CSS 429 Data Analysis Systems

Final project

Team 7 - Abraca-data:

- Muratbekov Maksatbek
- Sarkulov Yerden
- Tolbassy Bakdaulet
- Yegemberdi Nurgeldi
- Zhakypov Aslan

BUSINESS UNDERSTANDING

- **Identify, define, and motivate the business problem that you are addressing.**

  Suppose our team works as the Agency for Economic Development of Kazakhstan. Hence, it is important for us to forecast GDP per capita in order to develop the country's economy and plan for the next year in various economic sectors of the country. ***GDP per capita*** *is an important indicator of economic performance and a useful unit to make cross-country comparisons of average living standards and economic wellbeing.*

- **How (precisely) will a data mining solution address the business problem?**

  By predicting GDP and comparing our economic ties with neighboring countries, we can make a development plan for the future. In addition, by analyzing the data, we can learn from rich countries and move to their economic direction.

DATA UNDERSTANDING

- Identify and describe the data (and data sources) that will support data mining to address the business problem. Include those aspects of the data that we routinely talk about in class and/or in the assignments. (If appropriate highlight potential bias, full disclosure always better, and identify potential bias directions.)

  We used the Countries of the World dataset from Kaggle (link: https://www.kaggle.com/fernandol/countries-of-the-world)

  Firstly, we apologize for using old data in final project. We searched data of countries from Kaggle and it was corresponding despite being old. Actually, data was taken from 2007, but last updated 3 years ago.

  Data includes:

  - *Country name*

- *Region*
- *Population*
- *Area*
- *Pop. Density (per square meter)*
- *Coastline (coast/area ratio)*
- *Net migration*
- *Infant mortality (per 1000 births)*
- *GDP ($ per capita)*
- *Literacy (%)*
- *Phones (per 1000 people)*
- *Arable (%)*
- *Crops (%)*
- *Other (%)*
- *Climate*
- *Birthrate*
- *Deathrate*
- *Agriculture*
- *Industry*
- *Service*

DATA PREPARATION

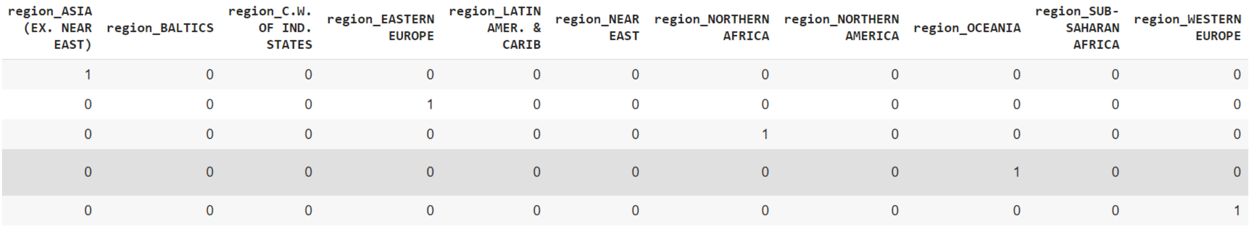
- **Specify how these data are integrated to produce the format required for data mining.**

  Data cleaning and inputting missing values:
  Firstly, we fixed column names removing brackets and also rephrasing to have shorter column names. After that, changed object type of all features to float except `country` and `region` as they converted to categorical type. Then, we found out missing values. Firstly, we searched it from `gdp_per_capita` as it is target variable. There was only one country (Western Sahara) with missed GDP per Capita, we input 2500$ using Google search.

  We replaced nulls with 0 for `net_migration`, `infant_mortality`, `arable`, `crops`, `other` and `climate` as these columns have not so many nulls. However, we input ***mean*** *and* ***median*** *for others.*

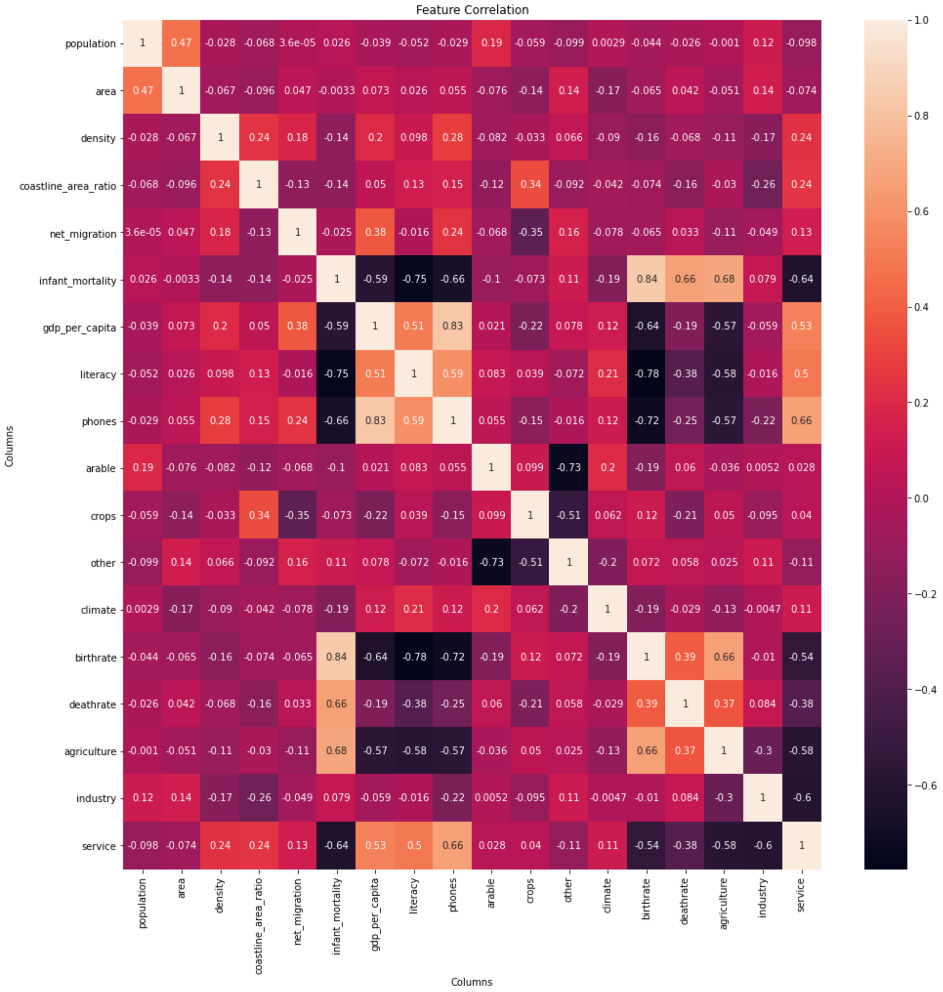
  Agriculture and service have 15 null values while industry has 16 null values. All belong to very small island nations. Those kind of nations usually have economies that rely heavily on services, with some agricultural and industrial *activities. So we replaced the missing values with the following: agriculture = 0.15, industry = 0.05. service = 0.8. For Monaco, we set the value for industry and service to be 0.05 and 0.78 respectively. For Western Sahara, we set the value for agriculture and industry to be 0.35 and 0.25 respectively*

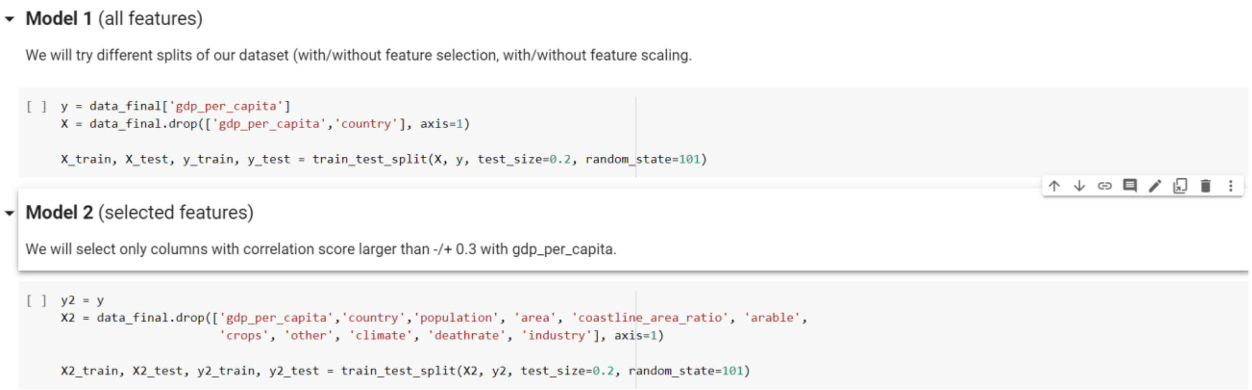
If you remember, there were two categorical variables, so we transformed `region` to numerical, while dropping `country` column.



region_ASIA (EX. NEAR EAST)	region_BALTICS	region_C.W. OF IND. STATES	region_EASTERN EUROPE	region_LATIN AMER. & CARIB	region_NEAR EAST	region_NORTHERN AFRICA	region_NORTHERN AMERICA	region_OCEANIA	region_SUB-SAHARAN AFRICA	region_WESTERN EUROPE
1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	1

Then we split our data to train and test creating two types of models, first one was created with all features, when second with selected features (if their correlation with GDP > 0.7)





Model 1 (all features)

We will try different splits of our dataset (with/without feature selection, with/without feature scaling.

```
[ ] y = data_final['gdp_per_capita']
    X = data_final.drop(['gdp_per_capita','country'], axis=1)

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=101)
```

Model 2 (selected features)

We will select only columns with correlation score larger than -/+ 0.3 with gdp_per_capita.

```
[ ] y2 = y
    X2 = data_final.drop(['gdp_per_capita','country','population', 'area', 'coastline_area_ratio', 'arable',
                          'crops', 'other', 'climate', 'deathrate', 'industry'], axis=1)

    X2_train, X2_test, y2_train, y2_test = train_test_split(X2, y2, test_size=0.2, random_state=101)
```

MODELING

- **Specify the type of model(s) built and/or patterns mined.**

  We used three regressions to predict our target variable. They were Linear Regression, Lasso Regression and Random Forest.

- **Discuss choices for data mining algorithm: what are alternatives, and what are the pros and cons?**

  Linear Regression:
  This was our baseline model, we used the result as a reference to compare with others. We could see that most features do not have a linear relationship with our labels (gdp_per_capita), yet we tried linear regression.

  Lasso Regression:
  Honestly, we used it because of your example in Final Instructions. And it resulted same as Linear.

  Random Forest:
  We used this model because it is robust to outliers, works well with non-linear data. It has lower risk of overfitting and better accuracy than other algorithms.

  If there was more time to do this project, we can improve our model using different optimizers (scaling, statemodel and etc.)

- **Discuss why and how this model should "solve" the business problem (i.e., improve along some dimension of interest to the firm).**

  As we said, our data has no so good linear relationship between them it is convenient to use Random Forest. It predicts GDP per capita that is close to a true the GDP of country.

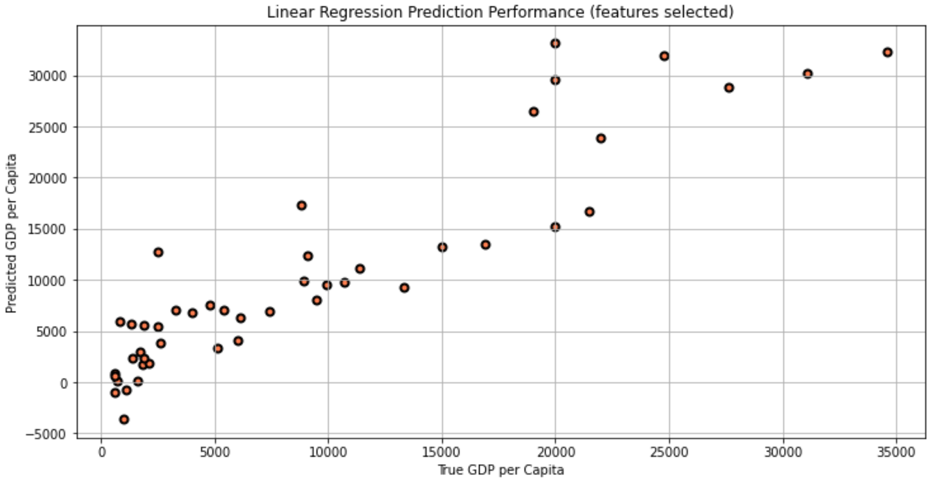
EVALUATION

- **Discuss how the result of the data mining is/should be evaluated. Provide good measures of the performance of predictive models. How should a business case be developed to project expected improvement? ROI? If this is impossible/very difficult, explain why and identify any viable alternatives.**

  As previous tasks, we used Mean Absolute Error (MAE), Root mean squared error (RMSE) and R-squared Score (R2_Score) because they are the best evaluators for regressions.

  Linear Regression:
  all features:
  MAE: 373268.60793178563
  RMSE: 1776237.7539696244
  R2_Score: -38182.43082897876

  selected features:
  MAE: 2921.00540806377
  RMSE: 4143.149069688224
  R2_Score: 0.7922534247797287



Lasso Regression:

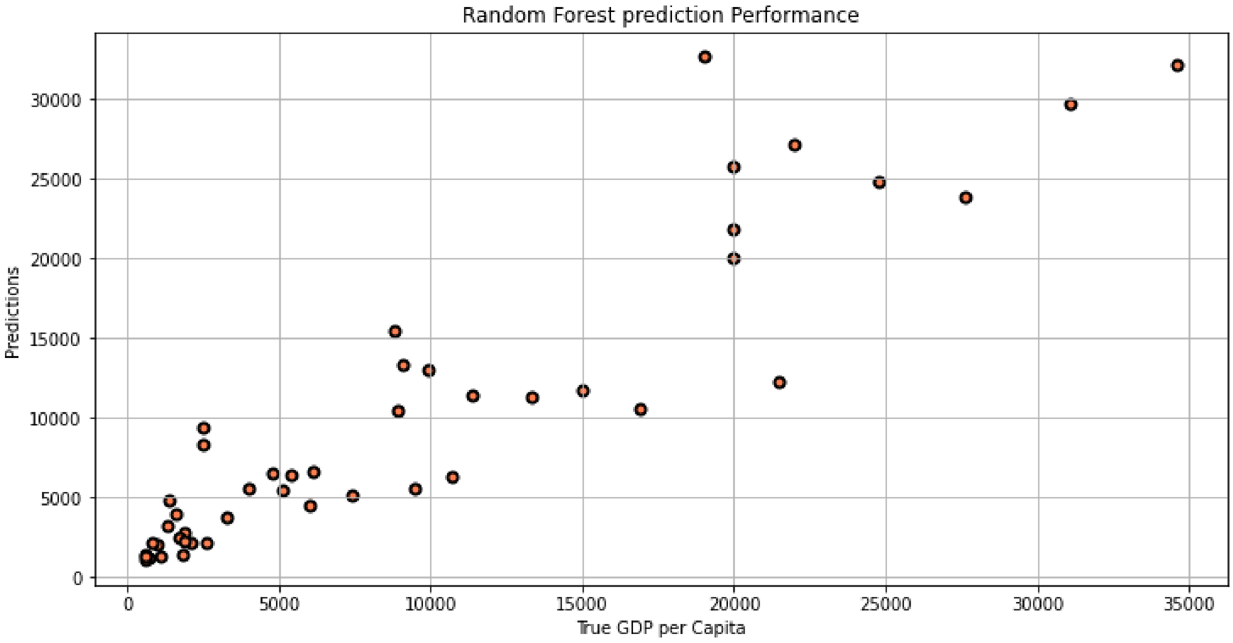
with selected features
MAE: 2921.0276000001822
RMSE: 4143.17544919213
R2_Score: 0.7922507793190192

Random Forest:

with all features
MAE: 2529.141304347826
RMSE: 3732.061477923428
R2_Score: 0.8314338636336989

with selected features
MAE: 2763.391304347826
RMSE: 4026.1710718746167
R2_Score: 0.8038188832529698

Random Forest performed better than linear regression. Random forest using all features performed significantly better than using selected features.



DEPLOYMENT

- **Discuss how the result of the data mining will be deployed.**

Firstly, we show predicted GDP per capita of Kazakhstan. (True GDP per capita: 6300$)
Random Forest with all features: 6466
Random Forest with selected features: 6185
Linear with selected features: 4491.527
Lasso with selected features: 4491.580

Random Forest with all features was closer to the real value, so this result can be used to understand how the economy is growing with population. GDP per capita analysis on a national level can provide insights on a country's domestic population influence. By this model we can predict our GDP for the next year inputting new data and create new useful reforms.

- **Discuss any issues the firm should be aware of regarding deployment.**

  We think the Agency for Economic Development of Kazakhstan should know what features they have to enter when they are using our best model (Random Forest) to predict, also should be aware of that our learning is supervised and for regression. By knowing the result, they must be aware of that it is not 100% true.

- **Are there important ethical considerations?**

  We think we have revealed everything. Still, there is something to be said. Since GDP directly assesses the daily situation of the population, we believe that it is very important to predict it and strive for progress.

- **Identify the risks associated with your proposed plan and how you would mitigate them.**

  If the prediction of our model is wrong, it will be a big risk. Therefore, to reduce this risk, we need to collect more and real data without nulls, improve their quality using statemodel and further increase the level of forecasting through various algorithms.