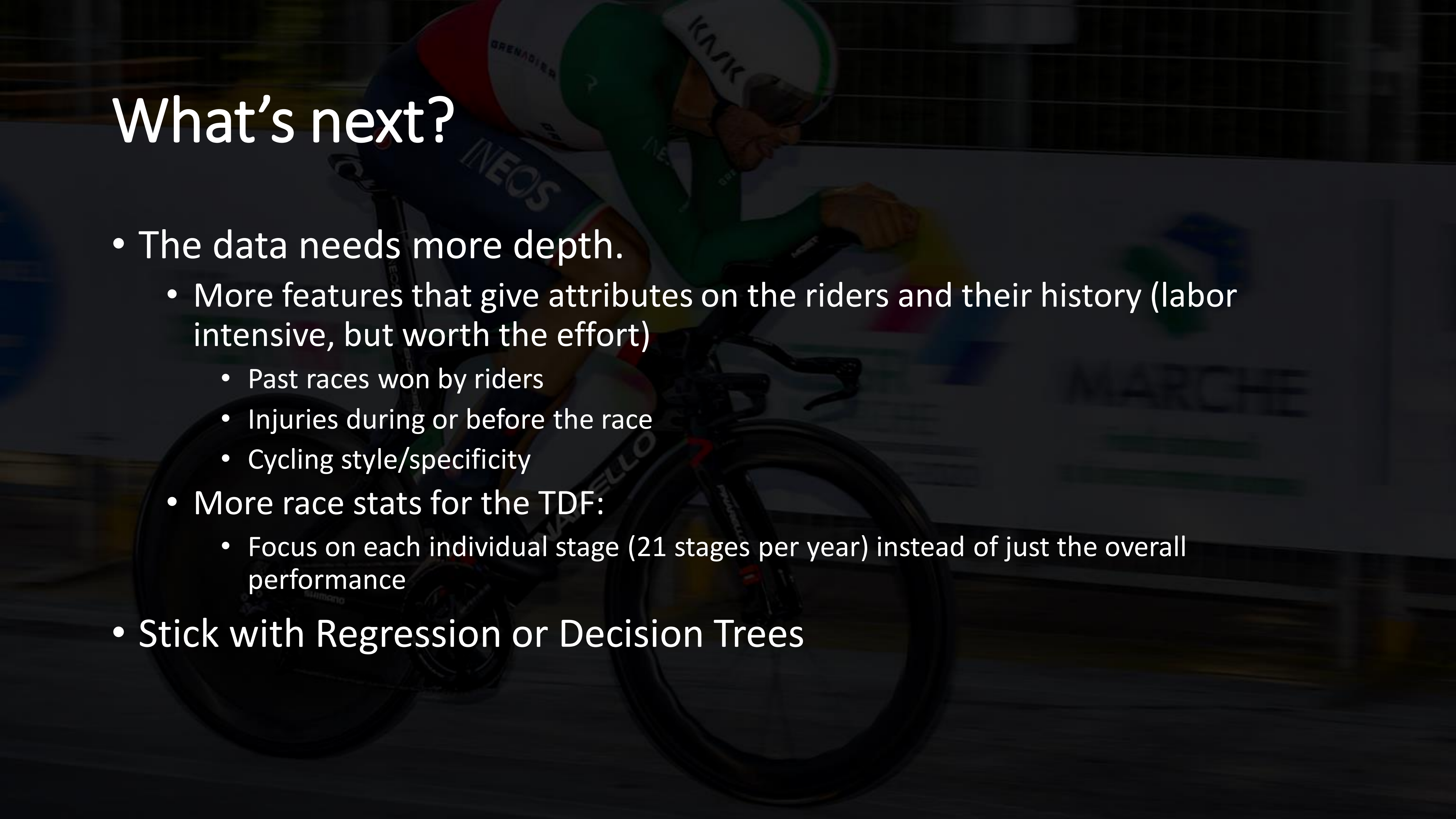
Modelling – Regression

- Linear Regression
  - Train score: 87.26%
  - Test score: 43.56%

Results Summary

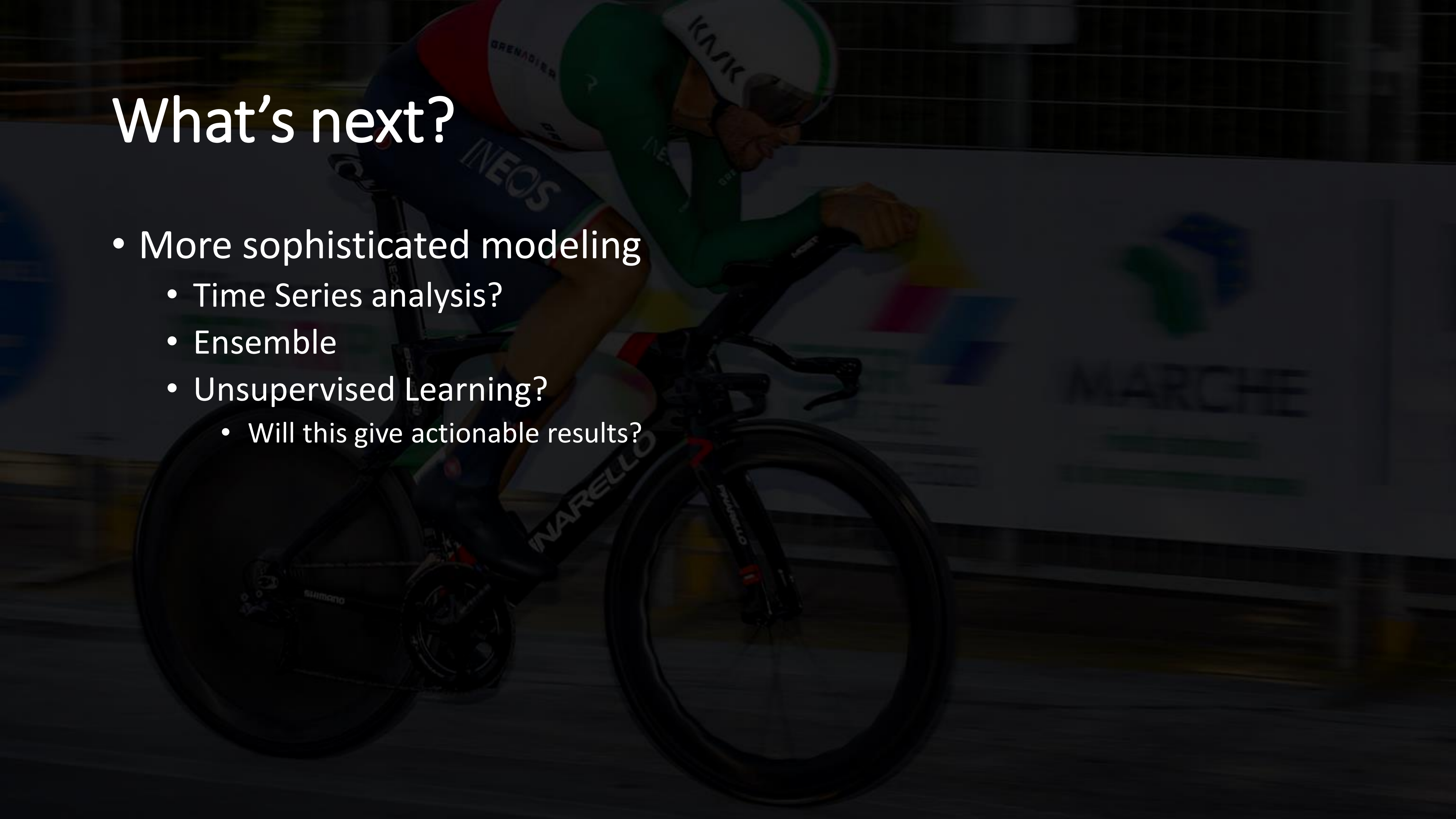
- The classification models are very finicky, and easily give out high scores without providing much actionable results/predictions.
- The regression model while trained well, due to its poor test results, it doesn’t to perform much better than a coin flip.





What’s next?

- The data needs more depth.
  - More features that give attributes on the riders and their history (labor intensive, but worth the effort)
    - Past races won by riders
    - Injuries during or before the race
    - Cycling style/specificity
  - More race stats for the TDF:
    - Focus on each individual stage (21 stages per year) instead of just the overall performance
- Stick with Regression or Decision Trees



What’s next?

- More sophisticated modeling
  - Time Series analysis?
  - Ensemble
  - Unsupervised Learning?
    - Will this give actionable results?

Questions?