In-Depth Analysis Report

Introduction

This capstone project seeks to identify additional indicators of a country's World Happiness Score. The factors investigated in this report are: gender disparities (development and inequality), economic freedom and population density. Outlined below, are the steps taken during machine learning for the target and features listed above. Please see the accompanying jupyter notebook for a detailed overview of the code, analysis and plots referenced below.

Preparing the Dataset

During EDA, it was noted countries of the same geographical regions had similar happiness scores. It would be advantageous to include regions in the feature set of the regression models. The regions included in the dataset are:

- Southern Europe
- Northern Africa
- Sub-Saharan Africa
- Western Asia
- Australia and New Zealand
- Western Europe
- Southern Asia
- Eastern Europe
- Latin America and the Caribbean
- South-eastern Asia
- Northern America
- Eastern Asia
- Northern Europe
- Central Asia

Linear regression models do not allow categorical data. In order to include regions, one hot encoding was utilized. Pandas .get_dummies() method was applied to create a column for each region and a value of 1 if the country was in the region and 0 if it was not.

	Country	Region	iso_a3	wh_score	gdi	gii	efs	pdn	year	region_1	region_2	region_3
0	Albania	Southern Europe	ALB	4.655	0.97	0.256	65.650	105.135	2015	1	0	0
1	Algeria	Northern Africa	DZA	6.355	0.858	0.442	48.881	16.680	2015	0	1	0
2	Angola	Sub-Saharan Africa	AGO	3.865	0.838	0.575	47.88	22.366	2015	0	0	1

Figure 1: Truncated example of dataset after One Hot Encoding

Feature Selection

Python module statsmodels was used to determine which features to include in the models. The summary method provides result statistics for each target/feature combination:

- R2: a measure of the proportion of the variance for the target variable that can be explained by the feature(s). A higher R2 is better
- P-value: the probability of obtaining test results at least as extreme as the ones observed. A p-value of less than 0.05 provides confidence that a feature has statistical significance
- AIC: provides an estimate on the quality of the model relative to the other models. A lower AIC indicates a better fitting model.

Based on the statistics above, it appears it would be beneficial to remove population density from the feature set. I chose not to include population density in the model because when grouped with the other features, it's p-value was higher than 0.05. Additionally, the AIC is higher, albeit not by much, than the model that omits population density. The features that will be included in the models are: gender development, gender inequality, economic freedom, and regions. Since the regions are a categorical feature, their p-value is not evaluated for significance. Another interesting observation is that the coef for gdi is negative when grouped with the other features.

	wh_score ~ gdi	wh_score ~ gii	wh_score ~ efs	wh_score ~ pdn	wh_score ~ region_x	wh_score ~ gdi+gii+efs+ region_x
R2	0.188	0.557	0.402	0.009	0.68	0.77
Adjusted R2	0.186	0.556	0.401	0.007	0.672	0.763
F-statistic	126.1	686.2	367.3	4.912	87.26	111.3
AIC	1565	1232	1397	1674	1078	902.2
coef	5.6736	-4.3129	0.0716	0.0001		
std err	0.505	0.165	0.004	6.75E-05		
P>|t|	0	0	0	0.027		

Figure 2: Key result statistics for target/feature combinations

OLS Regression Results

Dep. Variable:	wh_score	R-squared:	0.770
Model:	OLS	Adj. R-squared:	0.763
Method:	Least Squares	F-statistic:	111.3
Date:	Thu, 12 Mar 2020	Prob (F-statistic):	7.76e-158
Time:	07:54:11	Log-Likelihood:	-434.12
No. Observations:	548	AIC:	902.2
Df Residuals:	531	BIC:	975.5
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.5399	0.444	14.739	0.000	5.668	7.412
gdi	-1.0902	0.361	-3.021	0.003	-1.799	-0.381
gii	-3.6430	0.332	-10.961	0.000	-4.296	-2.990
efs	0.0131	0.004	3.641	0.000	0.006	0.020
region_1	-0.2579	0.087	-2.967	0.003	-0.429	-0.087
region_2	0.2653	0.142	1.869	0.062	-0.014	0.544
region_3	-0.0656	0.110	-0.594	0.553	-0.283	0.151
region_4	0.3440	0.084	4.075	0.000	0.178	0.510
region_5	1.1958	0.194	6.153	0.000	0.814	1.578
region_6	0.8465	0.115	7.384	0.000	0.621	1.072
region_7	0.1725	0.120	1.433	0.153	-0.064	0.409
region_8	0.1505	0.093	1.616	0.107	-0.032	0.334
region_9	1.2339	0.095	12.989	0.000	1.047	1.421
region_10	0.4259	0.112	3.800	0.000	0.206	0.646
region_11	1.2265	0.191	6.437	0.000	0.852	1.601
region_12	-0.2072	0.134	-1.548	0.122	-0.470	0.056
region_13	0.7753	0.103	7.559	0.000	0.574	0.977
region_14	0.4344	0.133	3.273	0.001	0.174	0.695

Omnibus:	16.383	Durbin-Watson:	2.138
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17.503
Skew:	-0.382	Prob(JB):	0.000158
Kurtosis:	3.427	Cond. No.	7.96e+17

Figure 3: Statsmodel summary results for selected model features (excluding Population Density)

OLS Regression Results

Dep. Variable:	wh_score	R-squared:	0.771
Model:	OLS	Adj. R-squared:	0.763
Method:	Least Squares	F-statistic:	104.7
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	6.37e-157
Time:	08:12:39	Log-Likelihood:	-433.84
No. Observations:	548	AIC:	903.7
Df Residuals:	530	BIC:	981.2
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.5463	0.444	14.744	0.000	5.674	7.419
gdi	-1.1174	0.363	-3.079	0.002	-1.830	-0.404
gii	-3.7012	0.342	-10.828	0.000	-4.373	-3.030
efs	0.0137	0.004	3.710	0.000	0.006	0.021
pdn	-2.84e-05	3.86e-05	-0.735	0.463	-0.000	4.75e-05
region_1	-0.2637	0.087	-3.021	0.003	-0.435	-0.092
region_2	0.2739	0.142	1.922	0.055	-0.006	0.554
region_3	-0.0472	0.113	-0.417	0.677	-0.270	0.175
region_4	0.3466	0.085	4.101	0.000	0.181	0.513
region_5	1.1718	0.197	5.943	0.000	0.784	1.559
region_6	0.8317	0.116	7.143	0.000	0.603	1.060
region_7	0.1942	0.124	1.566	0.118	-0.049	0.438
region_8	0.1473	0.093	1.578	0.115	-0.036	0.331
region_9	1.2436	0.096	12.961	0.000	1.055	1.432
region_10	0.4588	0.121	3.800	0.000	0.222	0.696
region_11	1.2079	0.192	6.281	0.000	0.830	1.586
region_12	-0.2108	0.134	-1.574	0.116	-0.474	0.052
region_13	0.7559	0.106	7.136	0.000	0.548	0.964
region_14	0.4364	0.133	3.286	0.001	0.176	0.697

Omnibus:	16.242	Durbin-Watson:	2.128
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17.460
Skew:	-0.376	Prob(JB):	0.000162
Kurtosis:	3.447	Cond. No.	8.88e+18

Figure 4: Statsmodel summary results for selected model features (including Population Density)

Fit Diagnostics

Statsmodel was used again to obtain the residuals and plot a residual and probability plot. The residual plot is fairly symmetrically distributed around 0. Both the probability plot and the residual histogram show that the data is normally distributed. Therefore, a linear regression model appears to be a good fit for this data.

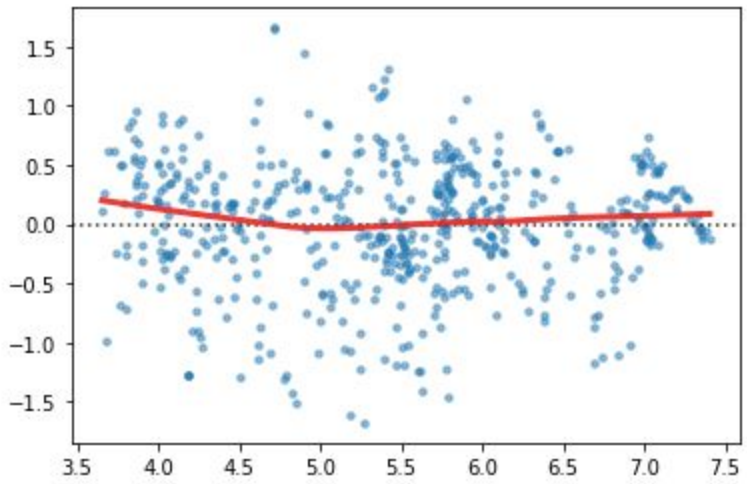

Figure 5: Residual plot: fitted values vs. residuals

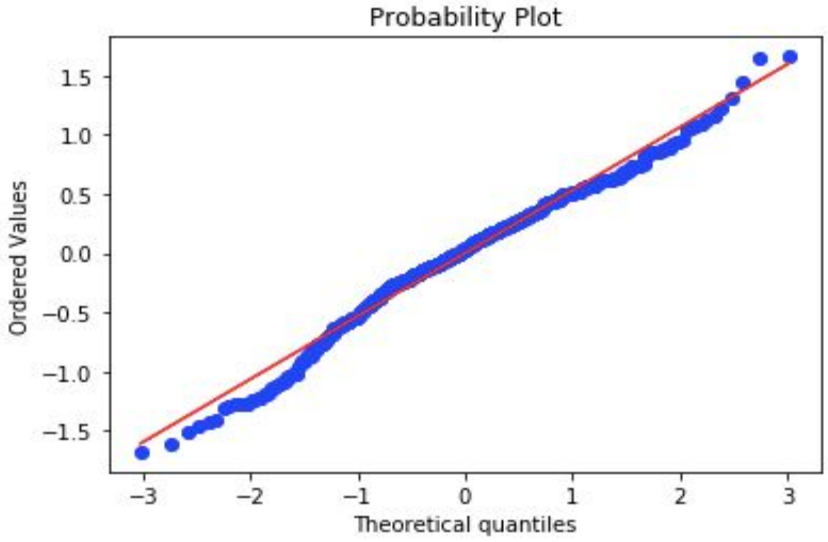

Figure 6: Probability plot

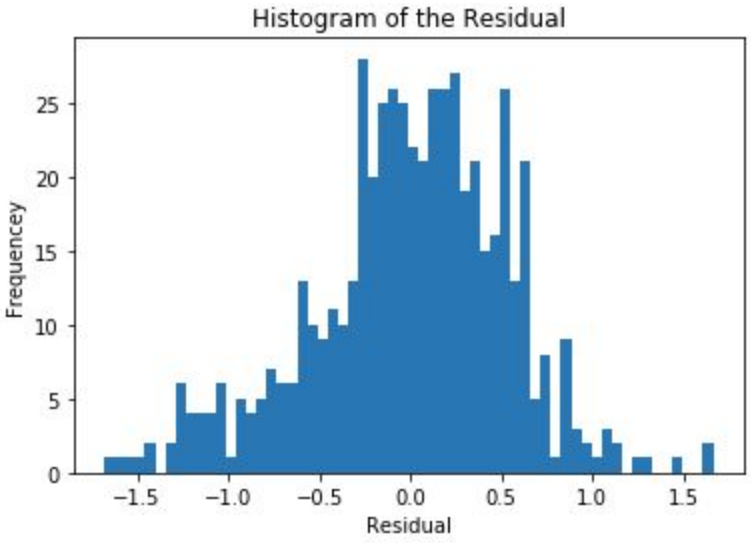

Figure 7: Histogram of the Residual

Modeling

I used the train_test_split() function to split my data into a training set and a testing set. 5 models were applied to the training data: Linear Regression, Ridge Regression, Lasso Regression, K-Nearest Neighbor Regressor, and Random Forest Regressor. Hyper-parameters' performance was selected and cross-validated using GridSearchCV. R2, mean squared error, and root mean squared error where the metrics used to evaluate the regression models against the training dataset and the testing dataset as well as check for overfitting.

```
Ridge Best params: {'ridge__alpha': 0.01, 'ridge__fit_intercept': True, 'ridge__normalize': True}
Ridge Best score: 0.7539017750475305
```

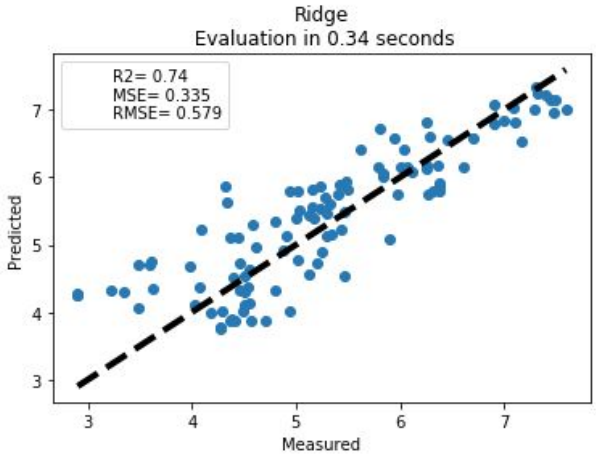

Figure 8: Ridge Regression Predicted vs. Actual Plot

```
Linear Best params: {'ols__fit_intercept': True, 'ols__normalize': True}
Linear Best score: 0.7537115490557087
```

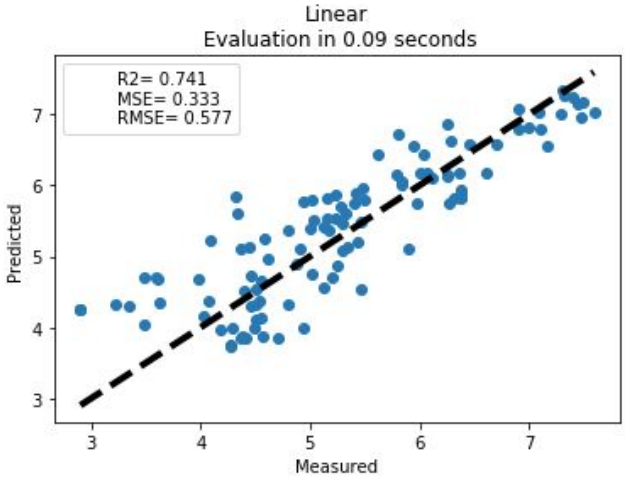

Figure 9: Linear Regression Predicted vs. Actual Plot

```
Lasso Best params: {'lasso__alpha': 0.001, 'lasso__fit_intercept': True, 'lasso__normalize': False}
Lasso Best score: 0.7537050651124357
```

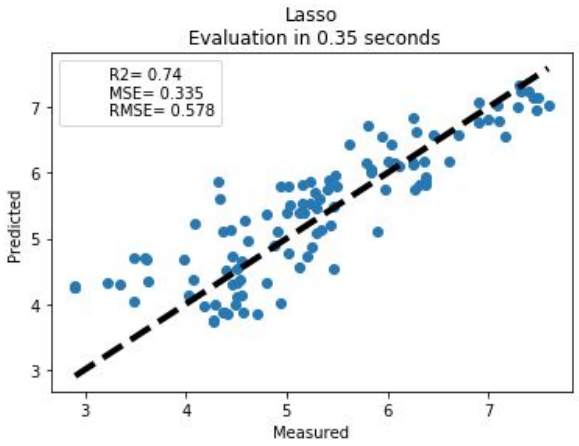

Figure 10: Lasso Regression Predicted vs. Actual Plot

```
KNN Best params: {'knn__n_neighbors': 2, 'knn__weights': 'distance'}
KNN Best score: 0.9049970413478693
```

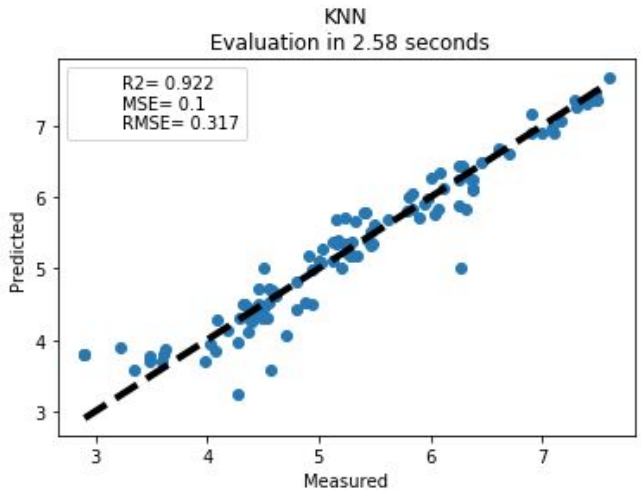

Figure 11: K-Nearest Neighbors Regressor Predicted vs. Actual Plot

```
RandomForest Best params: {'rf__criterion': 'mse', 'rf__max_depth': 90, 'rf__n_estimators': 70}
RandomForest Best score: 0.8757049990150947
```

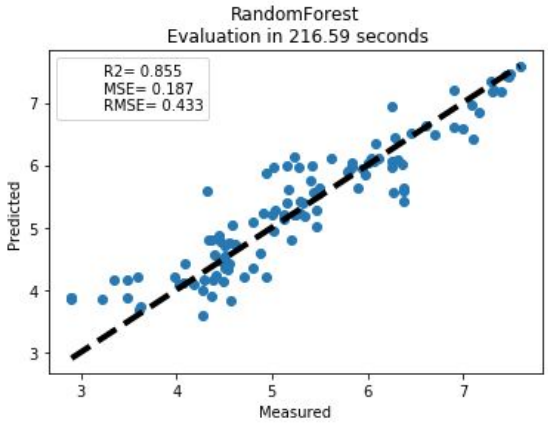

Figure 12: Random Forest Regressor Predicted vs. Actual Plot

Model Analysis

In all of the models, the delta between the R2 of the training data and the R2 of the testing data is less than 10%, therefore overfitting is not an issue. Random Forest performed well, but it's execution time is significantly higher than the other models and not ideal from a business perspective.

Conclusion

K-Nearest Neighbors Regressor model was selected for this dataset.
The model was selected based on the following criteria:

1. No issue with overfitting
2. Highest performance metric (R2)
3. Relatively low level of complexity