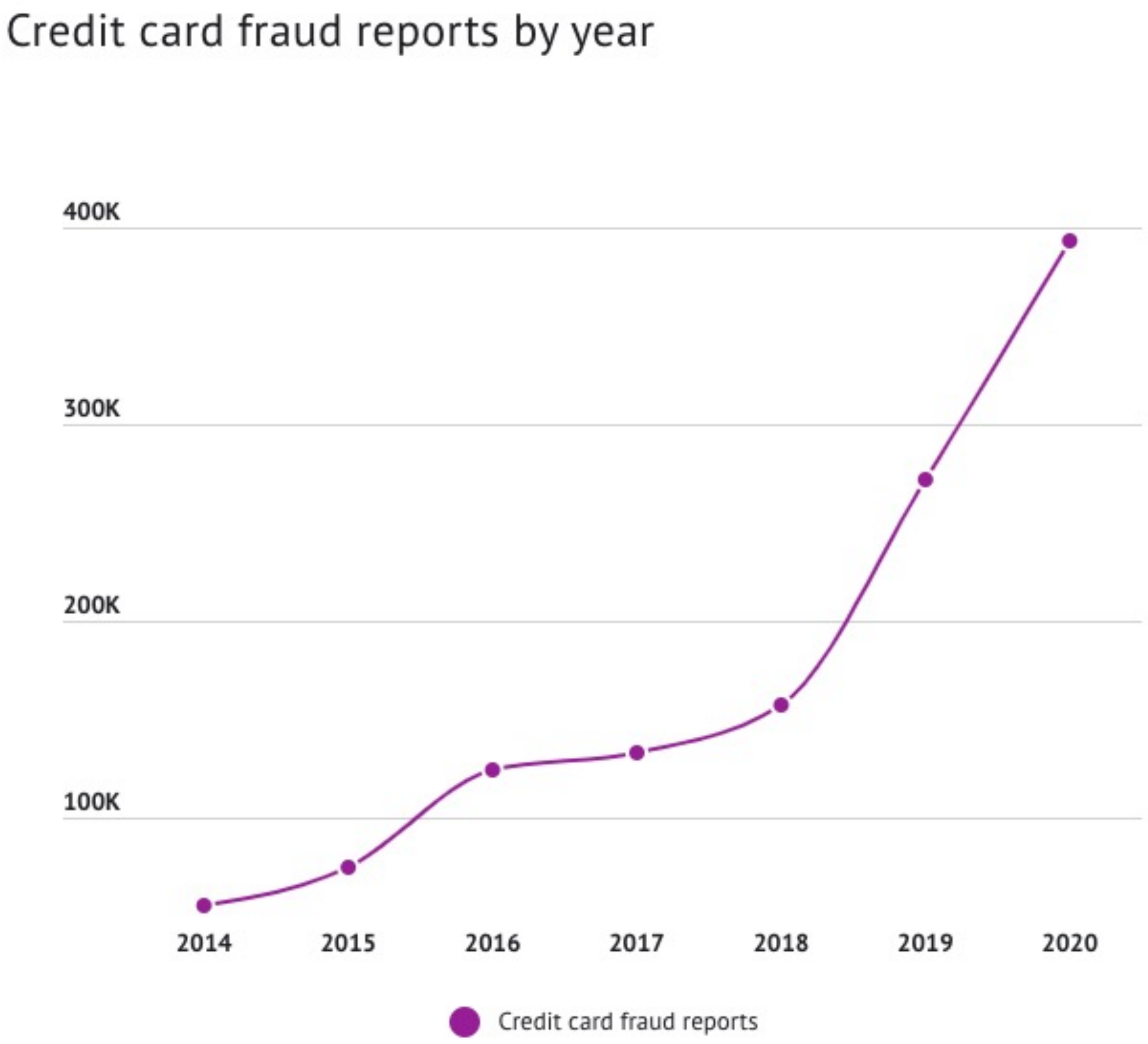
CREDIT CARD FRAUD DETECTION

PROBLEM STATEMENT

There were over 390 000 reports of credit card fraud in 2020

It is a 44,7 percent increase from the 2019



DATA

1.85 million of transactions

1.84 million – legit transactions

10 000 - fraudulent transactions

METHODOLOGY

- Data acquisition

METHODOLOGY

- Data acquisition

- EDA
  SQL and Pandas

METHODOLOGY

- Data acquisition

- EDA
  SQL and Pandas

- Model selection
  and training
  pyspark

METHODOLOGY

- Data acquisition

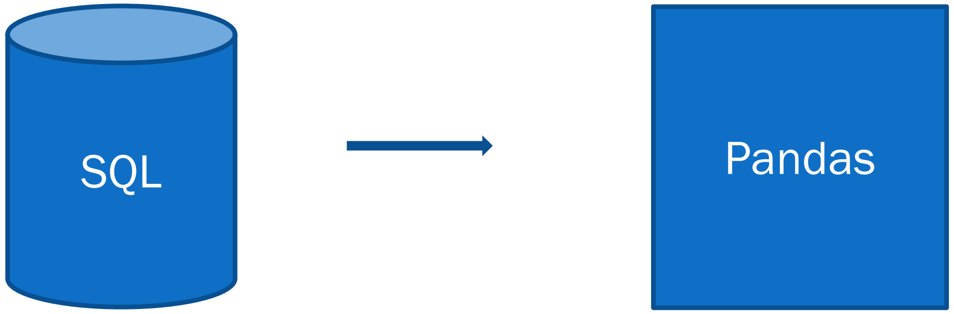
- EDA
  SQL and Pandas

- Model selection
  and training
  pyspark

- Model tuning
  pyspark

PROCESS FLOW

EDA



Dealing with missing values

Categorical features conversion

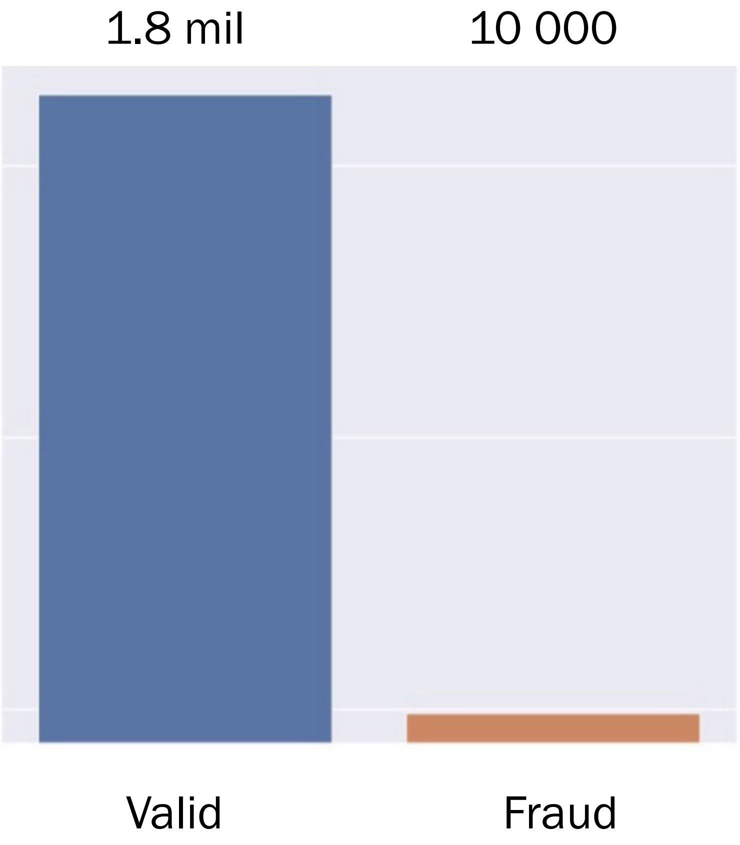
Feature selection

PROCESS FLOW

HANDLING CLASS IMBALANCE

Stratified train – test split

Oversample minority class



MODEL SELECTION AND TUNING

Logistic Regression	- ROC AUC	0.825670
Random Forest	- ROC AUC	0.857615
GBTClassifier	- ROC AUC	0.928379

MODEL SELECTION AND TUNING

Logistic Regression	- ROC AUC	0.825670
Random Forest	- ROC AUC	0.857615
GBTClassifier	- ROC AUC	0.928379

Grid Search with Cross Validation for tuning the model

FINAL MODEL

GBTClassifier

Accurracy - 0.974

Precision – 0.983

Recall – 0.926

F1 score – 0.953

ROC AUC – 0.923

FUTURE WORK

Train model on more data

Spend more time on model tuning

Deploy the model

Thank you