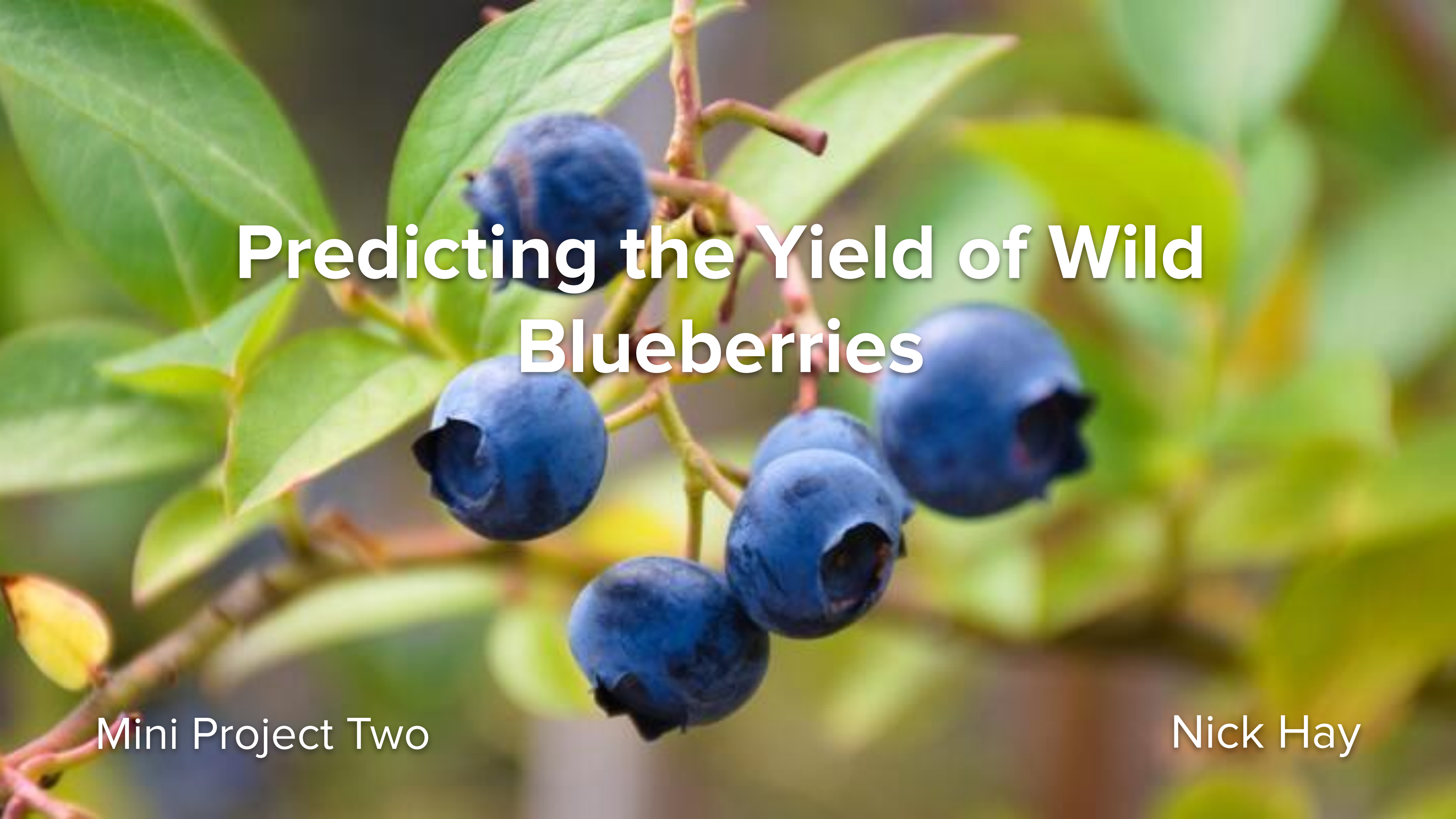


Predicting the Yield of Wild Blueberries

Mini Project Two

Nick Hay

Agenda

Problem Statement

Data Selection

Data Analysis

Model Training

Conclusion

What is the opportunity?

To predict the yield of wild **blueberries** based on a variety of complex variables

The business case (economic value)

Wild blueberries far superior in taste and nutritional content, which can command higher market prices.

Data Question?

What **variables** will influence the yield of wild blueberries and how can we use these to predict future yields.

Data Selection

- Source: kaggle [1]
- Synthetic data
- 18 columns
- No duplicates or null
- Target: **Yield**

DataFrame

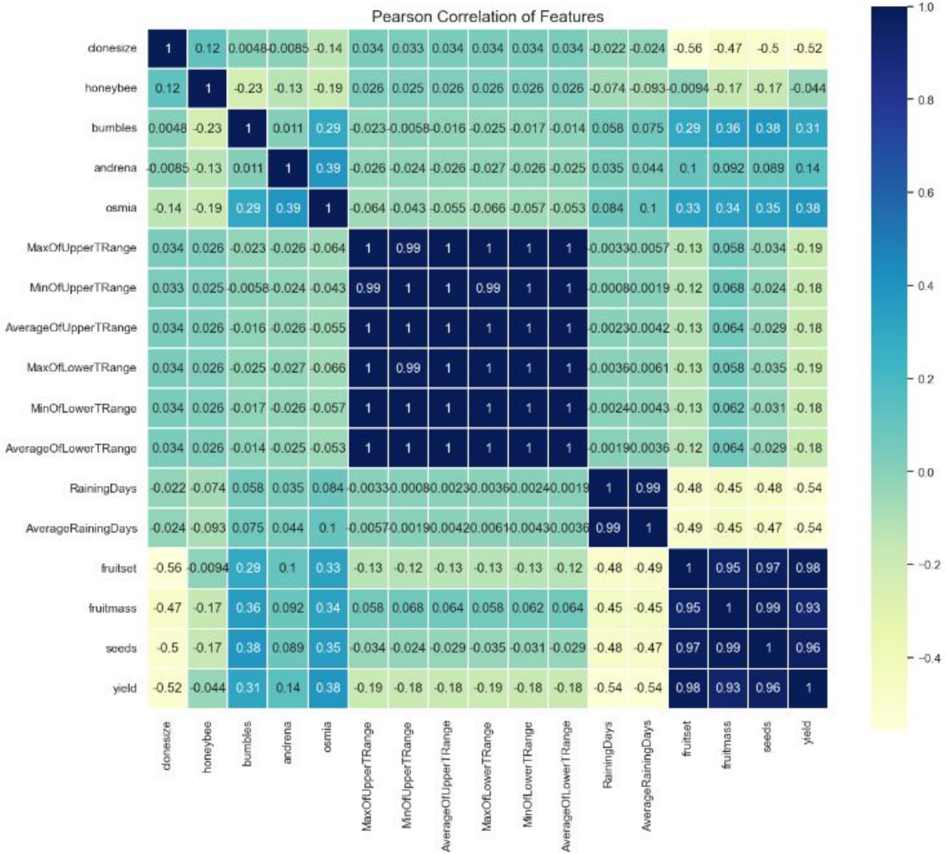
Variable / Features	Description	Unit
Clone size (CS)	The average blueberry clone size in the field	m^2
Honeybee (HB)	Honeybee (*Apis mellifera* (L.)) density in the field	bees/m^2/min
Bumblebee (BB)	Bumblebee (*Bombus* spp.) density in the field	bees/m^2/min
Andrena (AD)	*Andrena* spp. bee density in the field	bees/m^2/min
Osmia (OS)	*Osmia* spp. bee density in the field	bees/m^2/min
MaxOfUpperTRange (MaxUTR)	The highest record of the upper band daily air temperature during the bloom season	°F
MinOfUpperTRange (MinUTR)	The lowest record of the upper band daily air temperature number	°F
AverageOfUpperTRange (AvUTR)	The average of the upper band daily air temperature fare	°F
MaxOfLowerTRange (MaxLTR)	The highest record of the lower band daily air temperature	°F
MinOfLowerTRange (MinLTR)	The lowest record of the lower band daily air temperature	°F
AverageOfLowerTRange (AvLTR)	The average of the lower band daily air temperature	°F
Raining Days (RD)	The total number of days during the bloom season, each of which has precipitation larger than zero	Day
Average Raining Days (AvRD)	The average of raining days of the entire bloom season	Day
Fruitset	The ratio of flowers that successfully develop into fruits	Ratio or %
Fruitmass	The average mass of a single fruit produced	Ounces?
Seeds	The average number of seeds per fruit	Count
Yield	The total harvestable quantity of blueberries per unit area	lb/m^2?

Data Analysis

What methods did you use in your experiment?

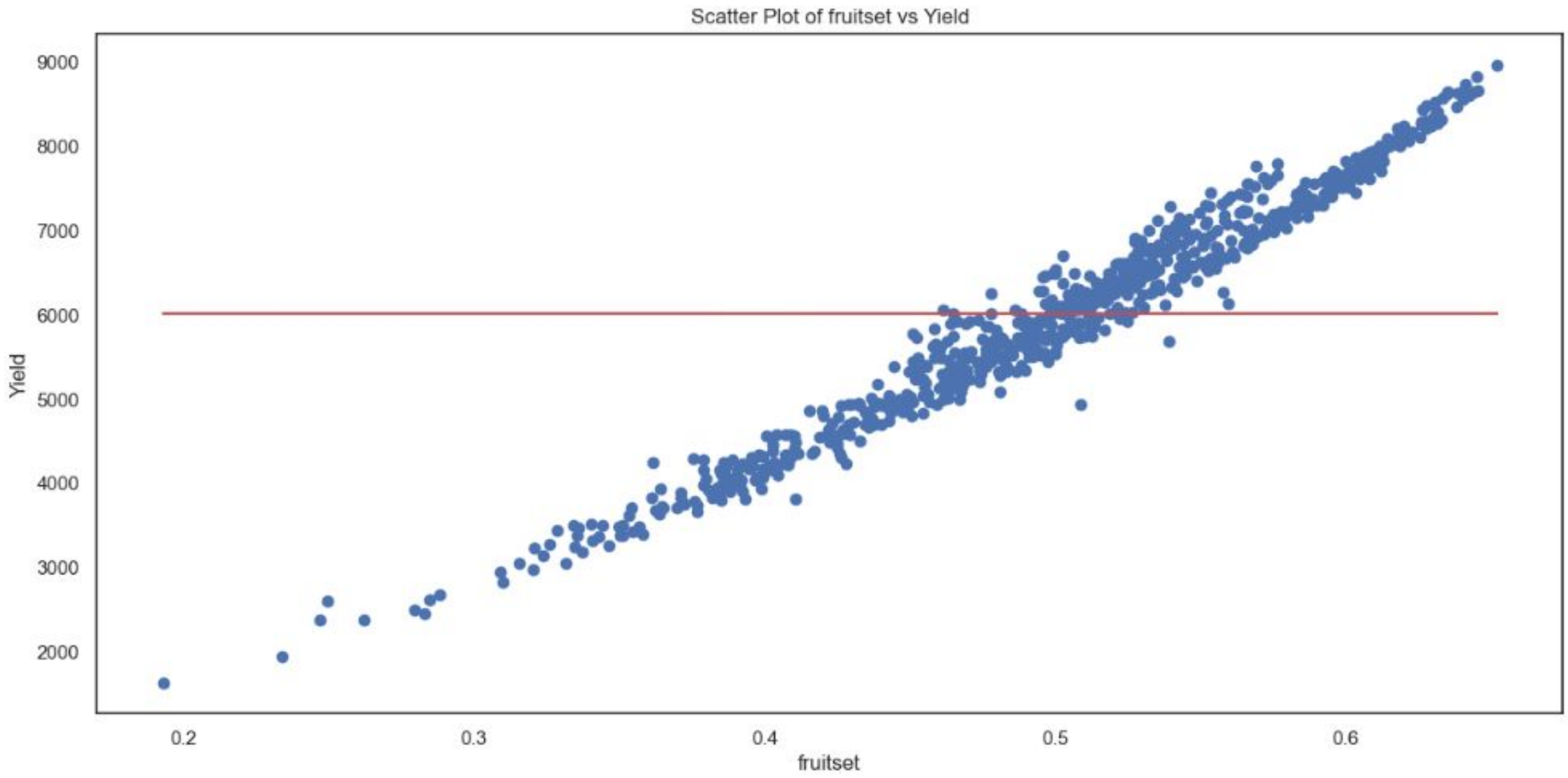
- Pearson Correlation
- EDA
- Feature Selection

Pearson Correlation



	Correlation
AverageRainingDays	-0.541215
RainingDays	-0.540069
clonesize	-0.516737
MaxOfLowerTRange	-0.187439
MaxOfUpperTRange	-0.187075
MinOfLowerTRange	-0.183339
AverageOfUpperTRange	-0.181774
AverageOfLowerTRange	-0.181293
MinOfUpperTRange	-0.175883
honeybee	-0.044250
andrena	0.140277
bumbles	0.309407
osmia	0.380892
fruitmass	0.930365
seeds	0.961249
fruitset	0.984081
yield	1.000000

EDA - Correlation between Fruitset & Yield



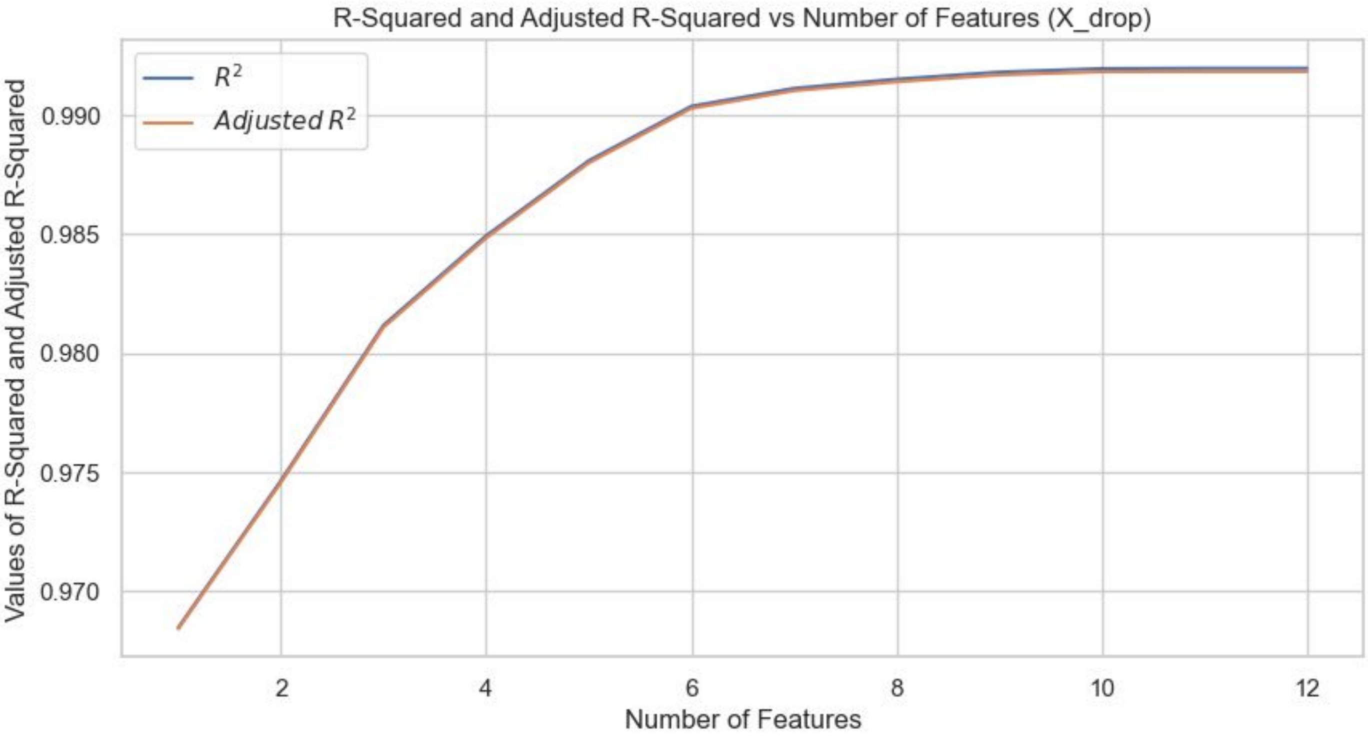
Feature Selection

```
Added feature fruitset with R^2 = 0.968 and adjusted R^2 = 0.968
Added feature RainingDays with R^2 = 0.975 and adjusted R^2 = 0.975
Added feature osmia with R^2 = 0.981 and adjusted R^2 = 0.981
Added feature AverageOfUpperTRange with R^2 = 0.985 and adjusted R^2 = 0.985
Added feature seeds with R^2 = 0.988 and adjusted R^2 = 0.988
Added feature fruitmass with R^2 = 0.990 and adjusted R^2 = 0.990
Added feature andrena with R^2 = 0.991 and adjusted R^2 = 0.991
Added feature honeybee with R^2 = 0.992 and adjusted R^2 = 0.991
Added feature clonesize with R^2 = 0.992 and adjusted R^2 = 0.992
Added feature AverageRainingDays with R^2 = 0.992 and adjusted R^2 = 0.992
Added feature AverageOfLowerTRange with R^2 = 0.992 and adjusted R^2 = 0.992
**************************************************

Resulting features:
fruitset, RainingDays, osmia, AverageOfUpperTRange, seeds, fruitmass, andrena, honeybee, clonesize, AverageRainingDays, Average
OfLowerTRange
```

These **11 features** would give the best prediction

Feature Selection



Modelling

- Multiple Linear Regression
- Ridge Regression
- Lasso Regression

Results

```
Multiple Linear Regression for 11 features  Metrics:
R-squared (R²): 0.9919746420219046
```

```
Ridge Regression for 11 features Metrics:
R-squared (R²): 0.9906832500974772
```

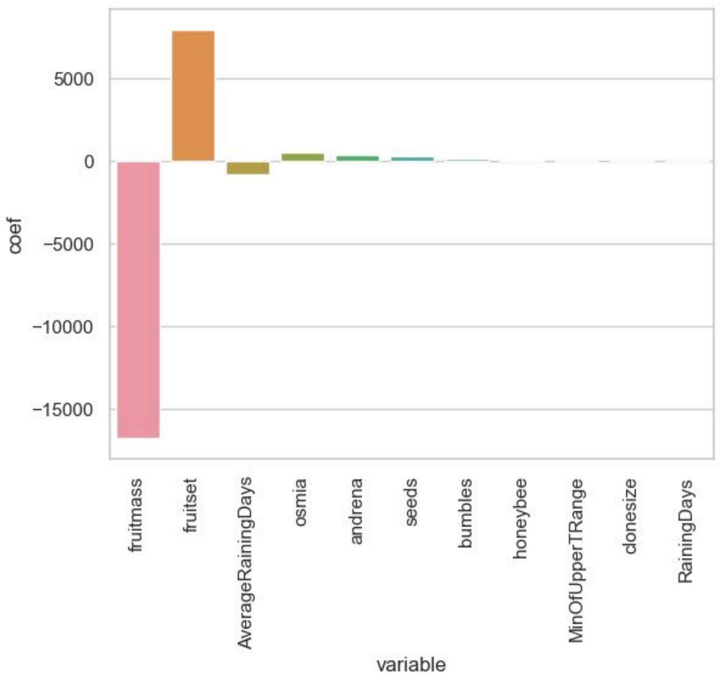
```
Lasso Regression for 11 features Metrics:
R-squared (R²): 0.9908032850488285
```

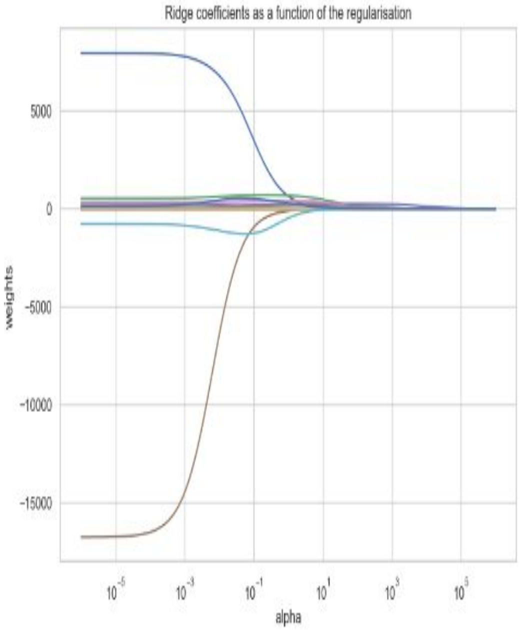
```
Multiple Linear Regression for all 16 features Metrics:
R-squared (R²): 0.9913733793291577
```

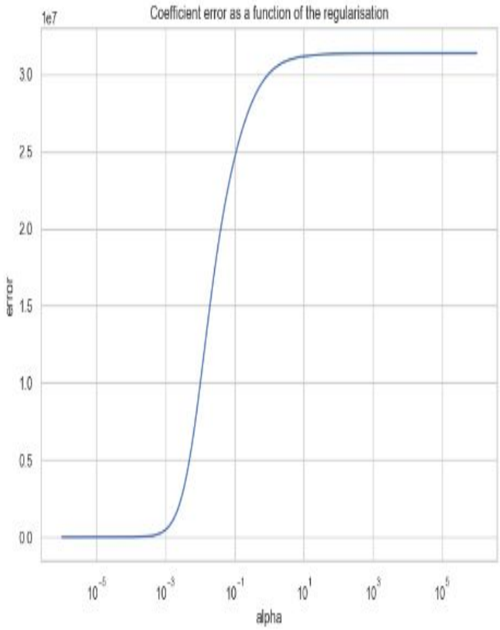
R-squared (Fruitset)

```
0.9684132908271317
```

Ridge Regression [3]

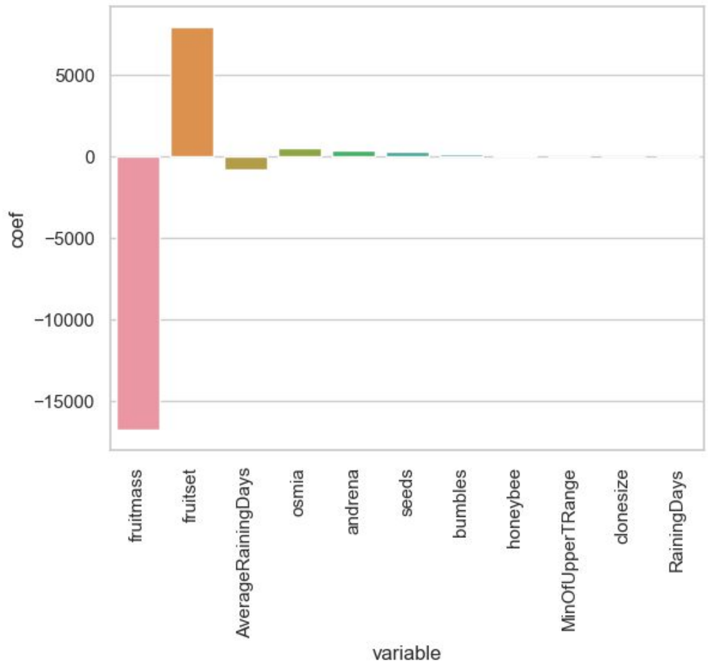


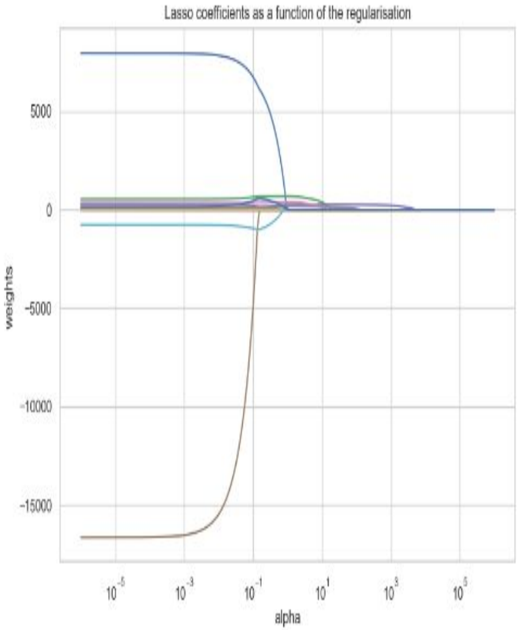


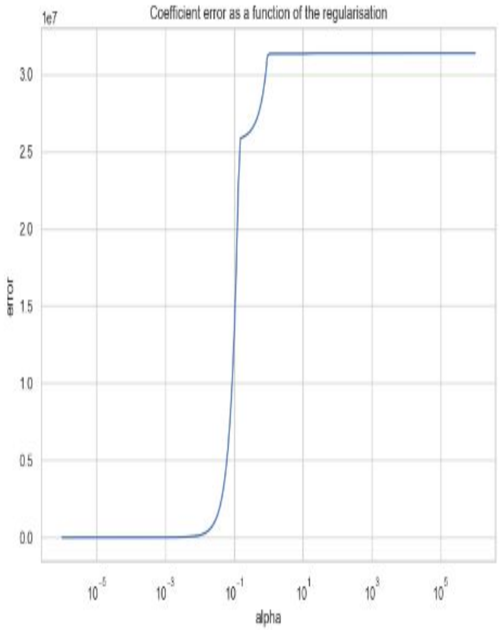


Influence of each variable on the model's predictions

Lasso Regression [4]







Influence of each variable on the model’s predictions

Conclusion

Linear regression is marginally better the Ridge & Lasso

Fruitset as a predictor is ok, using 11 features gives best result

Can **predict** the yield of wildberries using regression models:

Yield Prediction

```
AverageRainingDays = 1
RainingDays = 15
clonesize = 10
fruitmass = 0.384646
seeds = 0.392303
fruitset = 29.742583

# Create a dictionary
data_features = {'AverageRainingDays': [AverageRainingDays],
        'RainingDays': [RainingDays],
        'clonesize': [clonesize],
        'fruitmass': [fruitmass],
        'seeds': [seeds],
        'fruitset': [fruitset]}

# Convert the dictionary to a DataFrame
x_features = pd.DataFrame(data_features)

# Make a prediction
Ypred_features = linreg_features.predict(x_features)
print('Predicted yield: ', Ypred_features[0])

# Predicted yield:  344300.1975937406
```

Next Steps?

- Use other variables (soil temp)
- Compare to cultivated **blueberries**
- Other models (Tree-Based)

References

[1] https://www.kaggle.com/datasets/shashwatwork/wild-blueberry-yield-prediction-dataset

[2] Linear Regression Scikit-Learn: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[3] Ridge Regression Scikit-learn: https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression-and-classification

[4] Lasso Regression Scikit-learn: https://scikit-learn.org/stable/modules/linear_model.html#lasso

[5] IOD Labs 4

[6] Chat GPT and Microsoft Copilot

Questions?