HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN PYTHON

Introduction to HR analytics

Hrant Davtyan
Assistant Professor of Data Science
American University of Armenia

What is HR analytics?

- Also known as People analytics
- Is a data-driven approach to managing people at work.

Problems addressed by HR analytics

- Hiring/Assessment
- Retention
- Performance evaluation
- Learning and Development
- Collaboration/team composition
- Other (e.g. absenteeism)

Employee turnover

- Employee turnover is the process of employees leaving the company
- Also known as employee attrition or employee churn
- May result in high costs for the company
- May affect company's hiring or retention decisions

Course structure

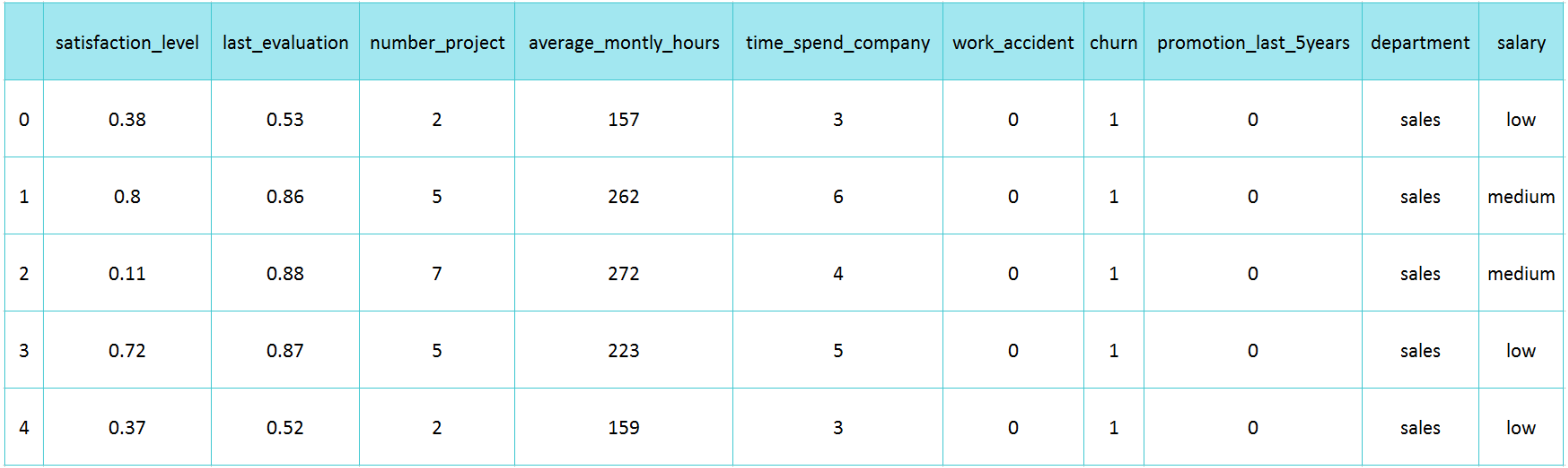
1. Describing and manipulating the dataset
2. Predicting employee turnover
3. Evaluating and tuning prediction
4. Selection final model

The Dataset

```
In  [1]: import pandas as pd
           data = pd.read_csv("turnover.csv")

In  [2]: data.info()

Out [2]:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
satisfaction_level       14999 non-null float64
last_evaluation          14999 non-null float64
number_project           14999 non-null int64
average_montly_hours     14999 non-null int64
time_spend_company       14999 non-null int64
work_accident            14999 non-null int64
churn                    14999 non-null int64
promotion_last_5years    14999 non-null int64
department               14999 non-null object
salary                   14999 non-null object
dtypes: float64(2), int64(6), object(2)
memory usage: 1.1+ MB
```

The Dataset (cont'd)

```
In [1]: data.head()
```



	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	work_accident	churn	promotion_last_5years	department	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.8	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low

Unique values

```
In [1]: print(data.salary.unique())


array(['low', 'medium', 'high'], dtype=object)
```

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN PYTHON

Let's practice!

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN PYTHON

Transforming categorical variables

Hrant Davtyan
Assistant Professor of Data Science
American University of Armenia

Types of categorical variables

- Ordinal - variables with two or more categories that can be ranked or ordered
    - Our example: **salary**
    - Values: low, medium, high
- Nominal - variables with two or more categories with **do not** have an instrinsic order
    - Our example: **department**
    - Values: sales, accounting, hr, technical, support, management, IT, product_mng, marketing, RandD

Encoding categories (salary)

```
In [1]: # Change the type of the "salary" column to categorical
        data.salary = data.salary.astype('category')

In [2]: # Provide the correct order of categories
        data.salary = data.salary.cat.reorder_categories(['low',
                                                          'medium',
                                                          'high'])

In [3]: # Encode categories with integer values
        data.salary = data.salary.cat.codes
```

Old values	New values
low	0
medium	1
high	2

Getting dummies

```
In [1]: # Get dummies and save them inside a new DataFrame
        departments = pd.get_dummies(data.department)
```

Example output

IT	RandD	accounding	hr	management	marketing	product_mng	sales	support	technical
0	0	0	0	0	0	0	0	0	1

Dummy trap

```
In [1]: departments.head()
```

IT	RandD	accounding	hr	management	marketing	product_mng	sales	support	technical
0	0	0	0	0	0	0	0	0	1

```
In [1]: departments = departments.drop("technical", axis = 1)
In [2]: departments.head()
```

IT	RandD	accounding	hr	management	marketing	product_mng	sales	support
0	0	0	0	0	0	0	0	0

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN PYTHON

Let's practice!

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN PYTHON

Descriptive Statistics

Hrant Davtyan
Assistant Professor of Data Science
American University of Armenia

Turnover rate

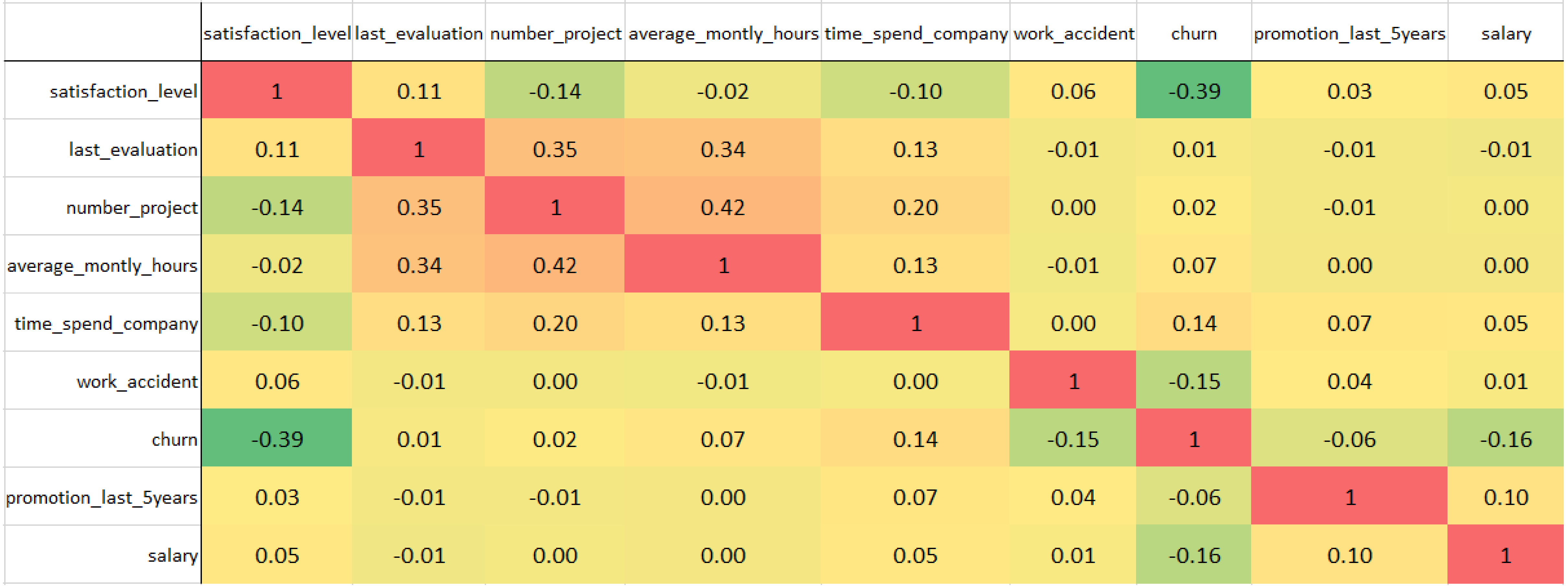
```
In [1]: # Get the total number of observations and save it
        n_employees = len(data)

In [2]: # Print the number of employees who left/stayed
        print(data.churn.value_counts())

In [3]: # Print the percentage of employees who left/stayed
        print(data.churn.value_counts()/n_employees*100)

Out [3]:

0    76.191746
1    23.808254
Name: churn, dtype: float64
```

Summary

Stayed	Left
76.19%	23.81%

Correlations

```
In [1]: import matplotlib.pyplot as plt
In [2]: import seaborn as sns
In [3]: corr_matrix = data.corr()
In [4]: sns.heatmap(corr_matrix)
In [5]: plt.show()
```



	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	work_accident	churn	promotion_last_5years	salary
satisfaction_level	1	0.11	-0.14	-0.02	-0.10	0.06	-0.39	0.03	0.05
last_evaluation	0.11	1	0.35	0.34	0.13	-0.01	0.01	-0.01	-0.01
number_project	-0.14	0.35	1	0.42	0.20	0.00	0.02	-0.01	0.00
average_montly_hours	-0.02	0.34	0.42	1	0.13	-0.01	0.07	0.00	0.00
time_spend_company	-0.10	0.13	0.20	0.13	1	0.00	0.14	0.07	0.05
work_accident	0.06	-0.01	0.00	-0.01	0.00	1	-0.15	0.04	0.01
churn	-0.39	0.01	0.02	0.07	0.14	-0.15	1	-0.06	-0.16
promotion_last_5years	0.03	-0.01	-0.01	0.00	0.07	0.04	-0.06	1	0.10
salary	0.05	-0.01	0.00	0.00	0.05	0.01	-0.16	0.10	1

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN PYTHON

Let's practice!