APPLICATION SCORING MODEL

'Data analysis in business'

PRESENTATION

HSE UNIVERSITY

OUR TEAM

Alena Iakovleva

Anastasiia Prokhorova

Vadim Khanin

HSE UNIVERSITY

DATA PREPROCESSING

main steps

- data understanding
- data cleaning and manipulation
- portraits

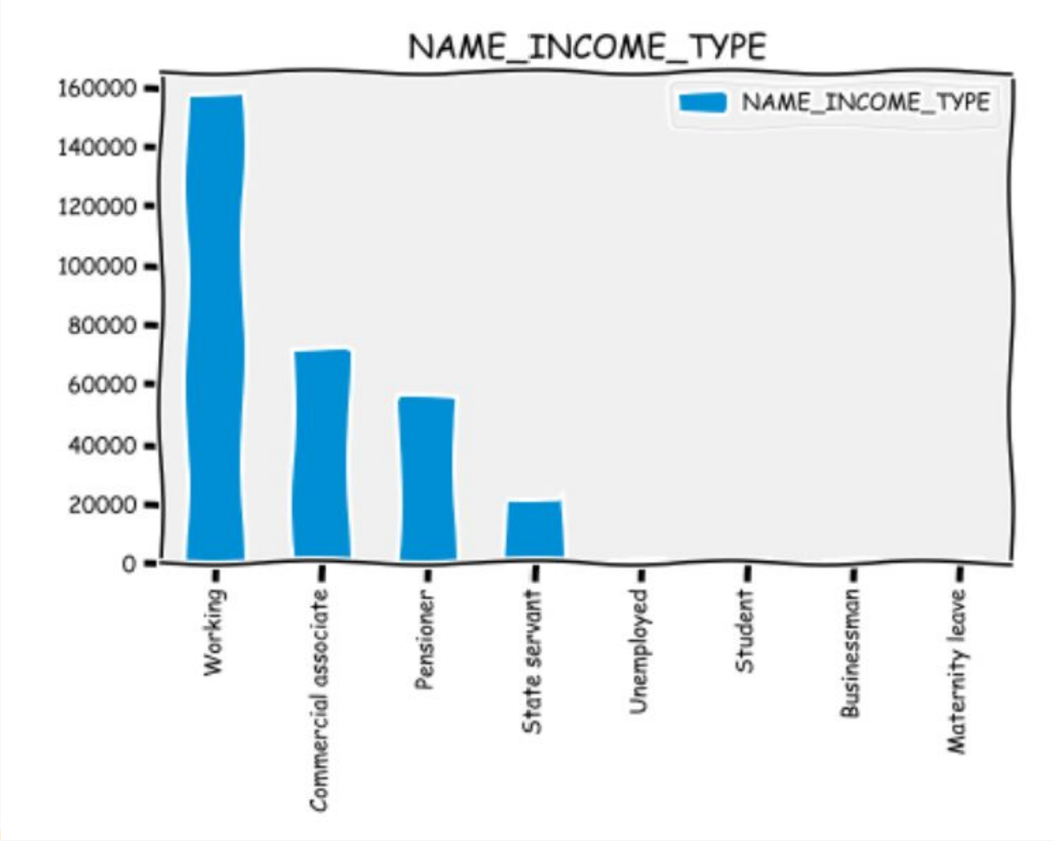
DATA PREPROCESSING

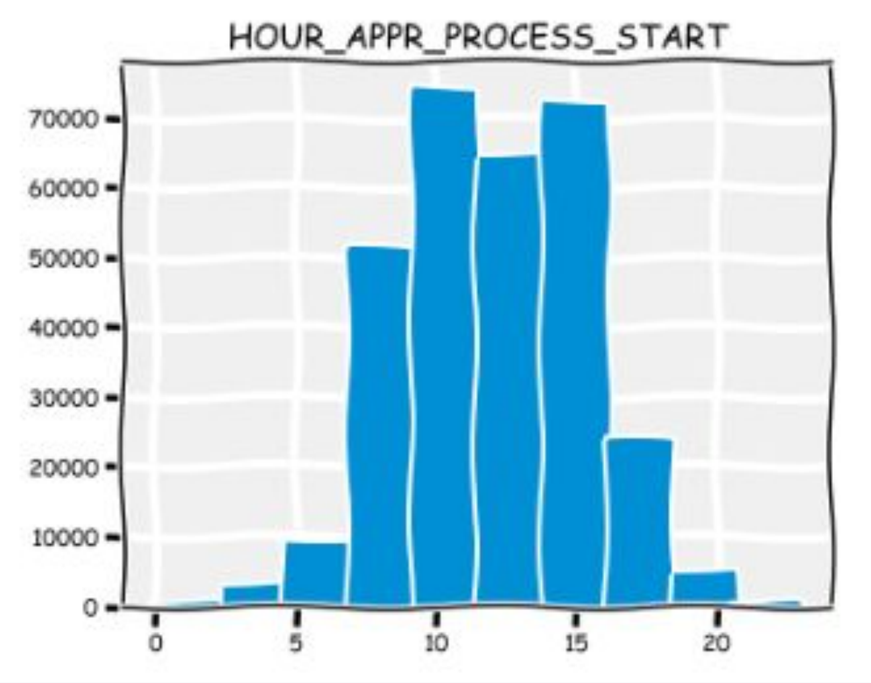
cleaning

- replaced values "XNA" and "XAP" with NaN
- replaced empty values in OCCUPATION_TYPE by "NaN" value
- created new variables:
  - **CREDIT_ACTIVE_CLOSED**
  - **CREDIT_ACTIVE_ACTIVE**
  - **CREDIT_ACTIVE_SOLD**
  - **CREDIT_ACTIVE_BAD_DEBT**
  - **CREDIT_DAY_OVERDUE**
  - **AVG_AMT_CREDIT_SUM**
  - **AMT_APPLICATION_APPROVED**
  - **AMT_APPLICATION_REFUSED**
  - **AMT_APPLICATION_CANCELED**
  - **AMT_APPLICATION_UNUSED**

DATA PREPROCESSING

distribution of data

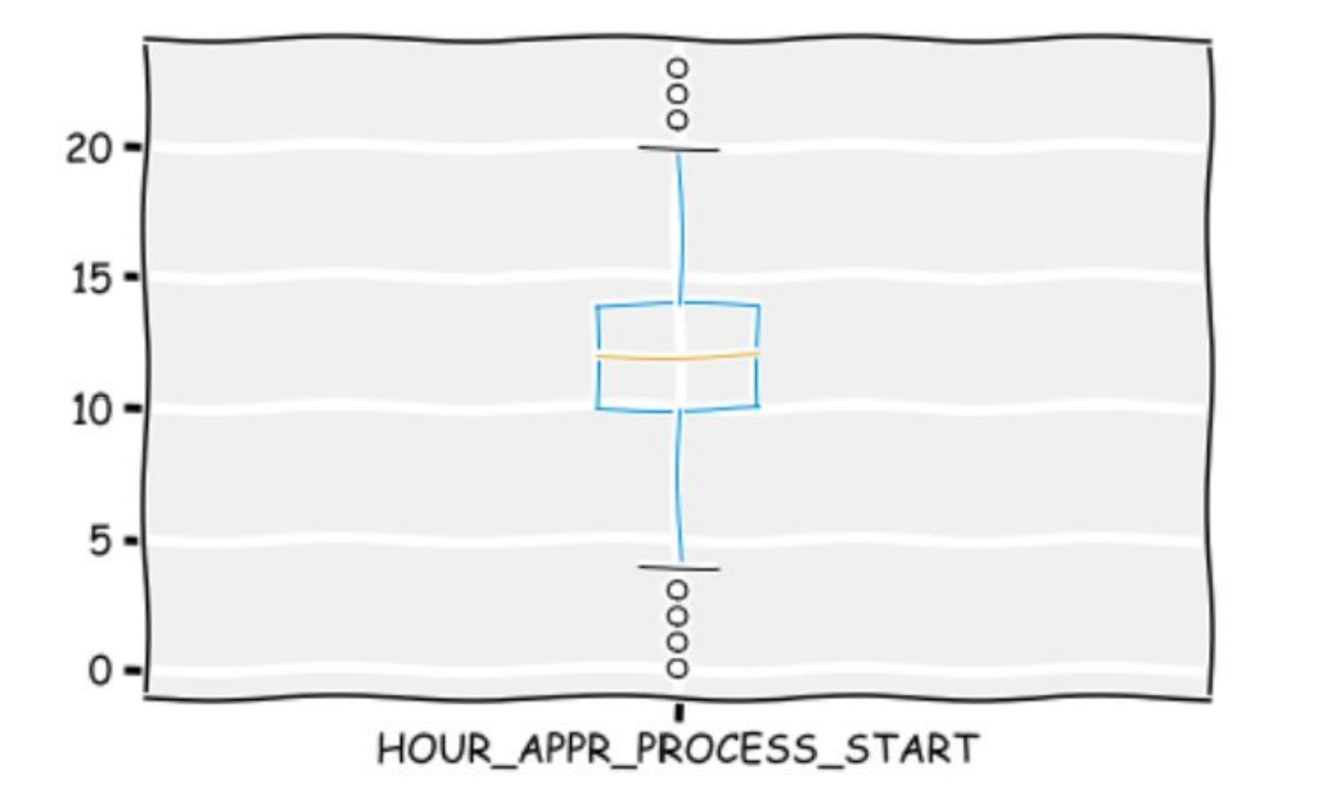




DATA PREPROCESSING

outliers

percentile[5,95]



DATA PREPROCESSING

outliers

- days columns -> to positive
- Age , Working_Exp , Id Doc Age Registration_Age column -> positive Years
- Income, Loan_Credit , Loan_Annuity, Total_Goods_Price into more understandable 'datasize' (/100000, round(2))
- created a age bucket column with values "20-30","30-40","40-50","50-60","60-70"

DATA PREPROCESSING

dataset info

12 object variables

12 binary varibles

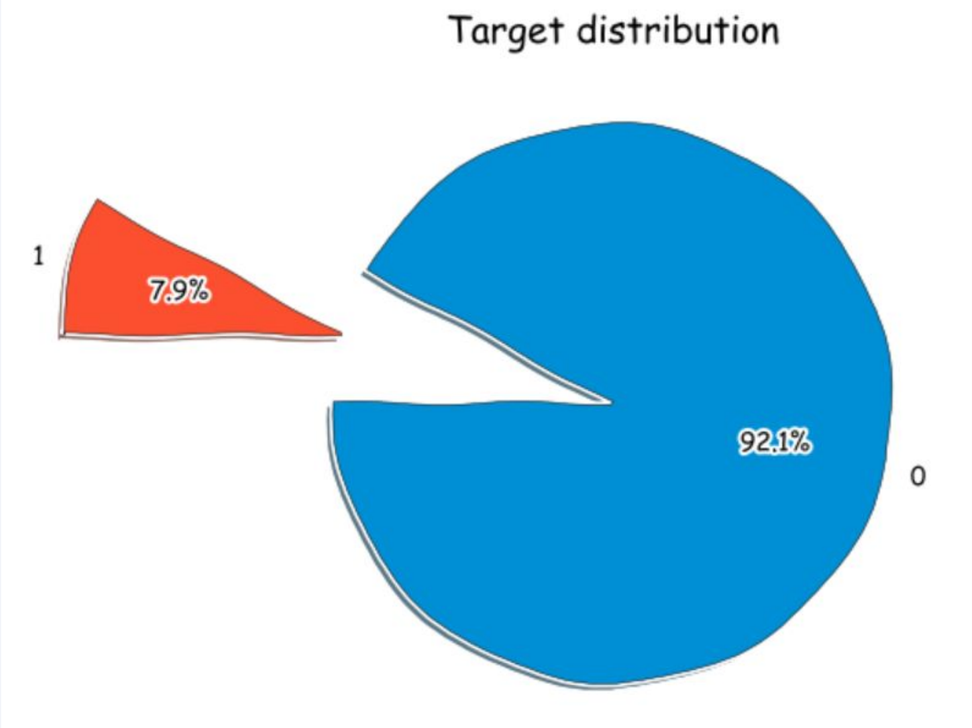
58 numeric variable

initially → **307511** rows & 132 columns

cleaning → **304531** rows & 81 columns

outliers → **258061** rows & 81 columns

final → **210843** rows & 81 columns



DATA PREPROCESSING

worth focusing on

- amount of active credits
- credit day overdue
- average amount of credit
- amount of applications refused
- age
- registration

HSE UNIVERSITY

LOGISTIC REGRESSION

	modification	ROC-AUC_liblinear	ROC-AUC_newton	Score_liblinear	Score_newton
0	base	0.605348	0.516300	0.920718	0.920723
1	PCA	0.711417	0.711417	0.920500	0.920500
2	balanced	0.597816	0.721580	0.572091	0.663087
3	balanced + PCA	0.708307	0.708307	0.653964	0.653964

LDA & QDA

ACCURACY SCORE

LDA

```
#evaluate model
scores = cross_val_score(lda, X, y, scoring='accuracy', cv=cv, n_jobs=-1)
print(np.mean(scores))
```

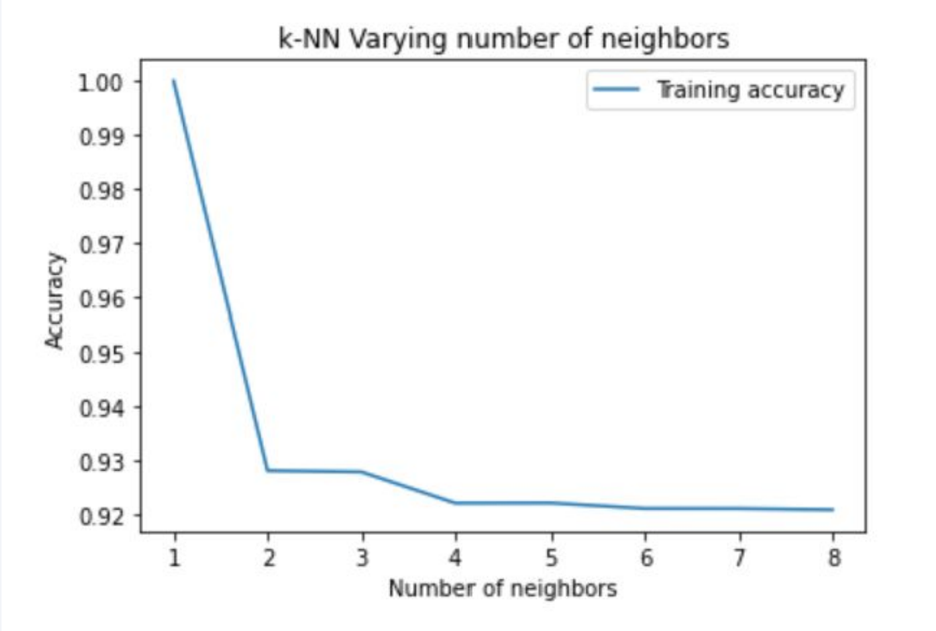
```
0.9918280042201194
```

QDA

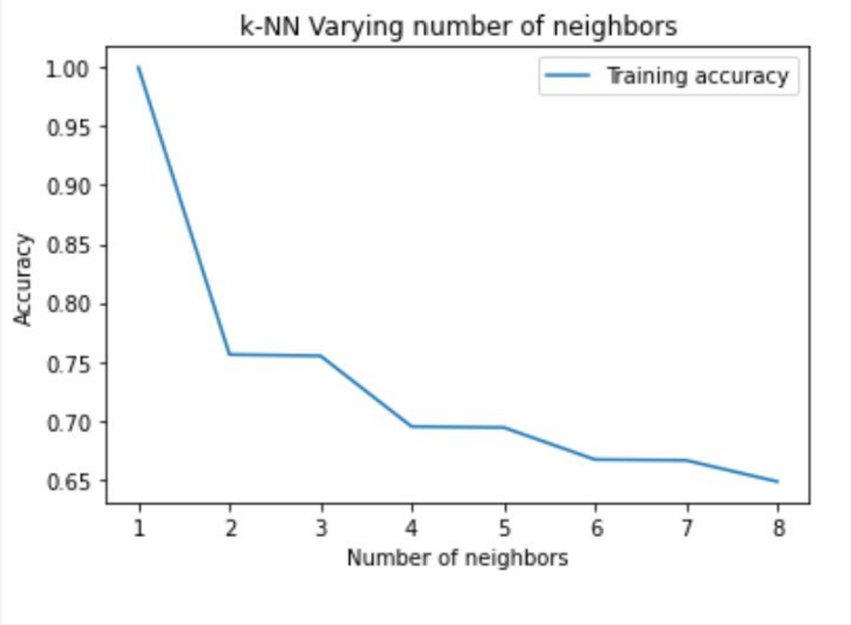
```
#evaluate model
scores = cross_val_score(qda, X, y, scoring='accuracy', cv=cv, n_jobs=-1)
print(np.mean(scores))
```

```
1.0
```

KNN



raw data with and without PCA



balanced data no PCA

GRADIENT BOOSTING

```
Result for gradboost(sklearn) for raw data
0.920296900566767
0.5049846734656602
0.7301664450540971
```

```
Result for gradboost(sklearn) for balanced data
0.6639246186060425
0.6640655986757982
0.7276554376323593
```

GRADIENT BOOSTING

```
Results of catboost on raw data
Accuracy -  0.9202731864639901
Balanced Accuracy -  0.5085577137348056
ROC-AUC -  0.7373347323666608
```

```
Results of catboost on balanced data
Accuracy -  0.947711582715418
Balanced Accuracy -  0.9394890788281144
ROC-AUC -  0.9621702040929335
```

INTEGRATING INTO THE COMPANY

1. RETRAIN THE MODEL ON COMPANY DATA.

2. Determine how the interaction between the model and existing systems of the company will be carried out.

3. Automation of the data update process.

4. Performance monitoring

5. staff training

HSE UNIVERSITY

THANK YOU!
WE ARE READY TO ANSWER YOUR QUESTIONS!

HSE UNIVERSITY