Turnover prediction

AutoMate

Turnover prediction (POC)

Visma AutoMate | August 2023

©AUTOMATE

Agenda

1. **Introduction**
    - Why we’re doing that?
2. **Data analysis**
    - Data preprocessing
    - Features engineering
3. **Modeling**
    - Algorithms
    - Evaluation metrics
4. **Results**
    - Model testing
5. **Improvements**
    - Next steps

Introduction

Retaining your talent is one of the most essential aspect of building a thriving business. The key element to avoid turning is to **spot the “red flags” at the right time**. Companies struggle with timing of taking proper actions to prevent employee turnover. And that’s why we’re exploring possibilities in that area. The main goal of turnover prediction analysis is to identify problem areas and **take preventive steps to retain your employees**.

Data analysis

- **Data source:**
  - SR-Bank
- **Data filtering:**
  - Keeping only regular “full-time” employees
  - Keeping only “Ordinært” form of employment
  - Testing on data that were not seen by the model during the training phase
- **Data cleaning:**
  - Drop not relevant information from data
- **Data transformation:**
  - All columns in data needs to be in numeric format, because computers understand only numbers, right ? :)

Feature engineering

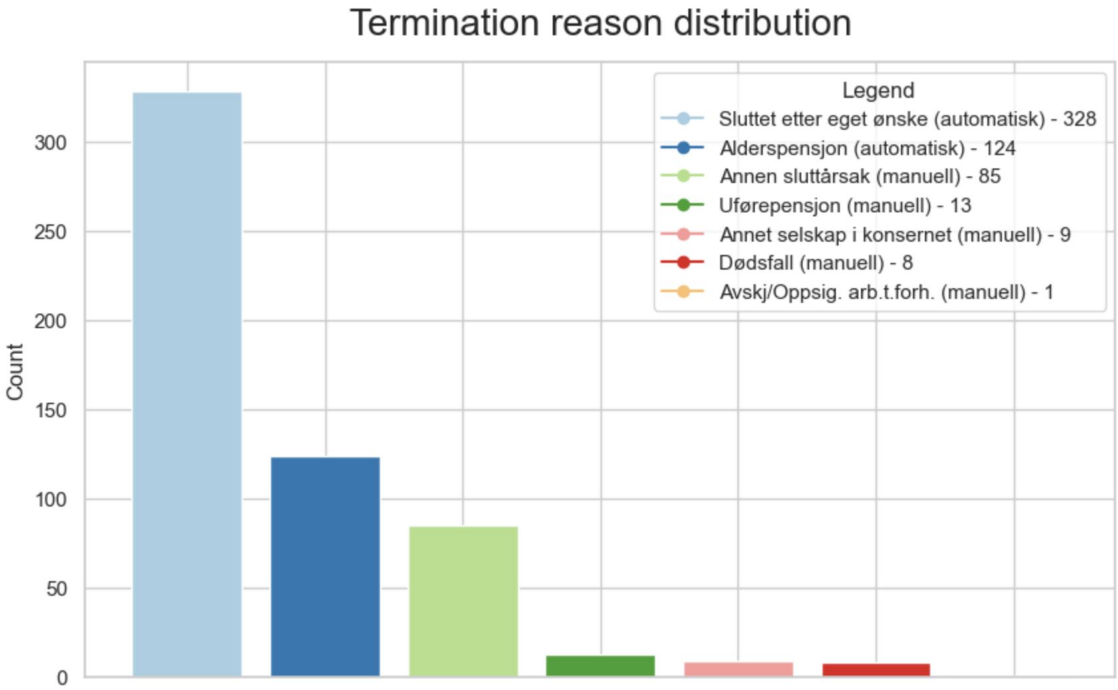
New fields derived from input data

- Employee groups based on salary and age
- Movement score - how many position codes has employee had
- Average salary on position, in position group, in age group etc..
- Average employee age on position, average job duration etc..

Dropped fields

- Unique columns (addresses, IDs)
- Constant columns
- Datetime columns

Why do employees leave?



Which ones leave the most?

```
original_df[original_df['position_group_code'] == '6']
```

	position_name	position_group_code
3	Aut. privatøkonomisk rådgiver	6
5	Aut. privatøkonomisk rådgiver	6

```
original_df[original_df['position_group_code'] == '8']
```

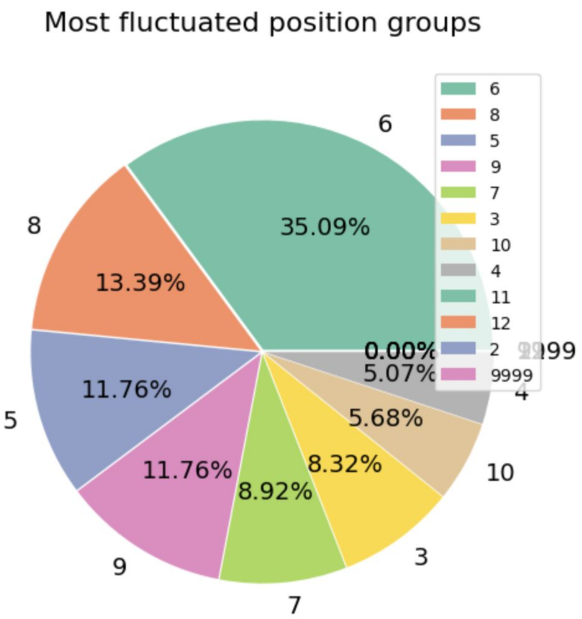
	position_name	position_group_code
1	Senior aut. privatøk. rådgiver	8
4	Senior bedriftsrådgiver SMB	8

```
original_df[original_df['position_group_code'] == '5']
```

	position_name	position_group_code
8	Fagrådgiver	5
13	Fagrådgiver	5

```
original_df[original_df['position_group_code'] == '9']
```

	position_name	position_group_code
10	Senior HR rådgiver	9
12	Leder profil	9



ML problem definition

Classification

- **Will employee quit in next 6 months?**
  - **output: decision (yes/no)**
- What is a confidence level of that prediction?
- Goal is to classify case as:
  - **Positive** - employee will quit in next 6 months
  - **Negative** - employee will stay in next 6 months

VISMA

Classification model

Goal:

- Predict if employee will quit in next 6 months

Algorithms that we used:

- Decision tree
- Random forest
- XGBoost

Evaluation metrics that we used:

- **Confusion matrix** - Performance measurement for classification model
- **Balanced accuracy** - How well a classifier identifies or excludes leaving employees
- **Precision** - probability of employee to leave in case of positive case classification
- **Recall** - probability of leaving employee detection
- **F1** - harmonic mean of precision and recall

What have we tested on?

```
df_test[df_test['resigned_in_6m']==1].tail()
```

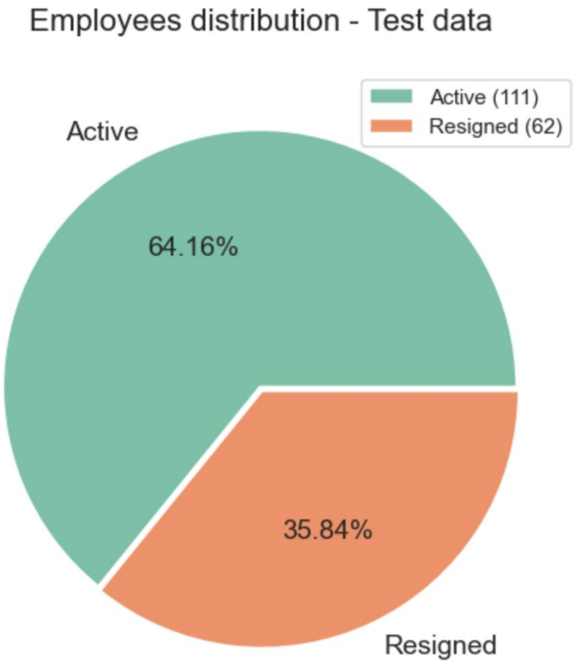
sex	age	nationality	civil_status	org_unit	position_group_code	position_code	movement_score	months_active	salary	salary_group	avg_pos_salary	avg_age_on_pos	resigned_in_6m
1	31	9	4	10290	7	11033	4	91	586253.0	5	646394	45	1
0	62	9	4	11560	6	11032	2	529	527182.0	5	571260	47	1
1	29	9	4	31000	6	10082	2	102	510987.0	5	638524	42	1
1	47	9	4	19850	8	11034	1	11	675436.0	6	771772	44	1
0	31	9	1	800	6	11002	1	60	352976.0	3	424558	38	1

```
df_test[df_test['resigned_in_6m']==0].tail()
```

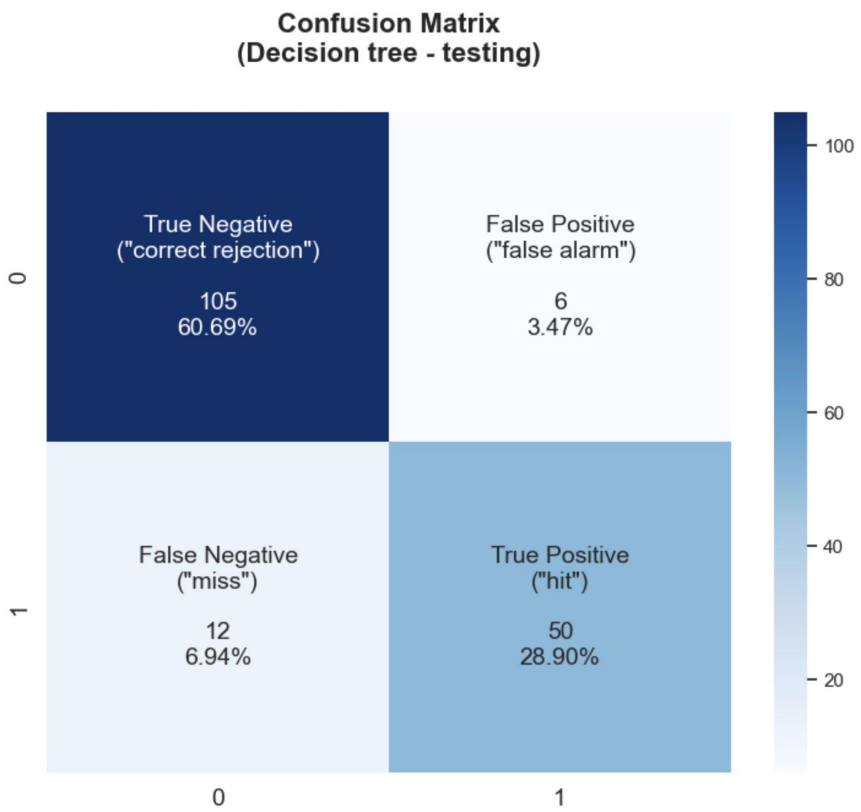
sex	age	nationality	civil_status	org_unit	position_group_code	position_code	movement_score	months_active	salary	salary_group	avg_pos_salary	avg_age_on_pos	resigned_in_6m
0	39	9	1	24500	8	10081	2	150	663814.0	6	764833	48	0
0	56	9	1	10290	8	11034	2	449	772935.0	7	784830	44	0
1	48	9	4	27020	7	11009	1	166	785321.0	7	520226	44	0
0	26	9	4	11240	5	10083	1	20	555851.0	5	536184	38	0
0	42	9	1	11910	10	10003	2	136	1100000.0	9	1179092	47	0

What have we tested on?

173 separated cases that were not used during a training phase of the model



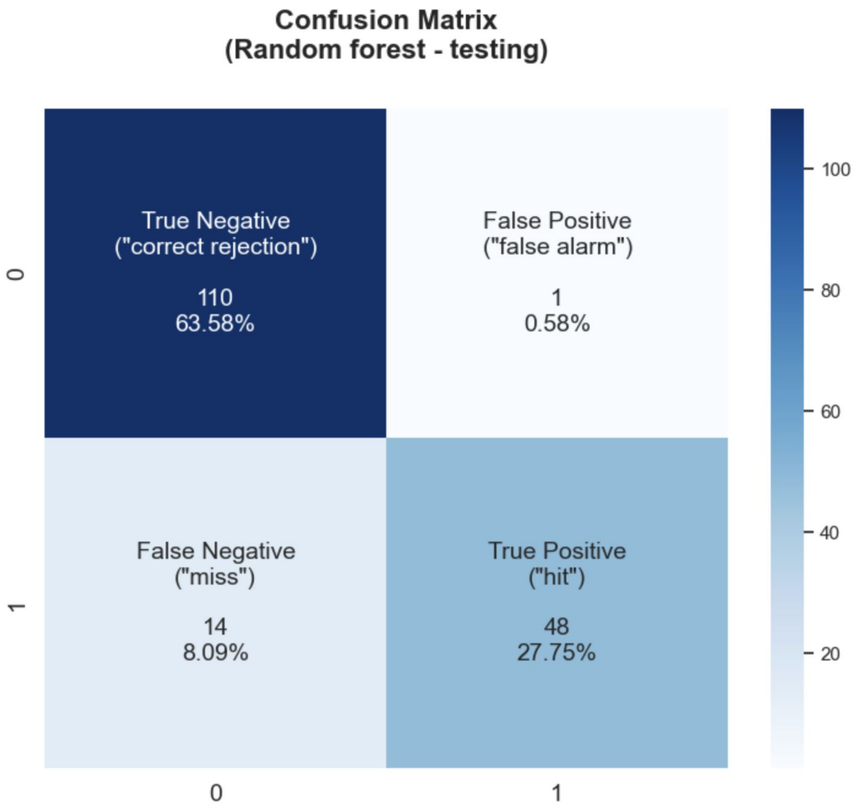
Classification - Decision tree



Classification Score Table
(Decision tree - testing)

Metric	Score
Balanced accuracy	0.876
Precision	0.893
Recall	0.806
F1	0.847

Classification - Random forest



Classification Score Table (Random forest - testing)

Metric	Score
Balanced accuracy	0.883
Precision	0.98
Recall	0.774
F1	0.865

Classification - XGBoost

Confusion Matrix
(XGboost - testing)

Classification Score Table
(XGboost - testing)

Metric	Score
Balanced accuracy	0.915
Precision	0.981
Recall	0.839
F1	0.904

What to do next?

- **Improvements:**
  - **Data from external sources -** How many similar open positions are available at the moment in specific region?
  - **More customer data** - larger dataset = more information to train model on (should result in better model)
- **Needs to be clarified:**
  - **What exactly we want to predict?**
    - Define unambiguous requirements on the solution
    - By being clear on this we'll be able to decide how to proceed and what techniques to use
  - **What is acceptable accuracy of model?**

VISMA

Respect

Reliability

Innovation

Competence

Team spirit