Supervised learning capstone

Predicting the trade volumes between countries by using past years data

Overview and problem statement International trade

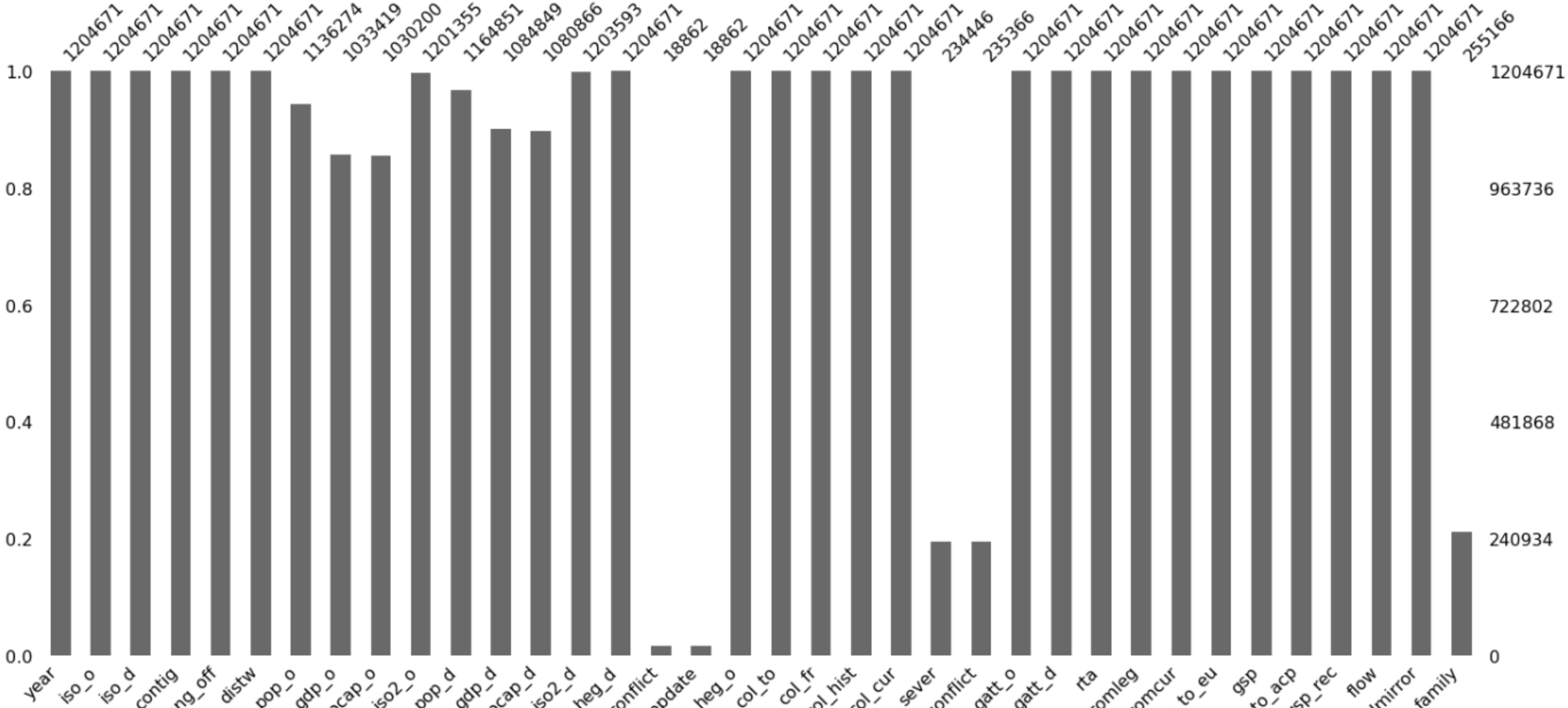
- Countries and stakeholders can benefit from. Such as Increased revenues, Decreased competition, Longer product lifespan, Easier cash-flow management, Better risk management, Benefiting from currency exchange, Access to export financing, and Disposal of surplus goods.

- Problem : shipping customs and duties, language barriers, cultural differences, servicing customers, returning products, and intellectual property theft

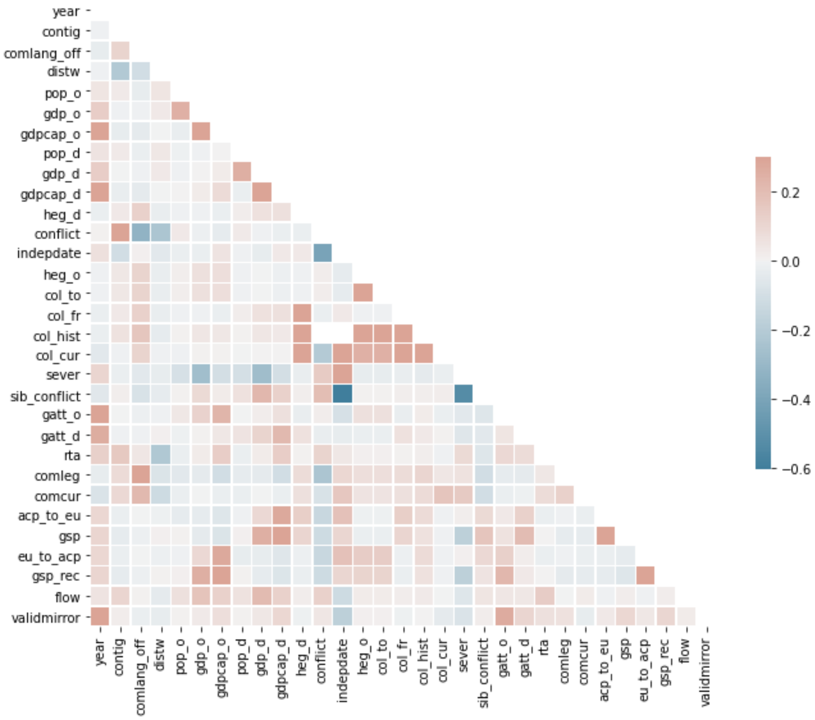
Data

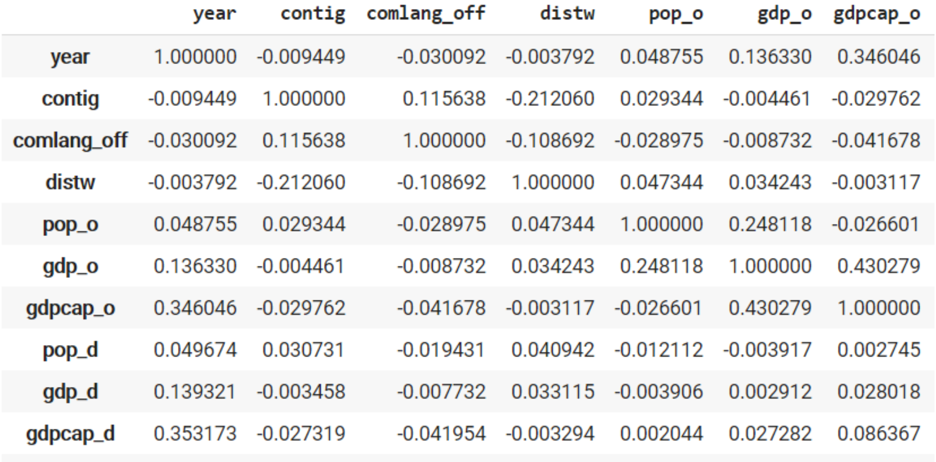
- This data is from all world pairs of countries (224), for the period 1948 to 2015. It would be safe to generalize the results to all counties in the data set.
- Total number of rows in dataset = 1204671 Total number of columns in dataset = 36
- Data includes gdp of origin and destination counries , population , amount of trade between countries etc.
- Raw data available here

Variables with missing data



Correlation between variables





	year	contig	comlang_off	distw	pop_o	gdp_o	gdpcap_o
year	1.000000	-0.009449	-0.030092	-0.003792	0.048755	0.136330	0.346046
contig	-0.009449	1.000000	0.115638	-0.212060	0.029344	-0.004461	-0.029762
comlang_off	-0.030092	0.115638	1.000000	-0.108692	-0.028975	-0.008732	-0.041678
distw	-0.003792	-0.212060	-0.108692	1.000000	0.047344	0.034243	-0.003117
pop_o	0.048755	0.029344	-0.028975	0.047344	1.000000	0.248118	-0.026601
gdp_o	0.136330	-0.004461	-0.008732	0.034243	0.248118	1.000000	0.430279
gdpcap_o	0.346046	-0.029762	-0.041678	-0.003117	-0.026601	0.430279	1.000000
pop_d	0.049674	0.030731	-0.019431	0.040942	-0.012112	-0.003917	0.002745
gdp_d	0.139321	-0.003458	-0.007732	0.033115	-0.003906	0.002912	0.028018
gdpcap_d	0.353173	-0.027319	-0.041954	-0.003294	0.002044	0.027282	0.086367

Data cleaning

- Dropping the features with more than 50 percent missing values.
- Dropping the categorical variables that have less impact on the target variable GDP of the origin country and not important in this project.
- Replacing missing values with simple imputer

Regression Models

- ## Decision Tree

Is a simple machine learning model for getting started with regression tasks. A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node (see here for more details(see here for more details).

- ## CatBoost

Is a recently open-sourced machine learning algorithm from Yandex. It can easily integrate with deep learning frameworks like Google's TensorFlow and Apple's Core ML. It can work with diverse data types to help solve a wide range of problems that businesses face today(see here for more details).

- ## Random Forest

Is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. (see here for more details).

- ## Linear Regression

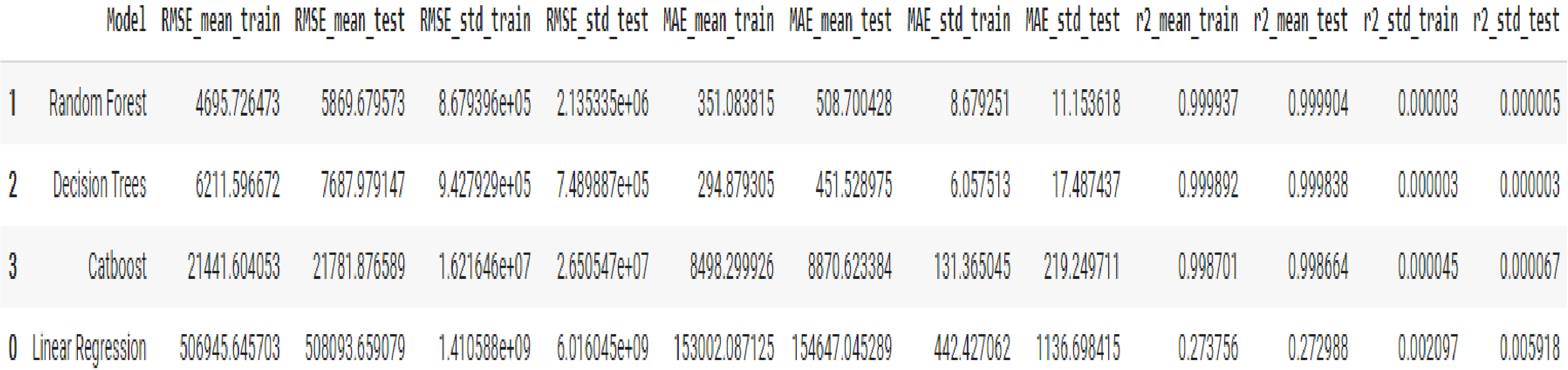
fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

Comparing Models

- Cross validation for regression models with train and test data

- RMSE, MAE and R squared calculated

Comparison results



	Model	RMSE_mean_train	RMSE_mean_test	RMSE_std_train	RMSE_std_test	MAE_mean_train	MAE_mean_test	MAE_std_train	MAE_std_test	r2_mean_train	r2_mean_test	r2_std_train	r2_std_test
1	Random Forest	4695.726473	5869.679573	8.679396e+05	2.135335e+06	351.083815	508.700428	8.679251	11.153618	0.999937	0.999904	0.000003	0.000005
2	Decision Trees	6211.596672	7687.979147	9.427929e+05	7.489887e+05	294.879305	451.528975	6.057513	17.487437	0.999892	0.999838	0.000003	0.000003
3	Catboost	21441.604053	21781.876589	1.621646e+07	2.650547e+07	8498.299926	8870.623384	131.365045	219.249711	0.998701	0.998664	0.000045	0.000067
0	Linear Regression	506945.645703	508093.659079	1.410588e+09	6.016045e+09	153002.087125	154647.045289	442.427062	1136.698415	0.273756	0.272988	0.002097	0.005918

Future improvements

- **Hyperparameter tuning for machine learning models**
- Compress features, reduce overfitting and noise and increase efficiency and performance