CAPSTONE FINAL

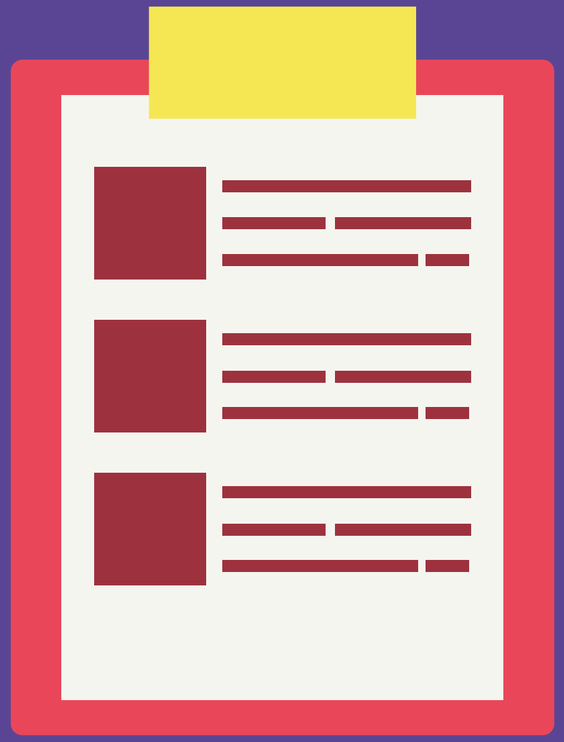
DONE BY BDA-03:

DAYANA KASSENOVA
DARIYA MAMAYEVA
RENAT ABDRAKHMANOV

OLDMEAD SECONDARY SCHOOL

SCIENCE CLASS

INTRODUCTION



Within the final exam there was a competition in Kaggle. We were given a dataset, containing 6 numeric features and 9 categorical features.

OLDMEAD SECONDARY SCHOOL

SCIENCE CLASS

OUTLINE OF THE PROJECT

- 1.TO CLARIFY REGRESSION TASK OF THE PROJECT
- 2.TO EXPLORE THE DATASETS
- 3. TO PREPARE DATA
- 4.TO CREATE DIFFERENT MODELS
- 5. TO COMPARE MODELS
- 6. TO CHOOSE BEST PREDICTION OF MSE

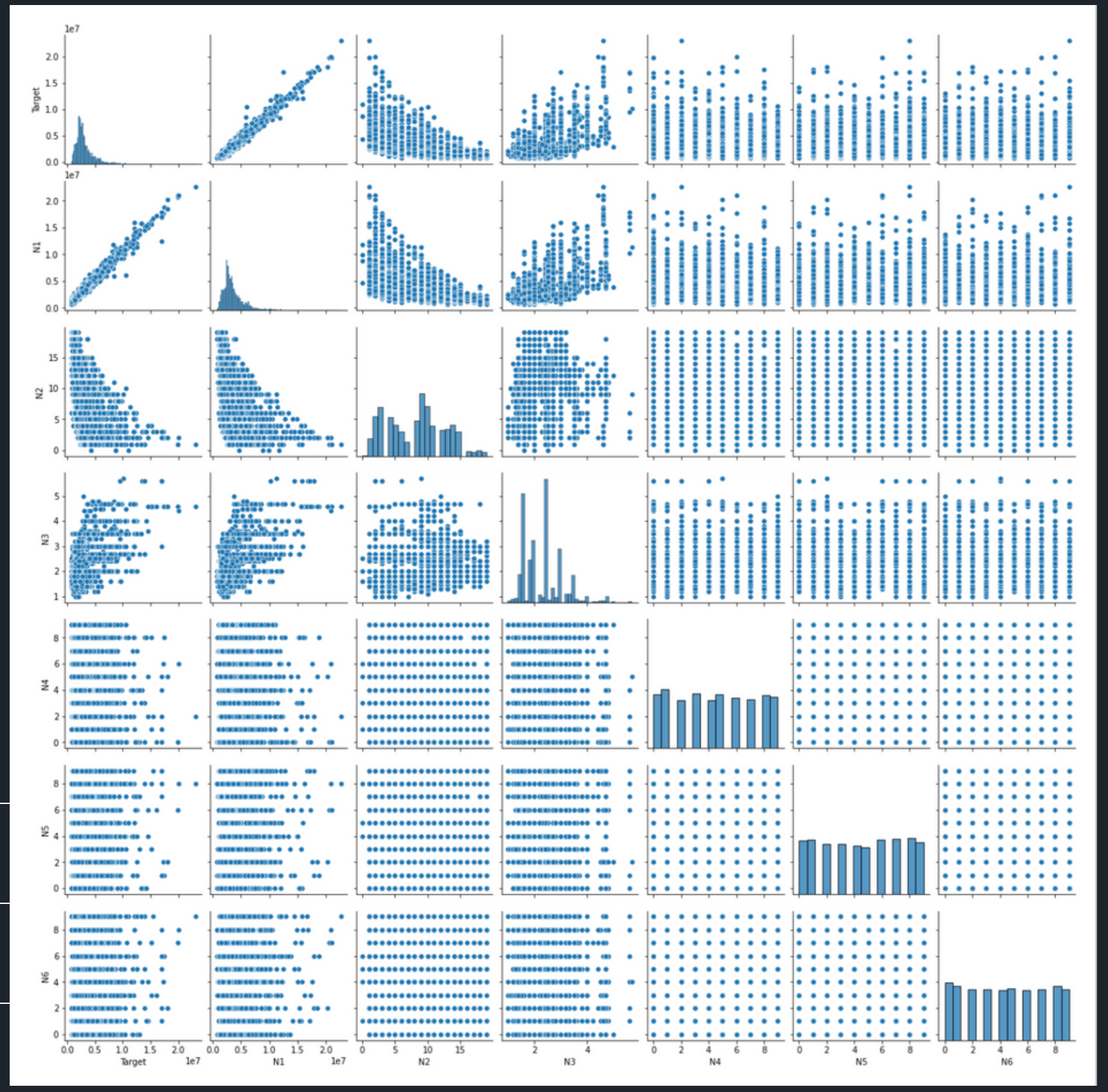
OLDMEAD SECONDARY SCHOOL

SCIENCE CLASS

DATASET EXPLORE

Dataset contains 16 columns, from which 6 columns represents numerical features and 9 columns with categorical features.

The image shows the correlation matrix between the variables.



OLDMEAD SECONDARY SCHOOL

SCIENCE CLASS

INFORMATION ABOUT COLUMNS

CATEGORICAL columns HAS UNIQUE VALUES :

C1 - has 3 nominal values: [1, 2, 3]

C2 -has 2 categorical values ['M', 'A']

C3 - has 3 categorical values['F', 'B', 'A']

C4 -has 3 categorical values ['V', 'K', 'B']

C5 - has 5 cat. values['B', 'BG', 'D', 'G', 'H']

C6 -has 7 cat. values ['S', 'C', 'V', 'L', 'M', 'U', 'P']

C7 - -has 21 categorical values

C8 - -has 28 categorical values

C9 - -has 32 categorical values

OLDMEAD SECONDARY SCHOOL

SCIENCE CLASS

DATA PREPARATION

1. ***CHECKING MISSING VALUES:***
2. **IN TRAIN DATASET, WE KNOW THAT THERE MISSING VALUES IN *N4,N5,N6 COLUMNS*. THEN WE DROPPED THESE CONTAINING ROWS.**
3. **IN TEST DATASET, NO NULL VALUES.**
4. **INCLUDE DUMMY VARIABLES**
5. **APPLYING LABEL ENCODING FOR VARIABLES WHERE NUMBER OF UNIQUE ELEMENTS IS MORE THAN 3**

MODEL CONSTRUCTION

Fitting train set to the RandomForest

```
randFor = RandomForestRegressor().fit(X_train, y_train)
```

Tuning Parameters with RandomizedSearchCV

```
randSearch = RandomizedSearchCV(reg, param, n_iter = 100, cv = 5, verbose=2, random_state=0,
                                scoring='neg_mean_squared_error', n_jobs = -1)
search = randSearch.fit(X_train, y_train)
print('Best features for Random Forest:', search.best_params_)
```

```
Fitting 5 folds for each of 100 candidates, totalling 500 fits
Best features for Random Forest: {'n_estimators': 1200, 'min_samples_split': 2, 'min_sampl
es_leaf': 2, 'max_features': 'auto', 'max_depth': 20, 'bootstrap': True}
```

CROSS-VALIDATION

SCORE 91295085639.41116

```
# Applying cross validation to evaluate tuned model's performance

randFor = RandomForestRegressor(n_estimators=1200, min_samples_split=2,
                                min_samples_leaf=2, max_features='auto',
                                max_depth=20, bootstrap=True).fit(X_train, y_train)

randForMSE = abs(cross_val_score(randFor, X_train, y_train, cv=5,
                                 scoring='neg_mean_squared_error')).mean()
```

```
randForMSE
```

```
91295085639.41116
```

Ridge Regression

1. Fitting Ridge Regression and obtaining cross-val score
2. Finding optimal alpha by using RidgeCV (alpha=1)
3. Set aplha to Ridge Regression and obtaining cross-val score

```
rid = Ridge().fit(X_train, y_train)
abs(cross_val_score(rid, X_train, y_train, cv=5, scoring='neg_mean_squared_error')).mean()
```

```
92482420801.94667
```

```
rid = RidgeCV(alphas=np.arange(1, 100001)).fit(X_train, y_train)
rid.alpha_
```

```
1
```

```
rid = Ridge(alpha=1).fit(X_train, y_train)

ridMSE = abs(cross_val_score(rid, X_train, y_train, cv=5, scoring='neg_mean_squared_error')

ridMSE
```

```
92482420801.94667
```

Lasso Regression

1. Fitting Lasso Regression
2. Finding optimal alpha using LassoCV
3. Fitting Lasso Regression with optimal alpha

```
las = Lasso().fit(X_train, y_train)
abs(cross_val_score(las, X_train, y_train, cv=5, scoring='neg_mean_squared_error')).mean()
```

```
92483112243.668
```

```
las = LassoCV(alphas=np.arange(1, 1001)).fit(X_train, y_train)
las.alpha_
```

```
987
```

```
las = Lasso(alpha=las.alpha_).fit(X_train, y_train)

lasMSE = abs(cross_val_score(las, X_train, y_train, cv=5, scoring='neg_mean_squared_error')

lasMSE
```

```
92404272948.96236
```

Model Comparison

Model	MSE
Random Forest	91295085639.41116
Ridge	92482420801.94667
Lasso	92404272948.96236

CONCLUSION

- To summarise the work done, we want to say that it was we want to say that it was an invaluable experience in practice of creating models by using different techniques.
- Furthermore, we find out the smallest MSE by RandomForest model with tuning hyperparameters.
- Finally, working in team lead us to achieve a shared goals in an effective way :)

OLDMEAD SECONDARY SCHOOL

SCIENCE CLASS

Thank you!

OLDMEAD SECONDARY SCHOOL

SCIENCE CLASS