Entity Recognition and Accounting Type Prediction: Simplifying Accounting on the Basis of Machine Learning

Hasham Munir and David Bläsi

Sevenit GmbH

sevDesk

About us

- David Bläsi
  - M.Sc. in Mathematics (2017) with focus on probability theory and medical statistics
  - Data Scientist / Machine Learning Engineer at Sevenit GmbH since September 2017: focus on automated invoice processing
- Hasham Munir
  - M.Sc. Communication & Media Engineering 2015 with focus on Database systems & Technologies
  - Data Scientist / Machine Learning Engineer at Sevenit GmbH since Oct.2016

Agenda

1. Introduction
2. Entity Recognition for invoice data
   - Task description
   - Workflow
   - Model comparison
3. Accounting type prediction
   - Idea of using accounting type
   - sevCleaner
   - sevProto
   - sevAuto
4. Live demo

Introduction: sevDesk and its functionalities

Digitize receipts

Cancellation bills / credits

Write bills

Write reminders

Online banking

VAT return

order processing

Net income method

and many more...

Entity Recognition for invoice data

David Bläsi – Data Scientist / Machine Learning Engineer

sevDesk

Task description - motivation

- Idea: facilitate process of digitalizing relevant information from invoices and receipts

- Our users:
  - have between 10 and 150 incoming invoices each month
  - come from a wide variety of domains

→ Goal: build universal invoice recognition system

Task description - entity recognition

- Goal: recognize and classify parts of natural language texts
- Named entity: e.g. persons, organizations, dates, etc.

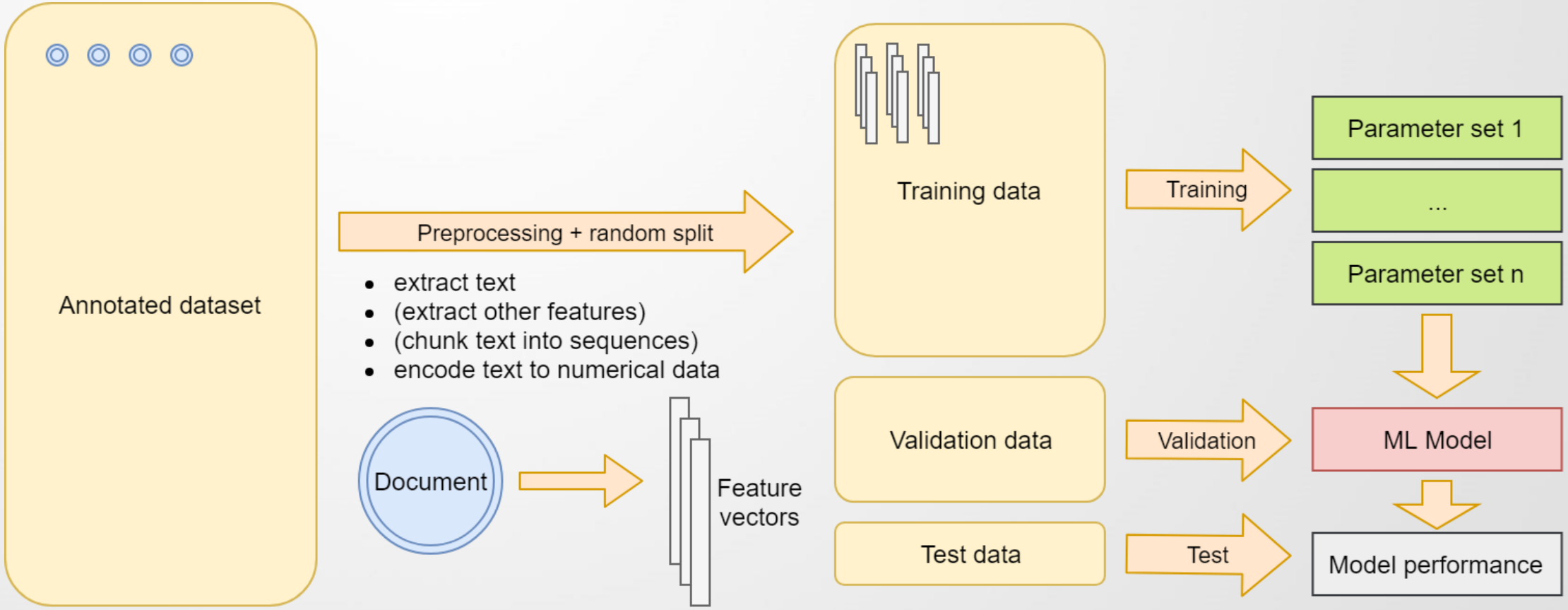
Organisation: Sevenit GmbH was founded in Time: October 2013 and is based in Location: Offenburg.

- Supervised learning algorithms have proven suitable for this task

Workflow - data preparation

- We use data provided by our users to train our models
- Raw data has to be annotated to create a dataset for supervised learning

Workflow - training ML models



Model comparison

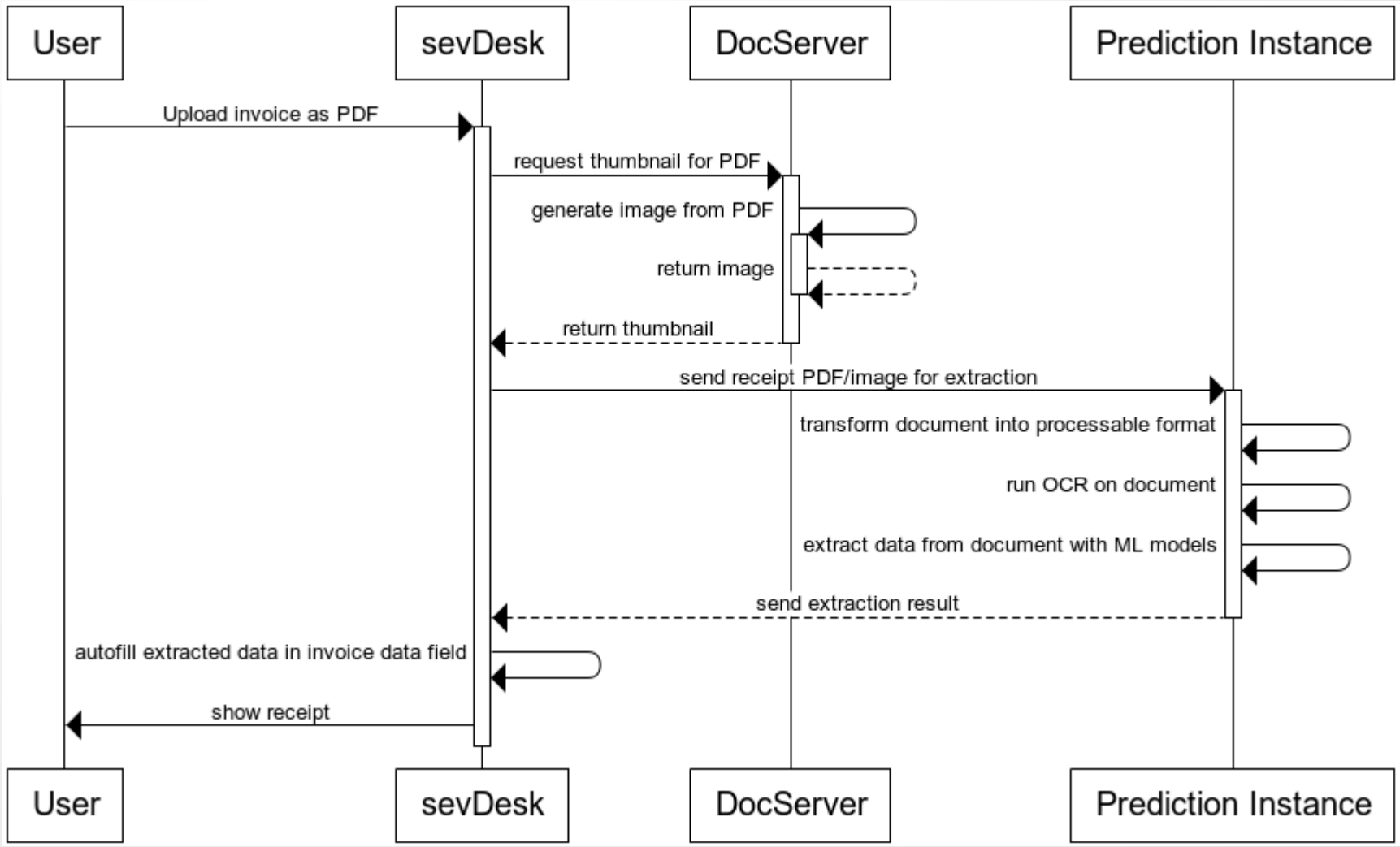
Conditional random fields (CRF)

- relies on hand-crafted features
- works well with small amount of data
- flexible document length
- comparatively few open source frameworks / libraries

Recurrent neural networks (RNN)

- no hand-crafted features
- scales well to large amounts of data
- expects rigid data format (document length)
- large choice of frameworks and active community

The model in production



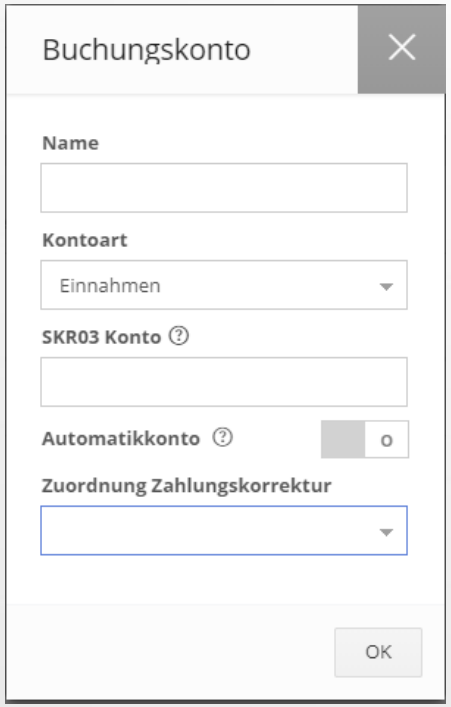
Accounting Type Prediction

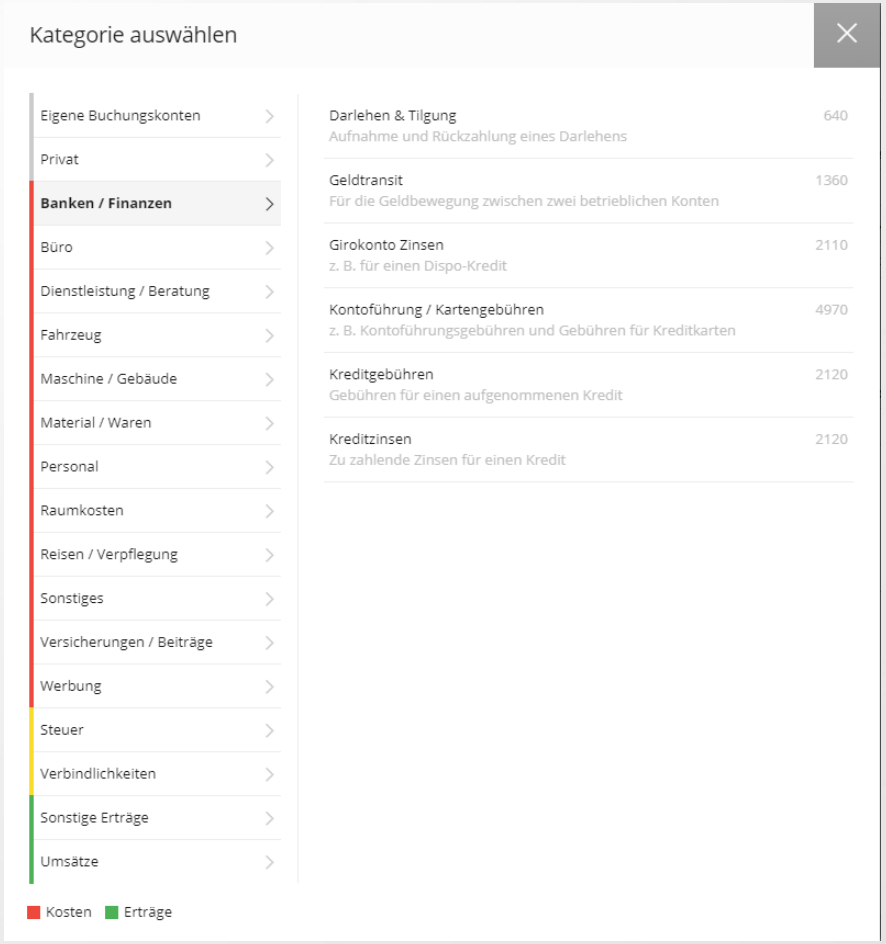
Hasham Munir – Data Scientist / Machine Learning Engineer

sevDesk

Accounting Type

- Accounting type helps to store the VAT
- Public accounting types
- Private accounting types



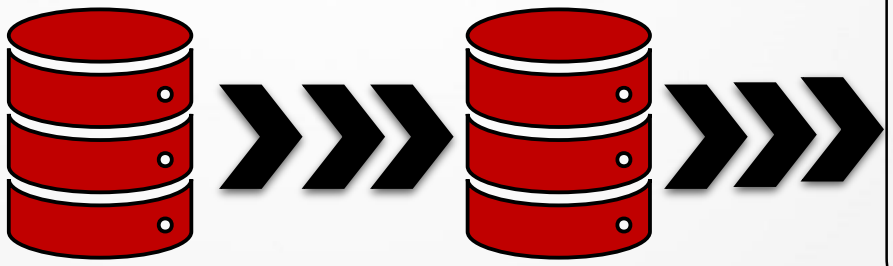


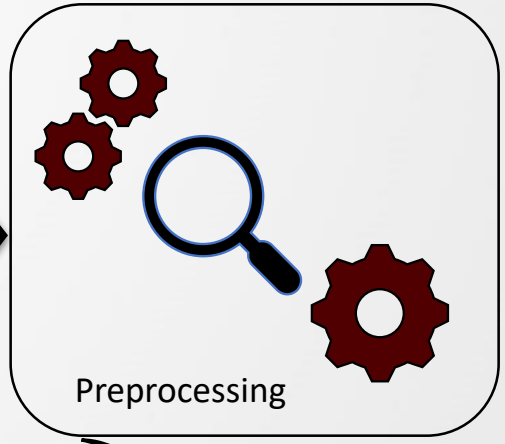
sevCleaner

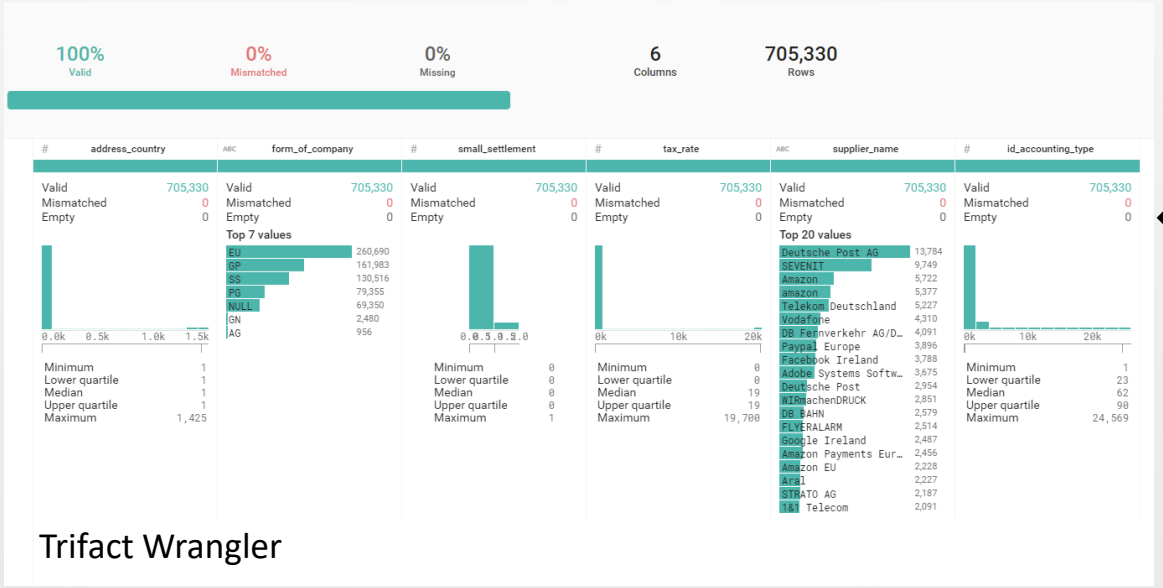
W.W Customers

Live Servers

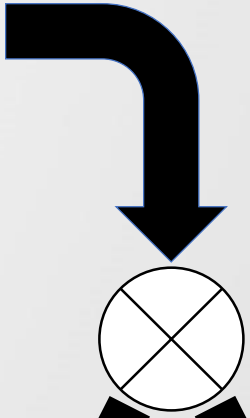
Test Server







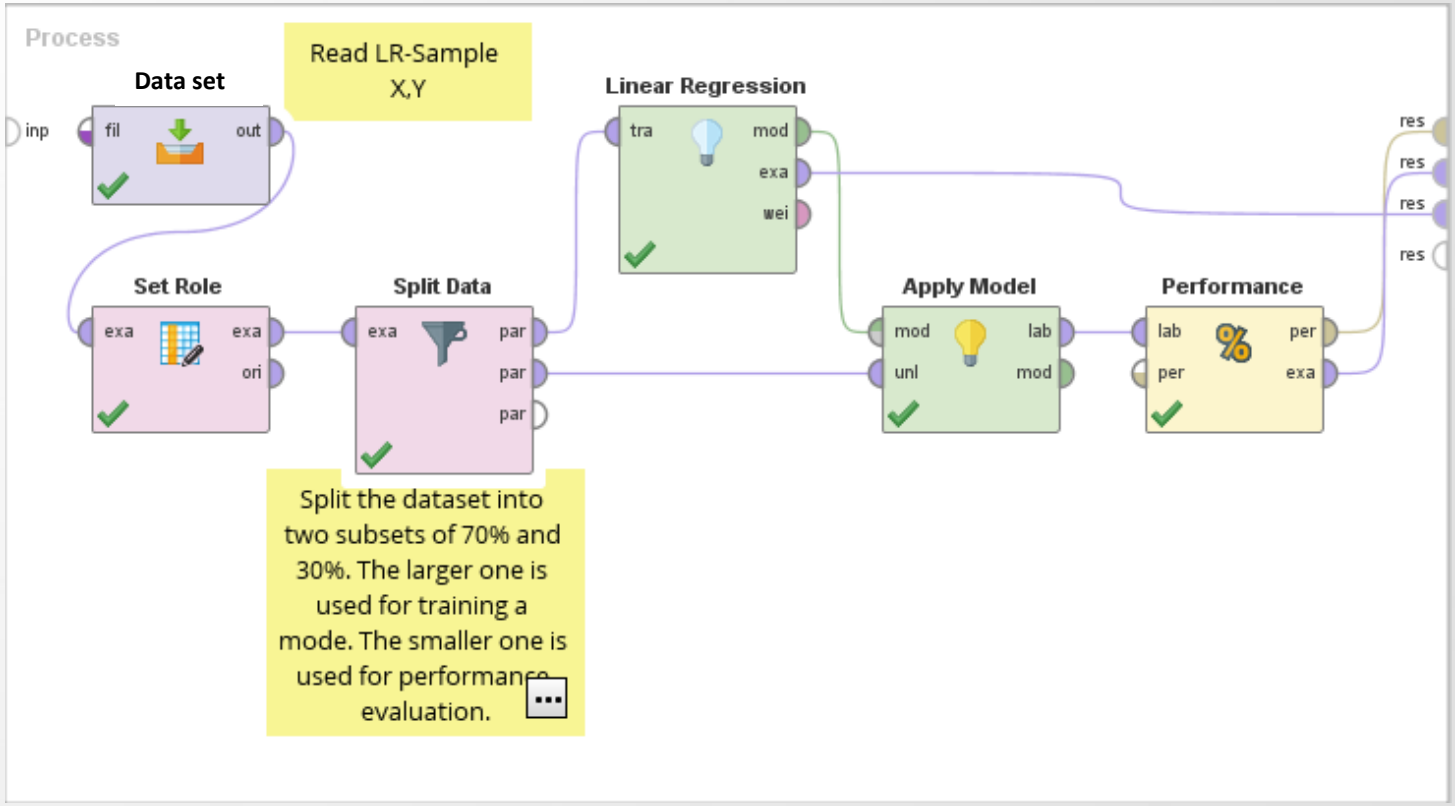
Trifact Wrangler



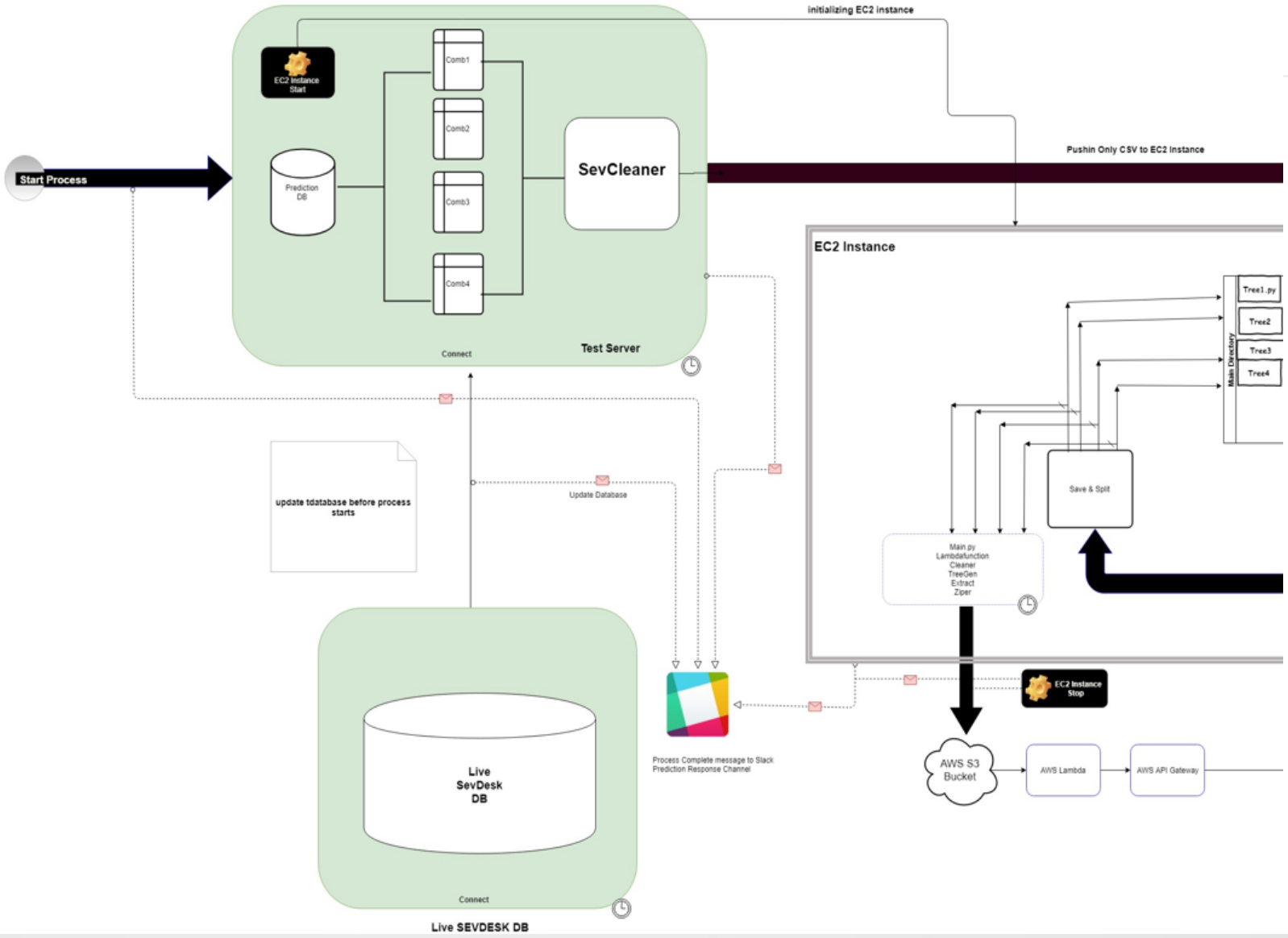
Data Spliter

sevProto

- Select Dataset
- Using Rapidminer
- Model Selection
- Model with high accuracy
- Testing with actual data



Prediction Automation with EC2 Instance



#pebe_prediction

☆ | 4 | 0 | Add a topic

Initializing EC2 Instance.....
.....
.....
EC2 instance started.....
Database update initialized...
process Data update start
Database updated
removing null accounting_types from Database
Database connection closed
Database Connection start
Tree1 Transfer process ON
Transfer complete
SUCCESS: All Folders Updated in EC2 INSTANCE
Tree Generation will start in EC2

```
Tree Generation Process Start
```

```
Reading from CSV file
```

CSV Loaded
Dataframe ready
Supplier Dictionary ready
FOC Dictionary ready
Train and Target ready

> KNN Score 75.82417582417582

KNN file stored with joblib protocol 2
Cross Validation Process
Cross Validation complete
RF Score 76.55677655677655

> Bagging with Bagging_knn file stored

> Bagging with Bagging_DT file stored

> RF file stored

> zip complete

```
Automation successfull
```

Boto connect ON

> S3 Bucket Pebe-rel-2 updated

EC2 Instance Shut Down

Live Demonstration

Thanks for your attention!

Questions?