Loan Classifier

ML course
Final Project

Bakhtovar Rahmatov

TABLE OF CONTENTS

About the Project

Background information, data sources, challenges and learning outcomes

01

Exploratory Data Analyses

Data content, analytics and main findings

02

Preprocessing

Data distribution, Correlation, Target Mean encoding

03

04

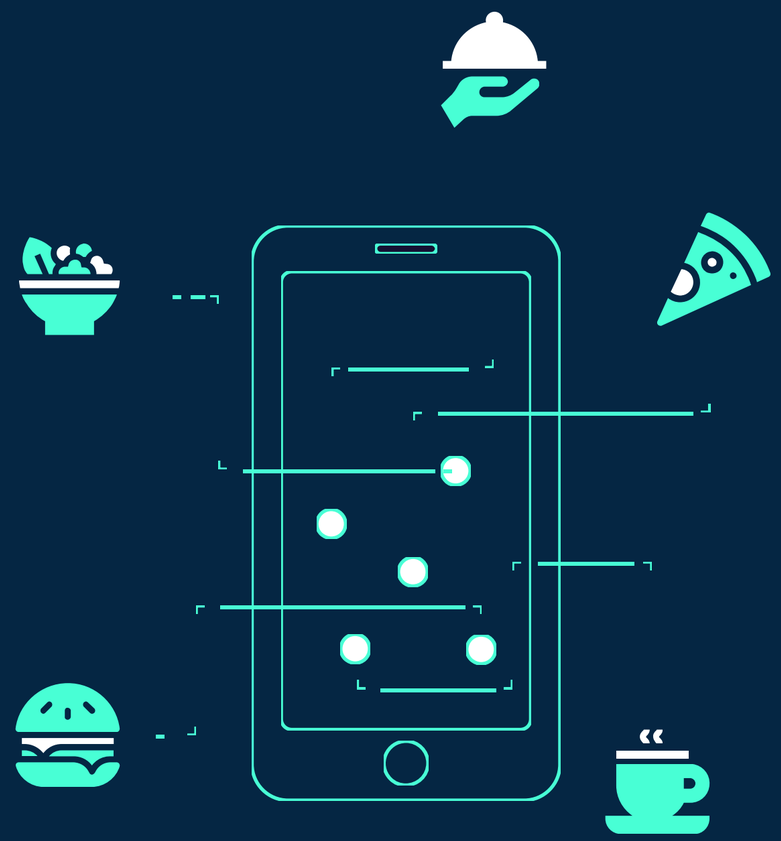
Models

Logistics Regression, Random Forest, Catboost, KNN

05

Conclutoion

models comparison and conclusion



ABOUT THE PROJECT

For the final project, I chose to build a classifier model for new loan requests. I chose this topic because back in 2015, one of my co-workers built a "lead-scoring" model, which gives "grades" to potential new customers, and first phone calls would be made based on the calculated "grade." This model showed significant improvement of lead conversion and sales. I was always curious how he used historical data to create this model on a simple Excel spreadsheet. Thanks to this course, I learned about ML and the power it gives to get insights from the data

Data source and materials used:

- Kaggle was used to find data for the project
- Class work opynb and materials from classroom

EDA

Libraries used

Initial data analytics

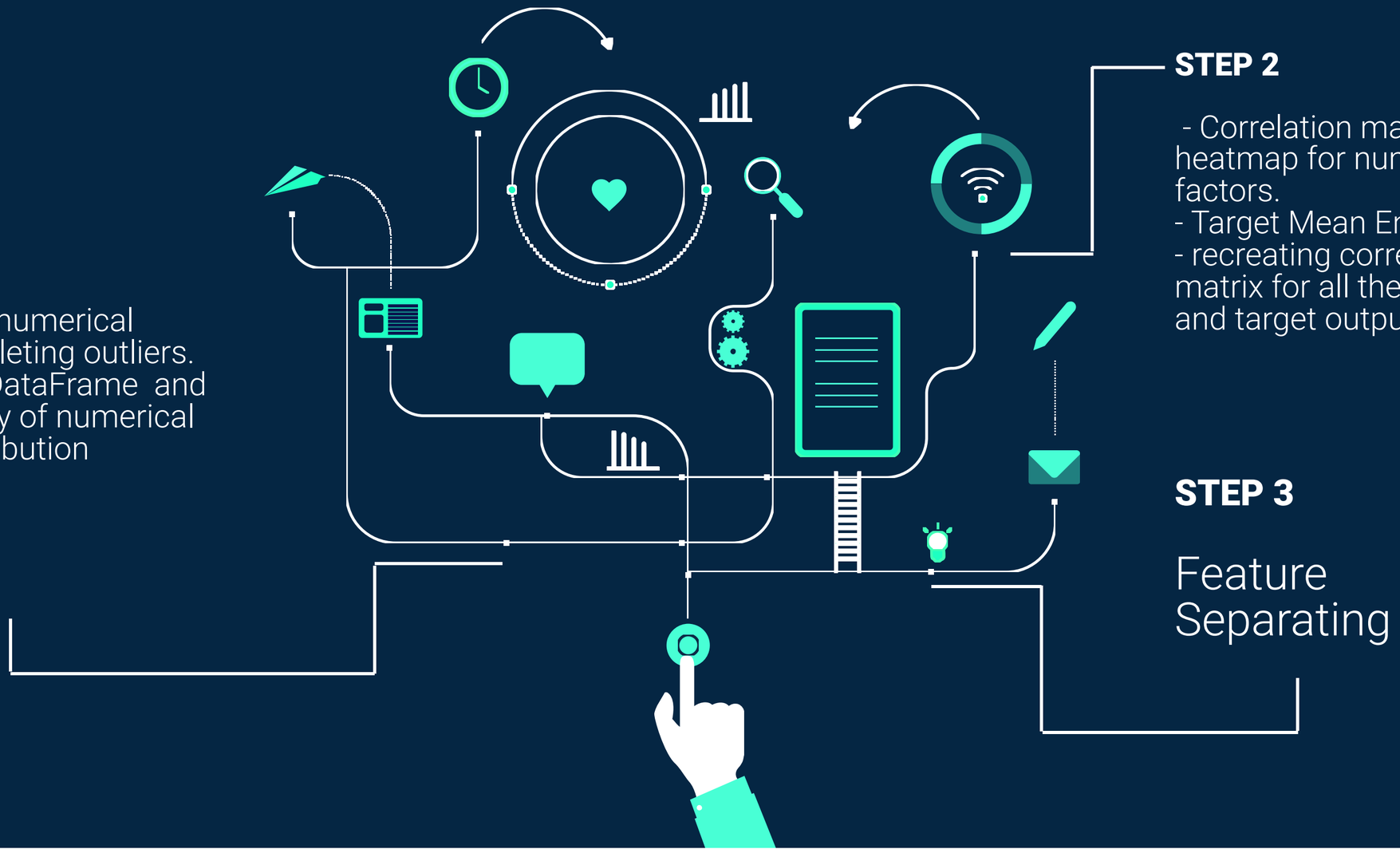
Visual analyses of some factors

NumPy
Pandas
matplotlib.pyplot
seaborn
sklearn.preprocessing
sklearn.model_selection
sklearn.linear_model
sklearn.metrics
sklearn.datasets
sklearn.ensemble
catboost
sklearn.neighbors

Preprocessing

STEP 1

Analyses of numerical columns. Deleting outliers. Recreating DataFrame and visual display of numerical factors distribution

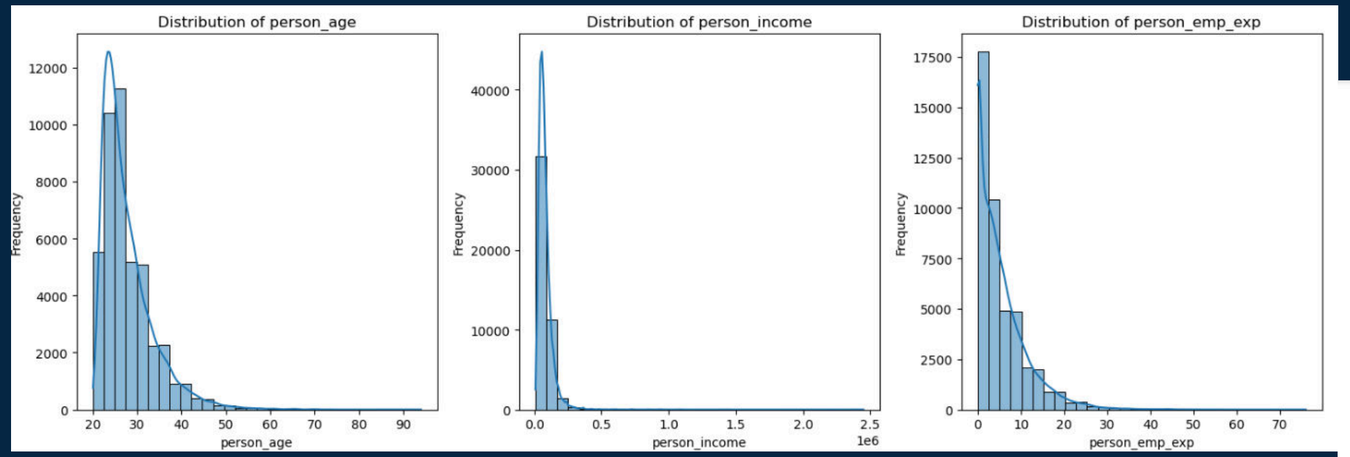


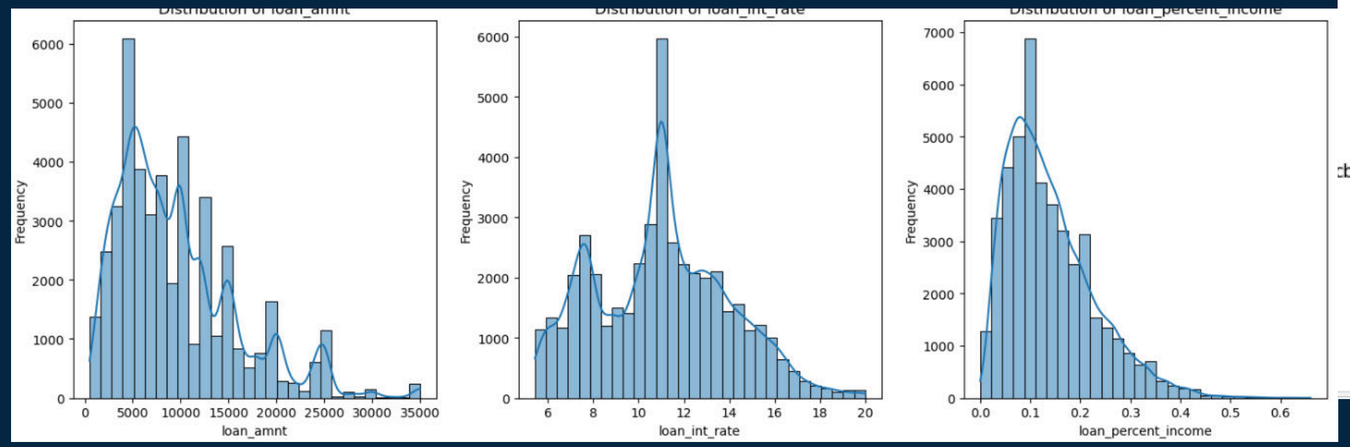
STEP 2

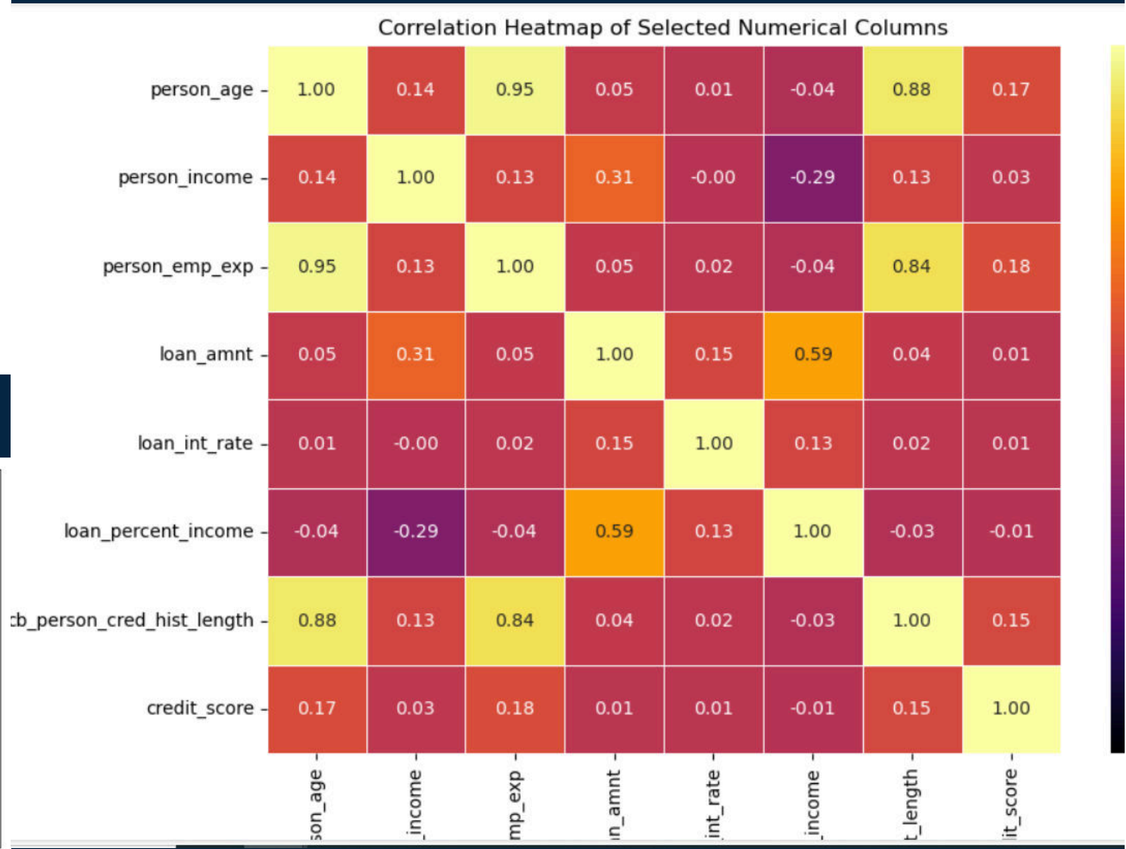
- Correlation matrix and heatmap for numerical factors.
- Target Mean Encoding
- recreating correlation matrix for all the factors and target output

STEP 3

Feature Separating


Distribution of person_age
Frequency
person_age
Distribution of person_income
Frequency
person_income
1e6
Distribution of person_emp_exp
Frequency
person_emp_exp


Frequency
loan_amnt
Frequency
loan_int_rate
Frequency
loan_percent_income


Correlation Heatmap of Selected Numerical Columns
person_age
person_income
person_emp_exp
loan_amnt
loan_int_rate
loan_percent_income
cb_person_cred_hist_length
credit_score
1.00 0.14 0.95 0.05 0.01 -0.04 0.88 0.17
0.14 1.00 0.13 0.31 -0.00 -0.29 0.13 0.03
0.95 0.13 1.00 0.05 0.02 -0.04 0.84 0.18
0.05 0.31 0.05 1.00 0.15 0.59 0.04 0.01
0.01 -0.00 0.02 0.15 1.00 0.13 0.02 0.01
-0.04 -0.29 -0.04 0.59 0.13 1.00 -0.03 -0.01
0.88 0.13 0.84 0.04 0.02 -0.03 1.00 0.15
0.17 0.03 0.18 0.01 0.01 -0.01 0.15 1.00

Models

- Logistics Regression
- Random Forest
- Catboost
- KNN

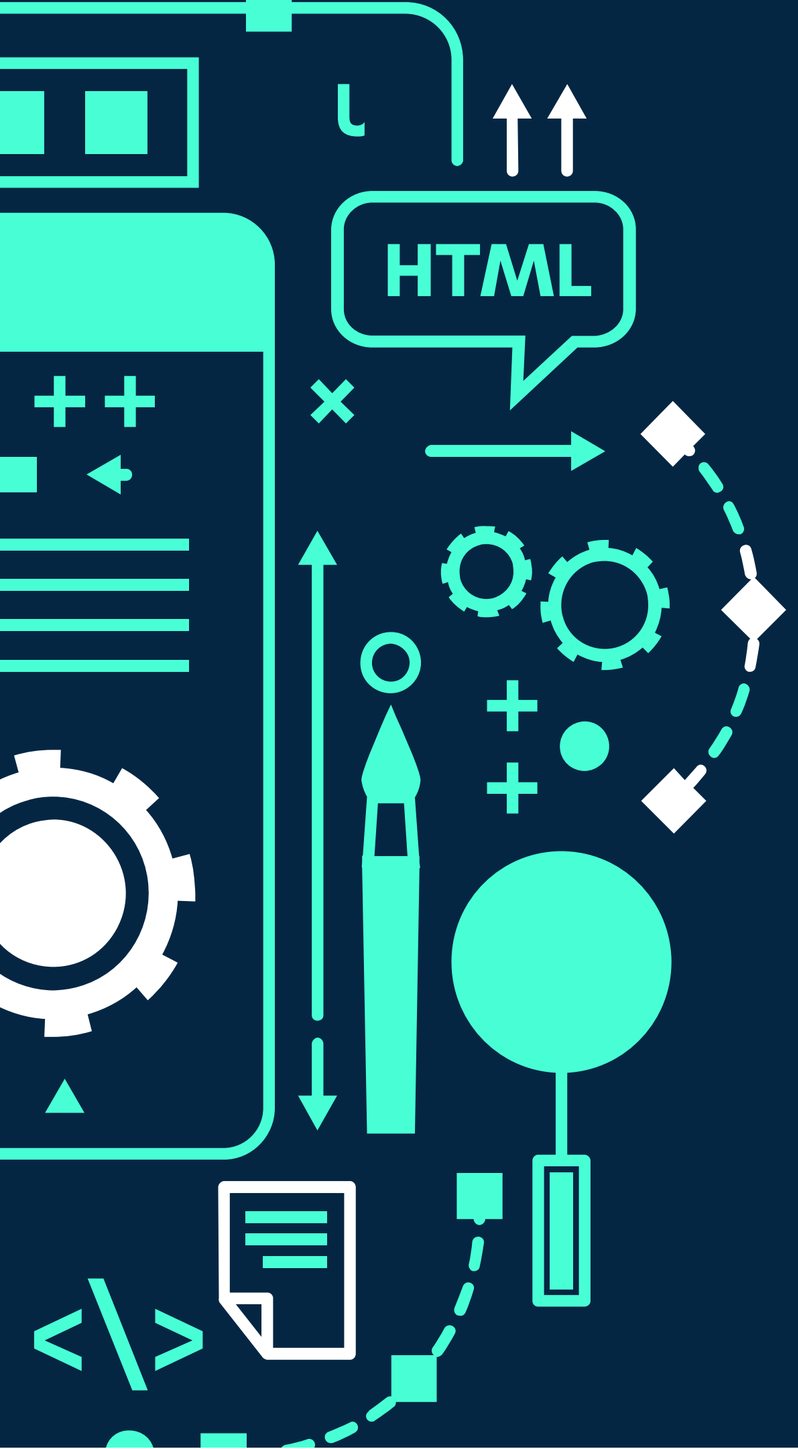
Conclution

]:

	Model	Accuracy
1	Random Forest	93.088121
2	Gradient Boost	90.187799
3	K Neighbors	90.187799
0	Logistic Regression	88.976553

All the models have an accuracy of 88% and higher

Random forest is the best model for out project with an accuracy of 93%



THANKS!