

**MINISTERUL EDUCAȚIEI, CULTURII ȘI CERCETĂRII AL REPUBLICII
MOLDOVA**

**Universitatea Tehnică a Moldovei
Facultatea Calculatoare, Informatică și Microelectronică
Departamentul Ingineria Software și Automatică**

**Admis la susținere
Șef departament
Ion Fiodorov, Conferențiar universitar, doctor în informatică**

_____ 2021
"_____"

Data-Driven HR: Understanding and predicting employee turnover

Master Thesis

Student: Brinza Ana-Maria, TIA-191M

Supervisor: Nistor Grozavu, lector sup.

Chișinău 2021

ADNOTARE

Brînză Ana-Maria. Data-Driven HR: Understanding and Predicting employee turnover

Chișinău, 2021

Structura tezei: Lucrarea conține adnotări în limbile română și engleză, cuprins, introducere, 4 capitole, concluzii, bibliografie.

Cuvintele-cheie: churn, predictive analytics, machine learning, NaN values, feature encoding, classification, regression, logistic regression, decision trees, random forest, prediction, accuracy, recall, ROC curve, AUC(area under the curve).

Domeniul de studiu: Analiza predictivă a datelor și învățare automată supravegheată

Scopul lucrării: Analiza datelor și dezvoltarea unui model de “machine learning” pentru a înțelege factorii care influențiază asupra plecării angajaților din companie și a prezice care sunt angajații cu o probabilitate mare de a părăsi compania.

Obiectivele lucrării: Analiza datelor istorice și extragerea informației relevante din date. Aplicarea modelelor statistice asupra datelor pentru a determina angajații cu un risc sporit de plecare din companie. Determinarea factorilor decisivi în plecarea angajaților. Studiarea și cercetarea a sel puțin două model statistice și analiza rezultatelor acestora. Propunerea unei strategii pentru reținerea angajaților în companie.

Valoarea teoretică a lucrării: Definirea datelor și a procesului de pregătire a datelor pentru modelare. Descrierea procesului de “machine learning” și a două modele bine cunoscute: logistic regression și random forest. Descrierea matematică a modelelor statistice.

Valoarea aplicativă a lucrării: Analiza și vizualizarea datelor prin intermediul tool-urilor și framework-urilor dedicate în limbajul de programare python. Procesarea datelor în python și pregătirea lor pentru modelare. Dezvoltarea modelelor de predicție în python (logistic regression și random forest).

ADNNOTATION

Brînză Ana-Maria. Data-Driven HR: Understanding and Predicting employee turnover

Chişinău, 2021

Thesis structure: The thesis contains annotations in Romanian and English language, contents, introduction, 4 chapters, conclusions, bibliography.

Keywords: churn, predictive analytics, machine learning, NaN values, feature encoding, classification, regression, logistic regression, decision trees, random forest, prediction, accuracy, recall, ROC curve, AUC(area under the curve).

Study domain: Supervised machine learning and predictive analytics.

Scope: Data analysis and development of a machine learning model for a better understanding of the most influential features of employee turnover. Predicting employee turnover.

Objectives: Analyzing historical data and extracting relevant information from data. Apply machine learning algorithms to the given dataset in order to predict employee turnover. Determine the most influential factors that cause employee turnover. Deep analysis of at least two machine learning algorithms. Analyzing the obtained results and evaluate the models. Define a high-level retention strategy for employee turnover.

Theoretical value of the thesis: Data set description and definition. Data preparation process description. Describing what machine learning is and types of machine learning. Mathematical description of two classification models: logistic regression and random forest.

Practical value of the thesis: Data analysis and visualization in python (dedicated tools and frameworks). Data pre-processing in python and preparing the data for machine learning modeling. Developing the prediction models in python (logistic regression and random forest).

TABLE OF CONTENTS

INTRODUCTION.....	1
I BUSINESS ANALYSIS AND UNDERSTANDING.....	2
1.1 Problem Definition	2
1.2 Employee Turnover State of Art.....	2
1.3 Outline of the report.....	4
II DATA UNDERSTANDING AND PREPARATION	5
2.1 Data Understanding	5
2.2 Data set description.....	5
2.3 Data insights	9
2.4 Data Visualization	11
2.4.1 Visualizing the employees who left vs those who stayed	11
2.4.2 Visualizing time spend at the company.....	12
2.4.3 Visualizing the distribution of the data for every feature	13
2.4.4 Visualizing Data Distribution by Department.....	15
2.4.5 Visualizing Data Distribution by employee's category	15
2.4.6 Visualizing the Data Distribution by employee's gender	16
2.4.7 Visualizing Data Distribution by employee's contract type	16
2.4.8 Visualizing Data Distribution by overtime.....	17
2.4.9 Visualizing Data Distribution for the promotion for the last 5 years	18
2.4.10 Visualizing Data Distribution by employee's age	18
2.4.11 Visualizing Data Distribution by average monthly hours	19
2.4.12 Visualizing Data Distribution by seniority in the company	20
2.4.13 Visualizing Data Distribution by training time in the last 5 years	21
2.4.14 Data Correlation.....	22
2.4.15 Data Analysis and Visualization Summary	24
2.5 Data Pre-Processing	25

2.5.1 Data Quality	25
2.5.1.1 Missing Data	26
2.5.1.2 Inconsistent values	30
2.5.1.3 Duplicate values.....	30
2.5.2 Feature Aggregation.....	30
2.5.3 Feature Encoding	30
2.5.4 Feature Scaling.....	33
2.5.4.1 Normalization	34
2.5.4.2 Standardization	34
2.5.5 Train and Test sets split	35
2.5.6 Data Pre-Processing Summary	36
2.6 Data anonymization and confidentiality	37
2.6.1 Pseudonymization	37
2.6.2 Data Masking.....	38
2.6.3 Generalization.....	39
2.6.4 Data swapping	39
2.6.5 Data perturbation	39
2.6.6 Synthetic data.....	40
2.6.7 Disadvantages of Data Anonymization	40
III METHODOLOGY	41
3.1 Logistic Regression	42
3.1.1 Model Development and Prediction	43
3.1.2 Model Evaluation	44
3.1.2.1 Accuracy.....	44
3.1.2.2 Precision	44
3.1.2.3 Recall	45
3.1.2.4 Confusion Matrix	46

3.1.2.5 ROC AUC curve.....	47
3.1.3 Advantages.....	48
3.1.4 Disadvantages	48
3.2 Random Forest.....	49
3.2.1 Model Development and Prediction	51
3.2.2 Model Evaluation	52
3.2.2.1 Accuracy.....	53
3.2.2.2 Precision	53
3.2.2.3 Recall	53
3.2.2.4 Confusion Matrix	54
3.2.2.5 ROC AUC Curve	55
3.2.3 Advantages.....	56
3.2.4 Disadvantages	56
3.3 Other Machine Learning Models	56
IV RESULTS AND DISCUSSION.....	59
4.1 Strategic Retention Plan.....	59
CONSLUSION.....	61
REFERENCES	62

INTRODUCTION

Artificial intelligence and Machine Learning tools made great progress for the past decade and the technological development registers unprecedented changes, at an exponential scale. Artificial Intelligence and ML is largely applied to 3D-printing, to Internet of Things (IoT), medical sciences ,transportation, political science, as well as many other fields.

The industry is being continuously challenged and improved, in the aim of making the life of populations across the globe better.

Emerging technologies such as “big data”, “cloud computing”, “artificial intelligence ” are taking over the market and have a significant impact on the operational model of many enterprises. For some, the business model might slightly change, some processes are automated, more data is available and real-data processing is possible. Given the volume of the available data, it’s variety, veracity and the value [Big Data 5V definition] it can bring - ‘data-driven’ decisions can be made and more and more enterprises tend to have a data-driven strategy.

The ‘data-driven’ term describes a decision-making process which involves collecting data, extracting patterns and facts from that data, and utilizing those facts to make inferences that influence decision-making. [1] Data-driven decision making is the process of making organizational decisions based on actual data rather than intuition or observation alone.

Companies all over the world tend to define data-driven strategies, as these can give them more insights over the business and can consequently improve the revenues of the company. While data-driven strategy is widely encountered on the market, data-driven HR is a new emerging term. HR domain is no exception. HR will have a unique role to play in this data- and AI-driven world.

When we say data-driven HR, we are referring to HR data and the valuable insights it can generate.

In this paper, one specific problem will be analyzed: employee turnover. It’s a problem most of the companies are facing. A deep analysis will be performed on data in order to understand and predict employee turnover.

REFERENCES

1. *Northeastern University*. Data-Driven Decision Making ©2019, [cited 12.11.2020]. Available: www.northeastern.edu/graduate/blog/data-driven-decision-making
2. HEATHER, Boushey, GLYNN, Sarah Jane. *There Are Significant Business Costs to Replacing Employees*[online]. 2012 [cited 02.12.2020]. Available: www.americanprogress.org/wp-content/uploads/2012/11/CostofTurnover.pdf
3. *Forbes*. Future Of People Analytics ©2020, [cited 02.12.2020]. Available: www.forbes.com/sites/bernardmarr/2020/02/14/future-of-people-analytics-what-lies-ahead-for-data-driven-hr
4. *IBM*. How chatbots can help reduce customer service costs by 30% ©2017, [cited 02.12.2020]. Available: www.ibm.com/blogs/watson/2017/10/how-chatbots-reduce-customer-service-costs-by-30-percent/
5. *SelectSoftware Reviews*. The Top 12 Best Recruiting and HR Chatbots ©2021, [cited 05.05.2021]. Available: www.selectsoftwarereviews.com/buyer-guide/hr-chat-bots
6. *XOR AI*. Watson Analytics ©2021, [cited 05.05.2021]. Available: www.xor.ai/
7. *IBM*. The Top 12 Best Recruiting and HR Chatbots ©2021, [cited 05.05.2021]. Available: www.ibm.com/partnerworld/program/benefits/watson-analytics-professional-extended-trial
8. Regulation (EU) 2016/679 (General Data Protection Regulation) in the current version of the OJ L 119, 04.05.2016
9. *imperva*. Data Masking ©2020, [cited 05.02.2021]. Available: www.imperva.com/learn/data-security/data-masking/
10. *imperva*. Data Security Anonymization ©2020, [cited 05.02.2021]. Available: <https://www.imperva.com/learn/data-security/anonymization/>
11. *KDnuggets*. How GDPR Affects Data Science ©2017, [cited 17.02.2021]. Available: <https://www.kdnuggets.com/2017/07/gdpr-affects-data-science.html>
12. DEBORAH J., Rumsey, *Statistics for Dummies* [online]. Los Angeles; London: SAGE Publications, 2009 [cited 02.03.2021]. Available: <https://www.dummies.com/education/math/statistics/statistics-for-dummies-cheat-sheet/>
13. The youth in the Republic of Moldova in 2019 from 11.08.2020 In: *National Bureau of Statistics of the Republic of Moldova*. Available: statistica.gov.md/

14. *Data Camp*. Machine Learning modeling ©2021, [cited 20.04.2021]. Available: www.datacamp.com
15. *Machine Learning Mastery*. Master Machine Learning Algorithms modeling ©2021, [cited 20.04.2021]. Available: <https://machinelearningmastery.com>