Mini Project 2 *Kevin Koljonen*

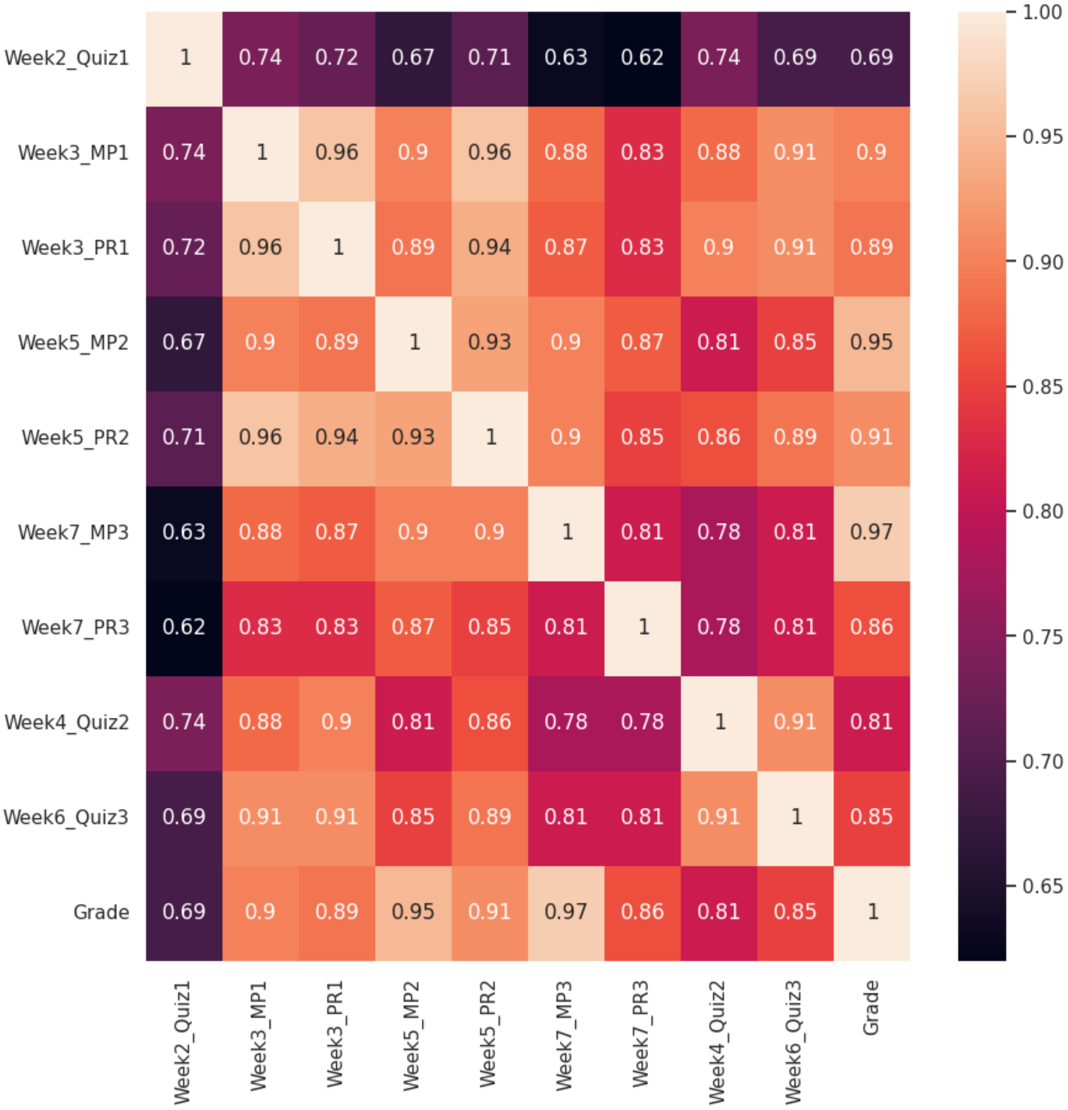
The program made for this project uses 2 supervised learning models, linear regression and ridge regression, to try and predict students' grades using data gathered from a 9-week long course.

1. Data processing

The features provided in the MP2_Data.csv file include data from course content to assignment related statistics. The Week1_Stat1 value only consists of zeroes so we can consider that a missing value. The features I've used in the two models tested are the 9 grade related values due to them having the biggest correlation between them and the grade. The rest of the features have been dropped.

2. Data split

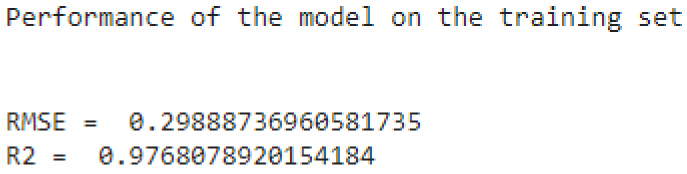
The data has been divided into train and test splits by having a 80% train split and 20% test split with the random state set to 3



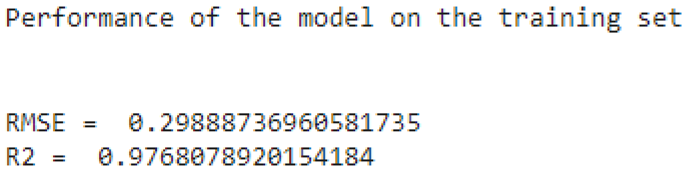
	Week2_Quiz1	Week3_MP1	Week3_PR1	Week5_MP2	Week5_PR2	Week7_MP3	Week7_PR3	Week4_Quiz2	Week6_Quiz3	Grade
Week2_Quiz1	1	0.74	0.72	0.67	0.71	0.63	0.62	0.74	0.69	0.69
Week3_MP1	0.74	1	0.96	0.9	0.96	0.88	0.83	0.88	0.91	0.9
Week3_PR1	0.72	0.96	1	0.89	0.94	0.87	0.83	0.9	0.91	0.89
Week5_MP2	0.67	0.9	0.89	1	0.93	0.9	0.87	0.81	0.85	0.95
Week5_PR2	0.71	0.96	0.94	0.93	1	0.9	0.85	0.86	0.89	0.91
Week7_MP3	0.63	0.88	0.87	0.9	0.9	1	0.81	0.78	0.81	0.97
Week7_PR3	0.62	0.83	0.83	0.87	0.85	0.81	1	0.78	0.81	0.86
Week4_Quiz2	0.74	0.88	0.9	0.81	0.86	0.78	0.78	1	0.91	0.81
Week6_Quiz3	0.69	0.91	0.91	0.85	0.89	0.81	0.81	0.91	1	0.85
Grade	0.69	0.9	0.89	0.95	0.91	0.97	0.86	0.81	0.85	1

Model training and performance evaluation

The models that I've used are linear regression and a ridge regression pipeline. According to the model evaluation RMSE and R2 scores for the training and test set of the linear regression model is:

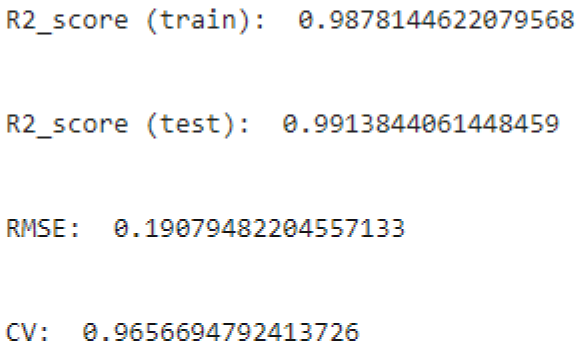


```
Performance of the model on the training set


RMSE =  0.29888736960581735
R2 =  0.9768078920154184
```



```
Performance of the model on the training set


RMSE =  0.29888736960581735
R2 =  0.9768078920154184
```

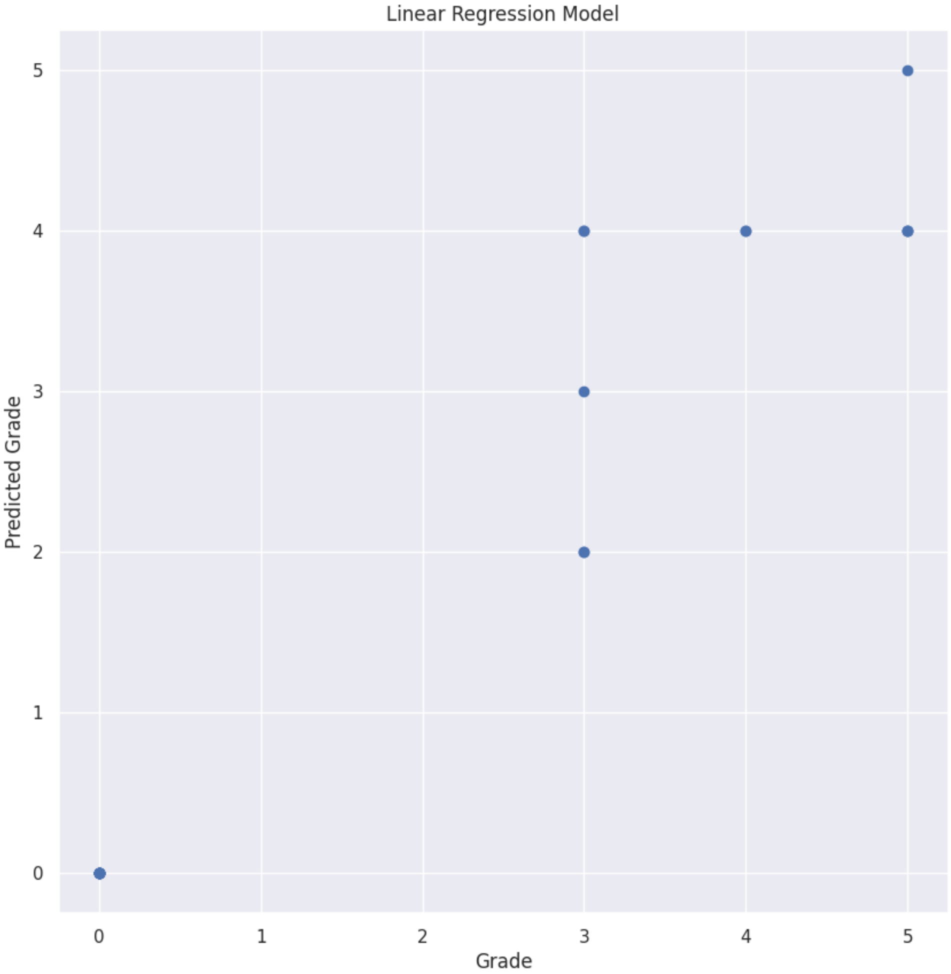
For the ridge regression a cross validation set was also used to check the effectiveness of the model. The RMSE, R2 and the cross validation score of the ridge regression is:



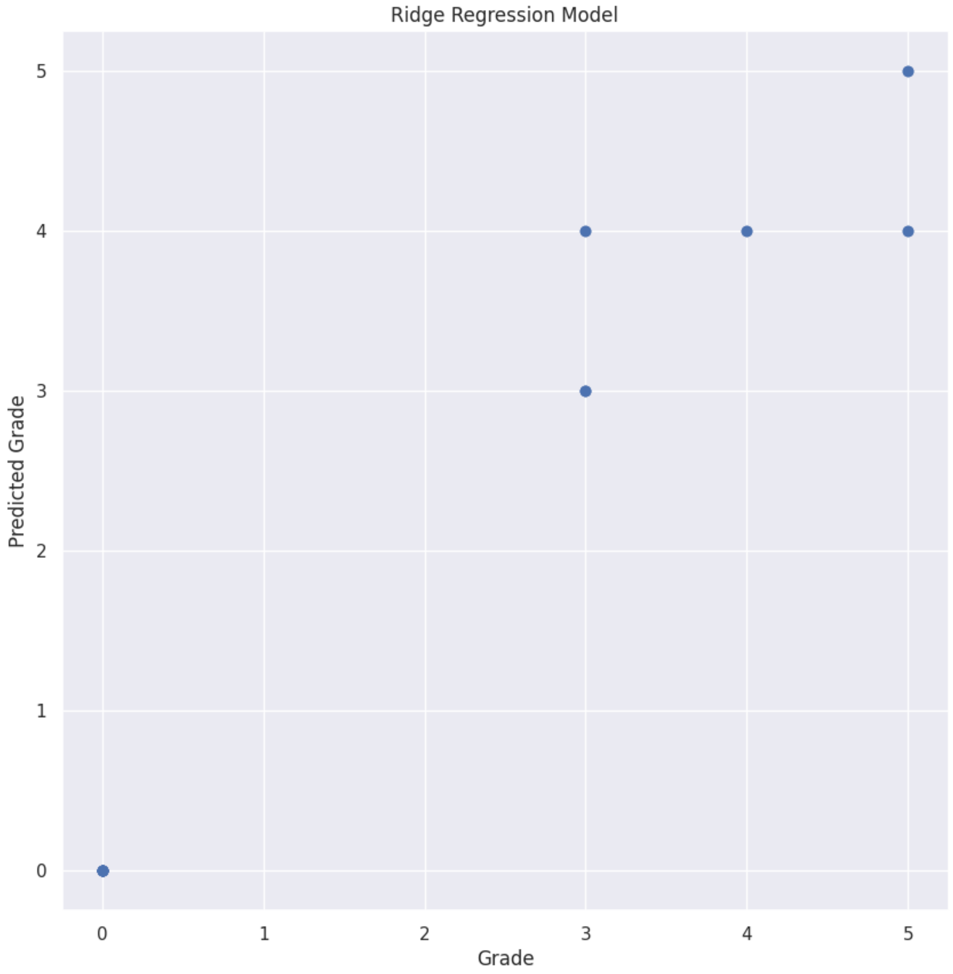
```
R2_score (train):  0.9878144622079568


R2_score (test):  0.9913844061448459


RMSE:  0.19079482204557133


CV:  0.9656694792413726
```

According to both of the models RMSE and R2 scores the ridge regression model is better with a lower RMSE score and a higher R2 score.

How the linear regression model predicted the students’ grade:



How the ridge regression model predicted the students' grade:



5. Important features

According to the correlation heatmap made in the code, the three most important features for predicting a grade are Week7_MP3, Week5_MP2 and Week3_MP1. Although Week8_Total also has the biggest correlation with the grade it is in itself the score for the grade which is why I did not use it.

Conclusion

In this project the ridge regression model was the better model to predict students' grades. Finding and figuring out the ridge regression pipeline was one of the "scientific" bottlenecks that I faced but that was solved by googling and not reinventing the wheel. Overall the models aren't perfect but they have a relative accuracy in predicting grades.