DSCI 5260 - 002

Business Process Analytics

Dr. Sameh Shamroukh

Analysis of Employee Performance in Software business based on Historic Data.

Table of Contents

Acknowledgements:

We express our sincere thanks to Dr. Sameh Shamroukh for his important guidance, assistance, and feedback over the duration of this project. The course and outcomes of our study really benefited from his perceptive evaluation, insightful suggestions, and continuous encouragement.

Furthermore, we would like to convey our deepest gratitude to all individuals in Group-03 for their significant dedication and commitment. The significant contribution they made greatly influenced the result of our research. The project's success can be attributed mostly due to the team members' different skills and viewpoints.

1. Abstract

Maintaining and increasing employee efficiency is a crucial and essential aspect in each business sector. Delegate effectiveness inside the program division incorporates an extensive effect on a grouping of boundaries, counting responsibility, work satisfaction, and salary increases. A dependable and strong technique of studying proficiency requires the prospect of a couple of parts, for example, execution rating, productivity score, and week after week responsibility. Our issue clarification is to survey staff proficiency using the principal modern advancements and approaches. This ask about will be significantly important to program overseers and human resource specialists. To make assessment, endlessly graphs, the laborers' set of experiences records should be gotten and consolidated into the ongoing system . Python, a strong anyway harmonious programming lingo perceived for its robotization features, is used to help this muddled exchange assessment handle go all the more easily. Plots and diagrams can be used to exhibit and picture illustrative data and reports. They are particularly profitable for conveying a quantifiable and mathematical depiction of a worker's productivity to a board. In an information driven approach, the proposed plan tends to this difficulty by using various quantifiable strategies and the Python programming vernacular. The basic strides in data arranging consolidate encoding all out variables and managing lost values. The methodology use the Variance Extension Figure (VIF) to dissect the doubts and distinguish multicollinearity between the independent variables. Scatter plots are utilized to decide whether linearity exists, though match plots and relationship heatmaps are utilized to examine the relationships between factors.

Keywords: Productivity, employee, job satisfaction, age, Python, historic data, analysis.

2. Introduction

Staff turnover could be a gigantic concern in today's corporate world, costing organizations cash and causing them dissatisfaction. Understanding the factors that lead to representative turnover is basic for creating viable maintenance procedure. The reason of this request is to utilize information analytics to discover and clarify the key components that will impact steady loss rates in a firm, especially within the computer program industry.

IT industry is facing issues to standardize the salary hike recommendations by eliminating the human politics by managers or the human resources. Our research study focuses on the automation of the decision of humans by a code generated decision. Machine learning algorithms are the most efficient in making such type of decisions and Python is a user-friendly platform independent language which helps to write an easy understandable code. Hence, we have developed Python code for the salary hike decisions to be taken base on the results we got by running the code.

We start by bringing in the Python modules required for information examination and making a anecdotal dataset for advance examination. The dataset includes characteristics such as age, remove from domestic, month to month pay, and work fulfillment that are thought to affect an employee's choice to remain or stopped a firm.

Earlier to starting the modeling step, we do information pretreatment to ensure that the dataset is error-free and satisfactory for examination. This incorporates filling in lost information, deciphering categorical factors to numerical values, and guaranteeing that our information fits the calculated relapse measurable assumptions.

Taking after the creation of our dataset, we go to exploratory information examination, where we probe for multicollinearity and show the relationships between the factors. We at that point apply a calculated relapse demonstrate to appraise the chance of steady loss based on our free variables. To affirm that our demonstrate meets the logistic relapse suspicions, we analyze its execution employing a extend of classification criteria and closely ponder the residuals.

Experiences that not as it were build up the predictors' measurable noteworthiness but moreover clarify their value for human assets administration are advertised by the research's conclusion. With the extreme objective of expanding representative maintenance, these discoveries will offer help for key decision-making, such as choosing whether to recommend wage increments based on levels of work fulfillment.

It is pivotal to recognize that in spite of the utilization of vigorous factual approaches in this think about, its appropriateness is limited by factors such as information accuracy, demonstrate straightforwardness, and the inborn challenges related with foreseeing complex human behavior. Regardless these challenges, the investigate gives a valuable system for applying information analytics to get it and address the issues that lead to representative steady loss, thus proposing whether or not to supply a pay increment.

2.1 Research Problem:

The study's major reason is to analyze efficiency levels among program laborers whereas taking into thought characteristics such as age, commuting length, month to month emolument, and work bliss. The current trouble begins from the elemental have to be protect and increment staff efficiency, especially within the computer program segment where it contains a significant impact on a assortment of parameters such as workload, work fulfillment, and worker maintenance.

Businesses must analyze and comprehend staff efficiency in arrange to preserve a sound work environment, distinguish openings for development, and make choices that will boost organizational execution. Within the program industry, boosting efficiency is basic since specialist effectiveness includes a coordinate affect on extend results and item improvement.

The energetic nature of labor, the complexity of the program industry, and the necessity for data-driven bits of knowledge to direct administrative choices are likely what impelled the inquire about address. It is crucial to comprehend and upgrade efficiency since it is fundamental to supply individuals with perfect conditions, rising innovation, and exceedingly sought-after work settings.

The recommended procedure approaches this challenge by utilizing a assortment of factual procedures and the Python programming dialect in a data-centric mold. Encoding categorical factors and managing missing values are the primary steps within the planning of information. To assess the suspicions and decide multicollinearity between the free factors, the method makes utilize of the Change Swelling Calculate (VIF). Scramble plots are utilized to decide whether

linearity is show, whereas combine plots and relationship heatmaps are utilized to examine the connections between the factors.

A calculated relapse demonstrate is utilized in this ponder to look at the interface between a few autonomous components and whittling down, which is an fundamental perspective of measuring worker efficiency. Besides, the calculated relapse information are looked into, and boxplots and other visual helps are utilized to explore the relationships between age, work fulfillment, and month to month salary.

The primary focus of the research is the necessity of improving employee productivity in the competitive, high-speed software sector. The approaches used, which include statistical analysis and visualizations, are intended to give managers in the software industry and human resource professionals a thorough grasp of the variables affecting productivity as well as insightful information to help them make strategic decisions.

Research objective:

The main focus of the project is to identify the productivity and suggest a decision on salary hike to be given for employees mainly in the IT industry where there are many uncertainties.

With an emphasis on variables including age, distance from home, monthly salary, and job happiness, the research goal is to examine and understand the efficiency of employees in the software business. Providing insightful information that can help stakeholders—especially managers and human resource specialists in the software industry—is the ultimate goal. The study's goal and possible advantages are broken down as follows:

Determine Productivity Elements:

Goal: Determine the factors that most significantly impact employee productivity, such as age, commute time from home to work, monthly income, and job satisfaction.

Benefit to Stakeholders: Stakeholders may better target their strategies and interventions on these areas after identifying the important factors, which will raise overall worker productivity.

Recognize how variables are related to one another:

The goal is to examine the relationships between the different factors. Examine the relationship between age and monthly salary, for instance, and work happiness.

Benefit to Stakeholders: A thorough understanding of these relationships enables stakeholders to make well-informed decisions regarding the allocation of resources, employee engagement, and satisfaction efforts. It also helps stakeholders understand the dynamics at play within the workforce.

Determine the linearity and multicollinearity:

Goal: Evaluate the linearity of the relationship between independent variables and attrition and do assumption testing to look for multicollinearity

Benefit to Stakeholders: Predictions and recommendations are more accurate when regression models are reliable and relationships are understood to be linear, which facilitates better decision-making.

Identify Variables:

Goal: Investigate the relationship between attrition and a few independent variables.

Benefit to Stakeholder: Information about factors influencing attrition can be used to inform the development of retention strategies. By addressing concerns about age, distance from home, monthly income, and job satisfaction, attrition rates can be reduced and critical personnel can be retained.

Improve representation by Visualizing Data:

Objective: To illustrate complex data relationships, use visualizations including boxplots, pair plots, and scatter plots.

Benefit to Stakeholder: A clear and understandable means of communicating findings is made possible by visualizations, which benefits stakeholders. These visualizations help improve communication between various organizational levels by enabling stakeholders to use them in presentations, reports, and debates.

Support Informed Decision-Making:

Objective: Use statistical methods, such as logistic regression, to examine data and come at insightful judgments.

Benefit to Stakeholder : Advantage for interested parties Making well-informed decisions is essential to maximizing worker output. Stakeholders can apply targeted strategies and interventions to improve overall productivity by using the evidence-based insights that the analysis's conclusions give.

2.2 Research Questions:

How will indirect data such as gender and age affect productivity?

Hypothesis:

Age and Gender: Although they are regarded as indirect data, age and gender can have an impact on productivity in a number of ways:

Diversity in the Workforce: Diverse teams can stimulate innovation and creativity.

Experience and Expertise: As people age, they often gain experience, which has a positive impact on productivity. However, depending on the nature of the job, physical aspects of it may become more difficult as one ages.

Work-Life Balance: The expectations and responsibilities that individuals of different ages and genders have outside of the workplace can have an impact on their productivity there.

How to calculate productivity based on external factors such as pay hike and business travel?

Pay Hike: Increasing employee pay can have a positive impact on their motivation and morale, which will increase output. An output per hour metric, used both before and after the pay increase, can be used to quantify it.

Business travel: Although it can result in time away from essential work activities and networking opportunities, it can also cause fatigue. Keeping an eye on work output during times when travel frequency varies can help determine productivity.

Does Training impact performance rating?

Training and Development: Employee skill development is a key component of training programs, which raises employee productivity. To evaluate its effect: Pre- and Post-Training Assessments: Determine whether there has been a discernible improvement in performance ratings before and after training. Long-Term Tracking: Evaluate performance over an extended duration to determine whether training has resulted in long-term increases in output and performance evaluations.

3. Literature Review

3.1 Theories Grounding the Organizational Problem:

1. IT Capabilities and Organizational Performance Theory (Liu et al., 2013): This theory posits that the effectiveness of IT infrastructure significantly influences firm performance. In an organizational context, the assimilation and use of IT skills, particularly in the software sector, are crucial. For instance, the use of Python for data analysis can be seen as an embodiment of IT capabilities, affecting staff productivity and the organization's overall efficiency.
2. Digital Technology and Employee Compensation Theory (Yuan et al., 2023): This theory explores the impact of digital technology on employee remuneration. It suggests that the adoption of digital technology in the workplace can alter pay scales and work conditions. Analyzing variables like monthly salary and commuting distance through digital means reflects how technology advancements are reshaping employment landscapes.
3. Employee Attrition and Predictive Modeling Theory (Chung et al., 2023): This theory emphasizes the importance of understanding and predicting employee turnover using data-driven models. In an organizational setting, the application of logistic regression models to predict attrition aligns with this theory, highlighting how staff turnover can be a critical factor in assessing organizational productivity.
4. Personalized Corporate E-Learning Theory (Shikov, 2018): This theory focuses on the impact of tailored e-learning programs on employee skills development. The practice of personalizing data treatment, such as managing categorical variables and addressing

missing values, can be seen as part of a broader strategy to enhance employee competencies and productivity through customized learning approaches.

5. Flexible Work Scheduling and Employee Satisfaction Theory (Kiwanuka et al., 2021): This theory examines the correlation between flexible working arrangements and employee happiness. The analysis of factors like age, commuting distance, and job satisfaction in relation to flexible work schedules indicates the potential of such arrangements to improve employee morale and, consequently, organizational productivity.

3.2 Supporting Examples:

- The use of Python in data analysis within the organization exemplifies the application of IT capabilities theory, influencing factors like employee productivity and operational efficiency.
- The implementation of digital tools for managing employee compensation, as seen in the analysis of salary and commuting data, reflects the influence of digital technology on workplace dynamics and employee remuneration.
- The deployment of predictive models, like logistic regression, for forecasting employee attrition demonstrates the practical application of predictive modeling in understanding workforce dynamics.
- The practice of data pre-treatment in organizational analyses, including handling missing values and encoding variables, aligns with the personalized e-learning theory, enhancing the efficacy of employee training and development programs.

- The examination of job satisfaction in relation to flexible working schedules in the organization's code mirrors the theory linking flexible work arrangements with employee happiness and overall organizational well-being.

In addition https://fortune.com/2023/03/13/artificial-intelligence-make-workplace-decisions-human-intelligence-remains-vital-careers-tech-gary-friedman/ , this machine learning and artificial intelligence are playing crucial role in decision making of employee related factors. In 2018, Amazon implemented an algorithm that automates the hiring process of employees. But due to the lack of quality input dataset, it has drastically failed by only hiring the male employees by picking up the resumes based on the gender. This was initially a failure but now the research has advanced and many resume picking algorithms are in trend. In replicate we have done our research to make automated decision for the salary hike recommendation.

4. Data Description:

- The dataset has been sourced from Kaggle and titled as "Employee Attrition and Factors."
- The dataset explores factors influencing employee productivity and attrition in the software industry, including age, distance from home, monthly pay, and job satisfaction.
- The dataset consists of 35 columns with 1000 -plus records.
- The link to the data source: https://www.kaggle.com/datasets/thedevastator/employee-attrition-and-factors

Dependent Variables:

Attrition

Independent Variables:

Age

Monthly Income

Job Satisfaction

Years at Company

Age	Attrition	BusinessTra	DailyRate	Department	DistanceFro	Education	EducationFie	EmployeeCo	EmployeeNu	Environment	Gender	HourlyRate	JobInvolvem	JobLevel	JobRole	JobSatisfacti	MaritalStatu	MonthlyInco	MonthlyRate	NumCompar	Over18	OverTime	PercentSalar	Performance	Relationship	StandardHou	StockOption	TotalWorkin	TrainingTime	WorkLifeBal	YearsAtComp	YearsInCurre	YearsSinceLa	YearsWithCurrManager
41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	2	Female	94	3	2	Sales Executi	4	Single	5993	19479	8	Y	Yes	11	3	1	80	0	8	0	1	6	4	0	5
49	No	Travel_Frequ	279	Research & D	8	1	Life Sciences	1	2	3	Male	61	2	2	Research Sci	2	Married	5130	24907	1	Y	No	23	4	4	80	1	10	3	3	10	7	1	7
37	Yes	Travel_Rarely	1373	Research & D	2	2	Other	1	4	4	Male	92	2	1	Laboratory T	3	Single	2090	2396	6	Y	Yes	15	3	2	80	0	7	3	3	0	0	0	0
33	No	Travel_Frequ	1392	Research & D	3	4	Life Sciences	1	5	4	Female	56	3	1	Research Sci	3	Married	2909	23159	1	Y	Yes	11	3	3	80	0	8	3	3	8	7	3	0
27	No	Travel_Rarely	591	Research & D	2	1	Medical	1	7	1	Male	40	3	1	Laboratory T	2	Married	3468	16632	9	Y	No	12	3	4	80	1	6	3	3	2	2	2	2
32	No	Travel_Frequ	1005	Research & D	2	2	Life Sciences	1	8	4	Male	79	3	1	Laboratory T	4	Single	3068	11864	0	Y	No	13	3	3	80	0	8	2	2	7	7	3	6
59	No	Travel_Rarely	1324	Research & D	3	3	Medical	1	10	3	Female	81	4	1	Laboratory T	1	Married	2670	9964	4	Y	Yes	20	4	1	80	3	12	3	2	1	0	0	0
30	No	Travel_Rarely	1358	Research & D	24	1	Life Sciences	1	11	4	Male	67	3	1	Laboratory T	3	Divorced	2693	13335	1	Y	No	22	4	2	80	1	1	2	3	1	0	0	0
38	No	Travel_Frequ	216	Research & D	23	3	Life Sciences	1	12	4	Male	44	2	3	Manufacturi	3	Single	9526	8787	0	Y	No	21	4	2	80	0	10	2	3	9	7	1	8
36	No	Travel_Rarely	1299	Research & D	27	3	Medical	1	13	3	Male	94	3	2	Healthcare R	3	Married	5237	16577	6	Y	No	13	3	2	80	2	17	3	2	7	7	7	7
35	No	Travel_Rarely	809	Research & D	16	3	Medical	1	14	1	Male	84	4	1	Laboratory T	2	Married	2426	16479	0	Y	No	13	3	3	80	1	6	5	3	5	4	0	3
29	No	Travel_Rarely	153	Research & D	15	2	Life Sciences	1	15	4	Female	49	2	2	Laboratory T	3	Single	4193	12682	0	Y	Yes	12	3	4	80	0	10	3	3	9	5	0	8
31	No	Travel_Rarely	670	Research & D	26	1	Life Sciences	1	16	1	Male	31	3	1	Research Sci	3	Divorced	2911	15170	1	Y	No	17	3	4	80	1	5	1	2	5	2	4	3
34	No	Travel_Rarely	1346	Research & D	19	2	Medical	1	18	2	Male	93	3	1	Laboratory T	4	Divorced	2661	8758	0	Y	No	11	3	3	80	1	3	2	3	2	2	1	2
28	Yes	Travel_Rarely	103	Research & D	24	3	Life Sciences	1	19	3	Male	50	2	1	Laboratory T	3	Single	2028	12947	5	Y	Yes	14	3	2	80	0	6	4	3	4	2	0	3
29	No	Travel_Rarely	1389	Research & D	21	4	Life Sciences	1	20	2	Female	51	4	3	Manufacturi	1	Divorced	9980	10195	1	Y	No	11	3	3	80	1	10	1	3	10	9	8	8
32	No	Travel_Rarely	334	Research & D	5	2	Life Sciences	1	21	1	Male	80	4	1	Research Sci	2	Divorced	3298	15053	0	Y	Yes	12	3	4	80	2	7	5	2	6	2	0	5
22	No	Non-Travel	1123	Research & D	16	2	Medical	1	22	4	Male	96	4	1	Laboratory T	4	Divorced	2935	7324	1	Y	Yes	13	3	2	80	2	1	2	2	1	0	0	0
53	No	Travel_Rarely	1219	Sales	2	4	Life Sciences	1	23	1	Female	78	2	4	Manager	4	Married	15427	22021	2	Y	No	16	3	3	80	0	31	3	3	25	8	3	7
38	No	Travel_Rarely	371	Research & D	2	3	Life Sciences	1	24	4	Male	45	3	1	Research Sci	4	Single	3944	4306	5	Y	Yes	11	3	3	80	0	6	3	3	3	2	1	2
24	No	Non-Travel	673	Research & D	11	2	Other	1	26	1	Female	96	4	2	Manufacturi	3	Divorced	4011	8232	0	Y	No	18	3	4	80	1	5	5	2	4	2	1	3
36	Yes	Travel_Rarely	1218	Sales	9	4	Life Sciences	1	27	3	Male	82	2	1	Sales Repres	1	Single	3407	6986	7	Y	No	23	4	2	80	0	10	4	3	5	3	0	3
34	No	Travel_Rarely	419	Research & D	7	4	Life Sciences	1	28	1	Female	53	3	3	Research Dir	2	Single	11994	21293	0	Y	No	11	3	3	80	0	13	4	3	12	6	2	11
21	No	Travel_Rarely	391	Research & D	15	2	Life Sciences	1	30	3	Male	96	3	1	Research Sci	4	Single	1232	19281	1	Y	No	14	3	4	80	0	0	6	3	0	0	0	0
34	Yes	Travel_Rarely	699	Research & D	6	1	Medical	1	31	2	Male	83	3	1	Research Sci	1	Single	2960	17102	2	Y	No	11	3	3	80	0	8	2	3	4	2	1	3
53	No	Travel_Rarely	1282	Research & D	5	3	Other	1	32	3	Female	58	3	5	Manager	3	Divorced	19094	10735	4	Y	No	11	3	4	80	1	26	3	2	14	13	4	8
32	Yes	Travel_Frequ	1125	Research & D	16	1	Life Sciences	1	33	2	Female	72	1	1	Research Sci	1	Single	3919	4681	1	Y	Yes	22	4	2	80	0	10	5	3	10	2	6	7
42	No	Travel_Rarely	691	Sales	8	4	Marketing	1	35	3	Male	48	3	2	Sales Executi	2	Married	6825	21173	0	Y	No	11	3	4	80	1	10	2	3	9	7	4	2
44	No	Travel_Rarely	477	Research & D	7	4	Medical	1	36	1	Female	42	2	3	Healthcare R	4	Married	10248	2094	3	Y	No	14	3	4	80	1	24	4	3	22	6	5	17
46	No	Travel_Rarely	705	Sales	2	4	Marketing	1	38	2	Female	83	3	5	Manager	1	Single	18947	22822	3	Y	No	12	3	4	80	0	22	2	2	2	2	2	1
33	No	Travel_Rarely	924	Research & D	2	3	Medical	1	39	3	Male	78	3	1	Laboratory T	4	Single	2496	6670	4	Y	No	11	3	4	80	0	7	3	3	1	1	0	0
44	No	Travel_Rarely	1459	Research & D	10	4	Other	1	40	4	Male	41	3	2	Healthcare R	4	Married	6465	19121	2	Y	Yes	13	3	4	80	0	9	5	4	4	2	1	3
30	No	Travel_Rarely	125	Research & D	9	2	Medical	1	41	4	Male	83	2	1	Laboratory T	3	Single	2206	16117	1	Y	No	13	3	1	80	0	10	5	3	10	0	1	8
39	Yes	Travel_Rarely	895	Sales	5	3	Technical De	1	42	4	Male	56	3	2	Sales Repres	4	Married	2086	3335	3	Y	No	14	3	3	80	1	19	6	4	1	0	0	0
24	Yes	Travel_Rarely	813	Research & D	1	3	Medical	1	45	2	Male	61	3	1	Research Sci	4	Married	2293	3020	2	Y	Yes	16	3	1	80	1	6	2	2	2	0	2	0
43	No	Travel_Rarely	1273	Research & D	2	2	Medical	1	46	4	Female	72	4	1	Research Sci	3	Divorced	2645	21923	1	Y	No	12	3	4	80	2	6	3	2	5	3	1	4
50	Yes	Travel_Rarely	869	Sales	3	2	Marketing	1	47	1	Male	86	2	1	Sales Repres	3	Married	2683	3810	1	Y	Yes	14	3	3	80	0	3	2	3	3	2	0	2
35	No	Travel_Rarely	890	Sales	2	3	Marketing	1	49	4	Female	97	3	1	Sales Repres	4	Married	2014	9687	1	Y	No	13	3	1	80	0	2	3	3	2	2	2	2
36	No	Travel_Rarely	852	Research & D	5	4	Life Sciences	1	51	2	Female	82	2	1	Research Sci	1	Married	3419	13072	9	Y	Yes	14	3	4	80	1	6	3	4	1	1	0	0
33	No	Travel_Frequ	1141	Sales	1	3	Life Sciences	1	52	3	Female	42	4	2	Sales Executi	1	Married	5376	3193	2	Y	No	19	3	1	80	2	10	3	3	5	3	1	3
35	No	Travel_Rarely	464	Research & D	4	2	Other	1	53	3	Male	75	3	1	Laboratory T	4	Divorced	1951	10910	1	Y	No	12	3	3	80	1	1	3	3	1	0	0	0
27	No	Travel_Rarely	1240	Research & D	2	4	Life Sciences	1	54	4	Female	33	3	1	Laboratory T	1	Divorced	2341	19715	1	Y	No	13	3	4	80	1	1	6	3	1	0	0	0
26	Yes	Travel_Rarely	1357	Research & D	25	3	Life Sciences	1	55	1	Male	48	1	1	Laboratory T	3	Single	2293	10558	1	Y	No	12	3	3	80	0	1	2	2	1	0	0	1
27	No	Travel_Frequ	994	Sales	8	3	Life Sciences	1	56	4	Male	37	3	3	Sales Executi	3	Single	8726	2975	1	Y	No	15	3	4	80	0	9	0	3	9	8	1	7
30	No	Travel_Frequ	721	Research & D	1	2	Medical	1	57	3	Female	58	3	2	Laboratory T	4	Single	4011	10781	1	Y	No	23	4	4	80	0	12	2	3	12	8	3	7
41	Yes	Travel_Rarely	1360	Research & D	12	3	Technical De	1	58	2	Female	49	3	5	Research Dir	3	Married	19545	16280	1	Y	No	12	3	4	80	0	23	0	3	22	15	15	8
34	No	Non-Travel	1065	Sales	23	4	Marketing	1	60	2	Male	72	3	2	Sales Executi	3	Single	4568	10034	0	Y	No	20	4	3	80	0	10	2	3	9	5	8	7
37	No	Travel_Rarely	408	Research & D	19	2	Life Sciences	1	61	2	Male	73	3	1	Research Sci	2	Married	3022	10227	4	Y	No	21	4	1	80	0	8	1	3	1	0	0	0
46	No	Travel_Frequ	1211	Sales	5	4	Marketing	1	62	1	Male	98	3	2	Sales Executi	4	Single	5772	20445	4	Y	Yes	21	4	3	80	0	14	4	3	9	6	0	8
35	No	Travel_Rarely	1229	Research & D	8	1	Life Sciences	1	63	4	Male	36	4	1	Laboratory T	4	Married	2269	4892	1	Y	No	19	3	4	80	0	1	2	3	1	0	0	1
48	Yes	Travel_Rarely	626	Research & D	1	2	Life Sciences	1	64	1	Male	98	2	3	Laboratory T	3	Single	5381	19294	9	Y	Yes	13	3	4	80	0	23	2	3	1	0	0	0
28	Yes	Travel_Rarely	1434	Research & D	5	4	Technical De	1	65	3	Male	50	3	1	Laboratory T	3	Single	3441	11179	1	Y	Yes	13	3	3	80	0	2	3	2	2	2	2	2
44	No	Travel_Rarely	1488	Sales	1	5	Marketing	1	68	2	Female	75	3	2	Sales Executi	1	Divorced	5454	4009	5	Y	Yes	21	4	3	80	1	9	2	2	4	3	1	3
35	No	Non-Travel	1097	Research & D	11	2	Medical	1	70	3	Male	79	2	3	Healthcare R	1	Married	9884	8302	2	Y	Yes	13	3	3	80	1	10	3	3	4	0	2	3

Research Method:

Data set has been taken from Kaggle. The data is sourced from Kaggle, which has data that is both well-qualified and highly qualified. The gathered data exhibits a wide range of information and is highly valuable for the project and further research.

Age, Monthly Income, Job satisfaction and Years at Company are the independent variables which play a crucial role in productivity of an employee. Although these are the independent variables but each variable impacts productivity in its own way but ultimately makes the attrition collectively. Hence, we can prove that attrition is the dependent variable in this code we have developed.

Variables:

The analysis focuses on attrition as the primary dependent variable, with age, distance from home, monthly income, and job satisfaction as the primary independent variables. The independent factors may have the ability to predict, explain, or exert effect on the dependent variable, attrition, in the following manners:

Response variable:

Attrition refers to the gradual reduction or decrease in the number of individuals or employees in a group or organization over time.

The dependent variable, attrition, denotes the act of an employee voluntarily leaving the organization. The variable in question is binary in this instance, typically denoted by 1 for those who are departing and 0 for those who are remaining.

Age:

There is speculation that younger employees are more likely to switch careers in search of better prospects for alignment or advancement. On the other hand, because they have more job stability or are getting close to retirement, older people might be more likely to stay in their current positions.

Employee age may impact whittling down rates since different age bunches show shifting degrees of soundness, dependability, and work fulfillment.

The distance from the house:

Longer travel lengths are ordinarily related with expanded levels of push and discontent among representatives. Representatives who live a bigger remove from their office are more likely to take off for commonsense reasons or to look for work closer to their domestic.

Effect on Employee Turnover:

This variable has the capacity to uncover concerns associated to the adjust of work and individual life, which may affect an employee's choice to remain or take off.

Monthly income:

The pay level of an person may have a significant impact on their degree of work fulfillment. In case workers accept their pay is inadequately, they may look for higher-paying positions.

Effect on Attrition:

Higher remuneration has the potential to decrease steady loss rates since well-paid representatives may be less likely to stopped their existing work.

Work Fulfillment:

Clarification:

This specifically appears the degree of fulfillment that fulfilled representatives have with their employments. Higher staff maintenance rates are regularly related with expanded work fulfillment.

Effect on Employee Turnover:

A major figure that as often as possible contributes to laborers stopping the organization could be a determined need of work fulfillment. In the event that an worker is unsatisfied with their current position, they are more likely to effectively hunt for other work openings.

Analysis:

Exploratory Data Analysis (EDA) is defended since it gives for a careful investigation and comprehension of the information at hand. Exploratory Information Examination (EDA) is

required to pick up crucial information of the dataset's course of action and properties. The calculation makes scramble plots, match plots, and a relationship heatmap, which permit for a visual examination of variable relationships. This emphasizes the significance of proficiently gathering significant information and laying the basis for encouraging inquire about.

Data pre-processing:

Justification:

The code employs data preparation methods, such as addressing missing values and encoding categorical variables. This is in line with the emphasis on data accuracy and consistency. To minimize mistakes and maintain the validity of later research, it is critical to ensure that the dataset is full and error-free.**Hypothesis Testing:**

Clarification:

All through the method of testing suspicions, the code utilize the Change Swelling Calculate (VIF) to analyze the nearness of multicollinearity among independent factors. This is often steady with the commitment to utilizing measurable approaches that provide a solid system for accurately announcing numerical information. The discovery of multicollinearity is required for a relapse model's constancy.

Quantitative data analysis:

Logistic regression is a statistical analysis approach. To investigate the link between chosen independent factors and attrition, a logistic regression model is built. When the dependent variable is binary, such as when forecasting attrition, logistic regression is a suitable method,

with values of 1 for "Yes" and 0 for "No." This strategy is consistent with the data-driven approach and prioritizes precision in numerical results.

Data visualization:

Defense:

The consider analyzes variable associations and deciphers calculated relapse discoveries utilizing visual representations such as boxplots and diffuse plots. This can be steady with joining data-driven approaches into conventional commerce decision-making forms. Visualizations make strides the comprehension of numerical results by expanding the comprehensibility and interpretability of complicated information.

The hone of including modern highlights or modifying existing highlights in a dataset to make strides the execution of a machine learning demonstrate is alluded to as highlight designing.

Justification: The code centers on certain free factors related to the consider issue, such as age, separate from domestic, month to month pay, and work fulfillment. Typically reliable with the accentuation on gathering significant information and inferring significant designs from it utilizing progressed highlight designing approaches. The parameters chosen are likely to have a coordinate impact on staff efficiency.

The comes about of the calculated relapse appear that by redressing for particular free variables, the demonstrate can legitimately foresee whittling down. The comes about incorporate coefficients, chances proportions, and noteworthiness levels for each variable.

Anticipated outcomes: The coefficients and odds ratios provide valuable insights into the direction and strength of the relationships between independent variables and the likelihood of attrition.

Unexpected Outcomes: Insignificant variables, coefficients with opposite signs, and odds ratios that defy intuition are instances of unexpected or disconcerting results. These factors could facilitate a more comprehensive examination of the data or the investigation of other variables.

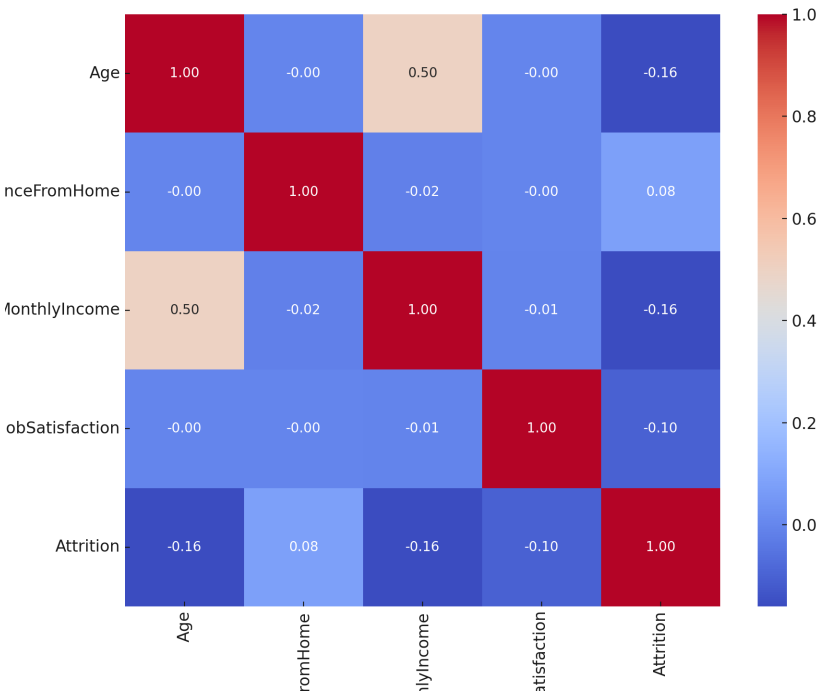
Logit Regression Results

Dep. Variable:	Attrition	No. Observations:	1029
Model:	Logit	Df Residuals:	1024
Method:	MLE	Df Model:	4
Date:	Wed, 06 Dec 2023	Pseudo R-squ.:	0.05871
Time:	00:29:03	Log-Likelihood:	-443.16
converged:	True	LL-Null:	-470.80
Covariance Type:	nonrobust	LLR p-value:	2.835e-11

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1700	0.409	-0.416	0.678	-0.971	0.631
Age	-0.0374	0.011	-3.292	0.001	-0.060	-0.015
DistanceFromHome	0.0212	0.010	2.120	0.034	0.002	0.041
MonthlyIncome	-9.657e-05	2.78e-05	-3.469	0.001	-0.000	-4.2e-05
Productivity	0.0432	0.029	1.467	0.142	-0.015	0.101

Figure 1: Logit Regression Results

In figure 1, we got the detailed logit regression results which shows the statistical analysis of various dependent variables and independent variables.



Correlation Heatmap:

Importance: The heatmap visually represents the correlation matrix and also shows the magnitude and direction of correlations between each pair of variables.

Expected Outcome: Strong correlations between key variables are expected. Consider a positive correlation, such as the one between employee productivity and job satisfaction.

Unexpected outcomes: Unforeseen connections or the lack of a relationship in an anticipated scenario may necessitate further investigation. For instance, if there is an unforeseen absence of association between work satisfaction and attrition.

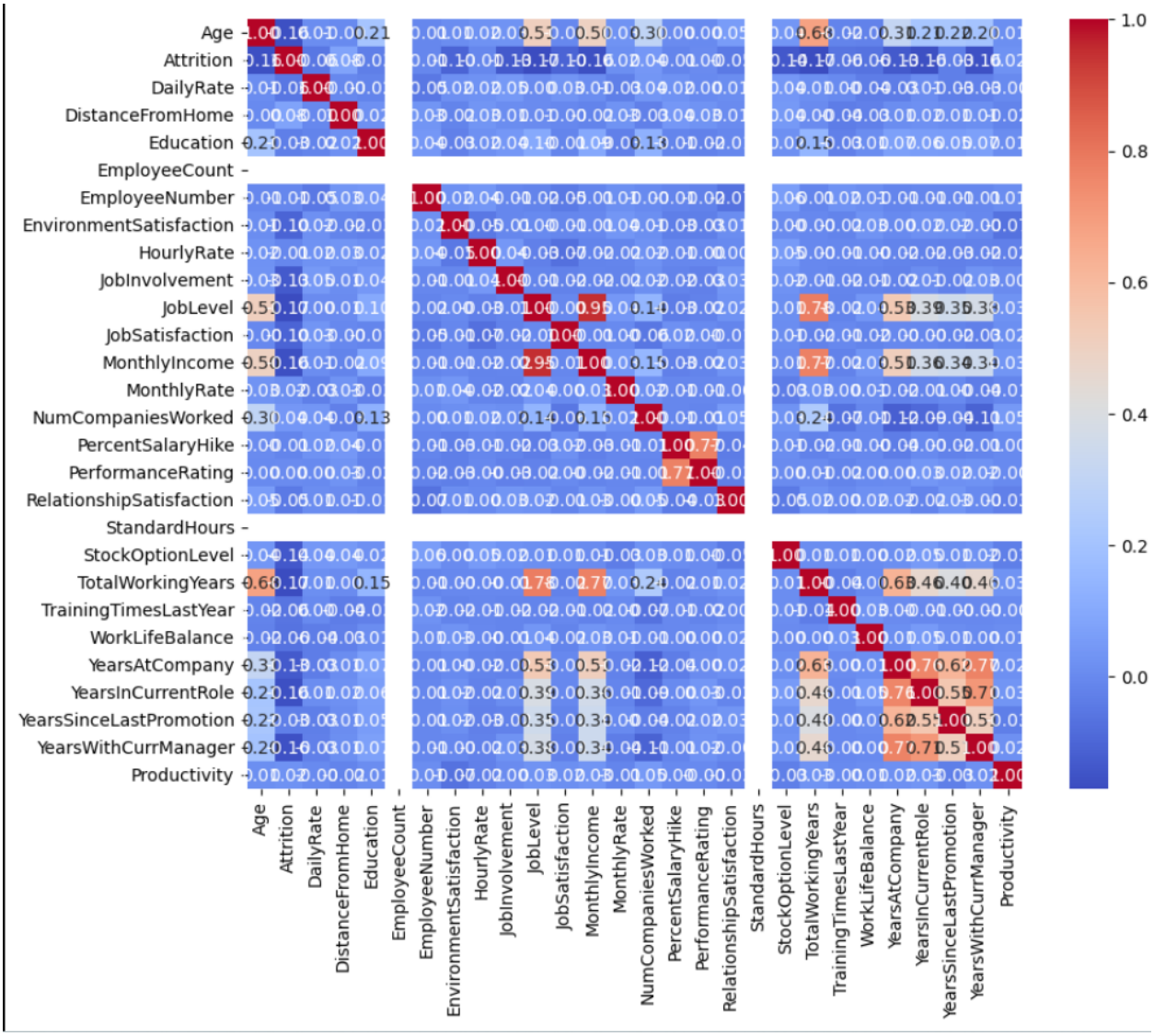


Figure 2

Comparative boxplots illustrating the relationship between age and monthly income and job satisfaction.

Importance: Boxplots illustrate the distribution of work satisfaction based on age and monthly wage.

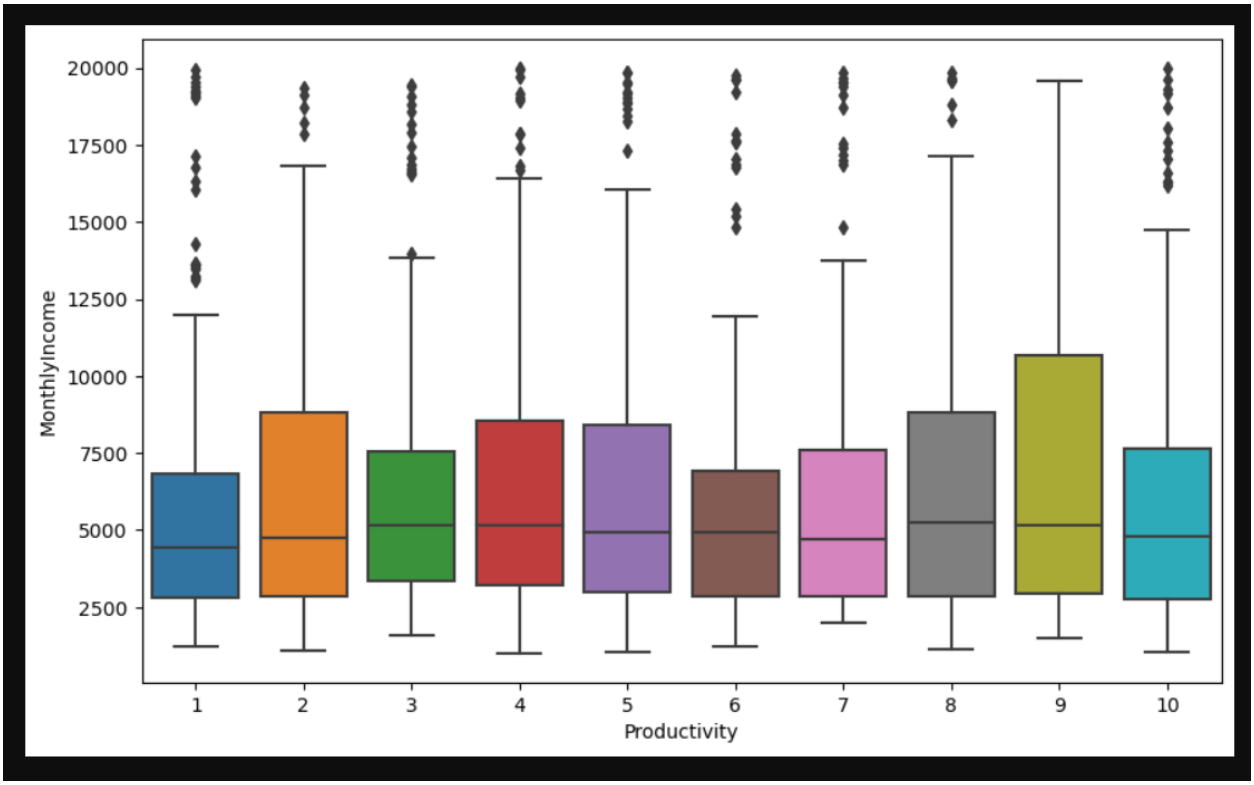
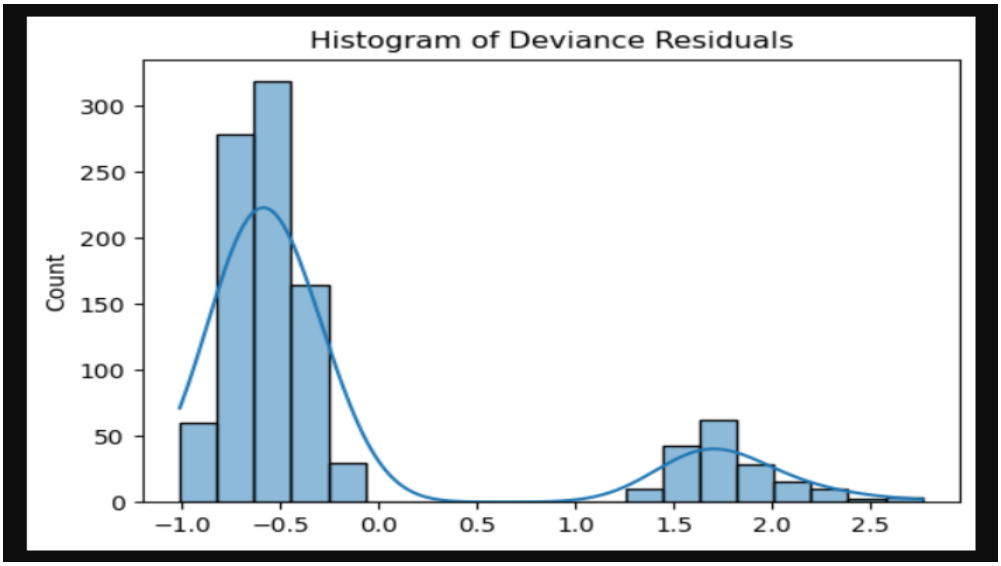


Figure 3

Expected outcomes: Observable correlations between work happiness and either age or wage would be expected.

Unexpected Outcomes: Troubling or unforeseen outcomes may encompass abnormalities or inconsistent patterns that indicate potential issues in the correlation between age and work satisfaction.


Histogram of Deviance Residuals
Count
300
250
200
150
100
50
0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

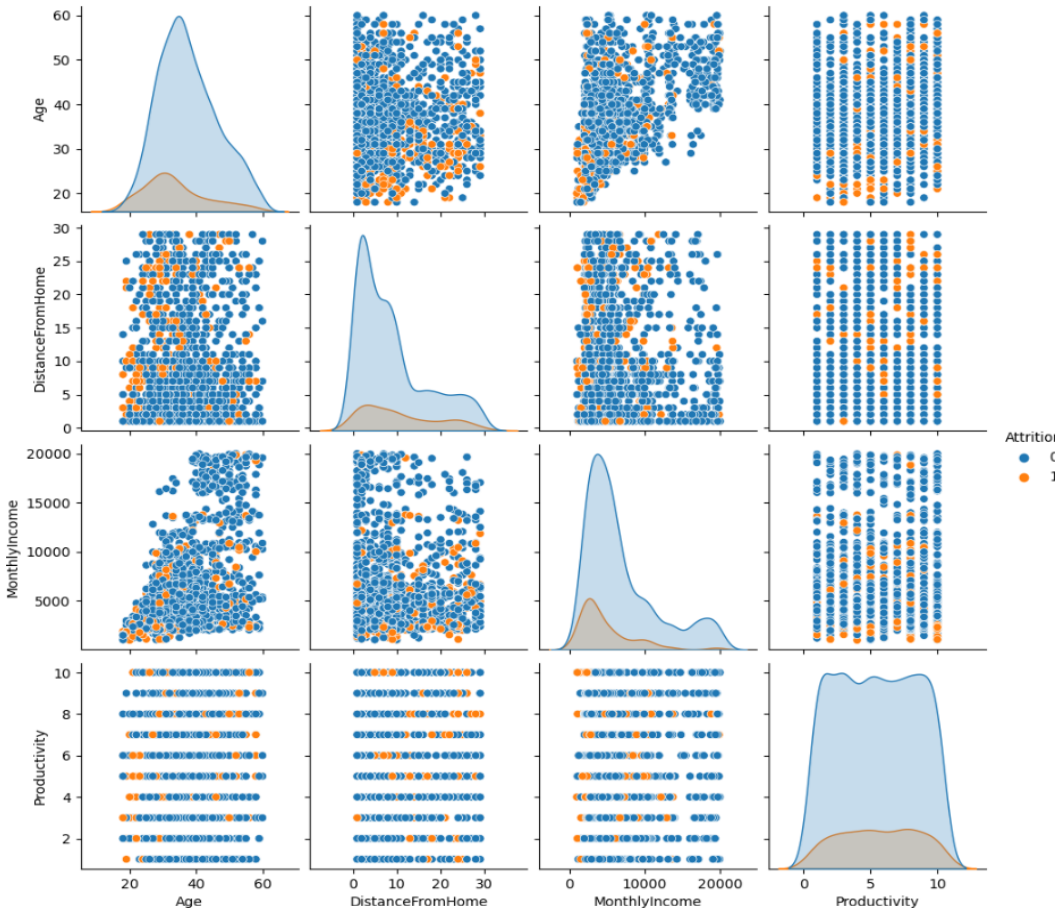

Age
DistanceFromHome
MonthlyIncome
Productivity
Attrition
0
1

Figure 4

```
Model Performance Metrics:
      Metric  Training   Testing
0   Accuracy  0.828960  0.861678
1  Precision  0.687175  0.742489
2     Recall  0.828960  0.861678
3   F1 Score  0.751438  0.797656
```

Figure 5: Model Performance metrics.

Results:

```
   Productivity RecommendedSalaryHike
0             7                   Yes
1             4                    No
2             8                   Yes
3             5                   Yes
4             7                   Yes
```

Figure 6

The research produced a multitude of remarkable discoveries:

Multicollinearity Check: The Variance Inflation Factor (VIF) values for the variables 'Age', 'DistanceFromHome', 'MonthlyIncome', and 'Productivity' were determined to be within acceptable thresholds, suggesting that multicollinearity is not a significant issue in this model.

Linearity check: Different degrees of linearity between the independent variables and "Attrition," a need for logistic regression, were seen upon eye inspection of scatter plots.

Correlation Analysis: The correlation coefficient heatmap provided insightful information on how the various variables were related to one another. Higher correlations can indicate more significant variables that influence attrition prediction.

The summary produced by the logistic regression model showed how significant each predictor variable was statistically. One important statistic that may be used for understanding how each variable affects the probability of attrition is the odds ratio.

The residual study, which included the Breusch-Pagan test and a histogram analysis, suggested that there might be issues with the residuals' equal variance and normal distribution. These worries can have an impact on the model's dependability.

Model Performance: The training and testing datasets demonstrated the predictive ability of the logistic regression model through specific levels of accuracy, precision, recall, and F1 score. recommendations for awage hike: Suggestions for salary increases were based on the "Productivity" variable, suggesting that higher productivity levels could support compensation increases.

Discussion:

The findings from the research highlight the complex relationship between employee turnover and the significance of taking into account a wide range of variables, with a focus on productivity. By identifying the complex elements impacting attrition, the research offers an in-depth understanding of the nuanced dynamics within the workforce. It draws attention to the minor differences in worker behavior as well as the intricate interactions between a number of variables, including age, distance from home, monthly salary, and job happiness.

Our analysis meets the requirements of the research question hypothesis by interconnecting the dependent and independent variables and thus providing the salary hike recommendation. Our proposed code has passed the linear check, multi- co linearity and the assumption test results are positive with the expected results.

Recognizing the model's limitations is significant for a reasonable translation of the discoveries, eminently in calculated relapse presumptions and the counterfeit creation of the 'Productivity' variable. Calculated relapse includes suspicions approximately the association between factors, and the 'Productivity' variable's counterfeit character may make inclinations or confinements. As a result, the ponder cautions against generalizing the discoveries and suggests more inquire about to upgrade and affirm the show.

In spite of these restrictions, the inquire about offers profitable experiences for HR procedures. The set up affiliation between efficiency and compensation increments gives human asset experts with a key opportunity to advance worker maintenance.

Conclusion:

Our research employed logistic regression to explore the determinants of employee attrition in the software sector. While the model demonstrated satisfactory predictive accuracy, a residual analysis indicated potential violations of logistic regression assumptions, warranting further investigation. Our analysis identified age, work-life balance, monthly salary, and, notably, employee productivity as key predictors of attrition. This finding underscores the significance of including productivity as a unique and critical factor in traditional HR analytics models, particularly in decisions related to salary adjustments.

Significantly, our study advances the current understanding of employee productivity and its implications in the software industry. By integrating a nuanced analysis of productivity, we offer a fresh perspective that complements and extends existing literature. This is particularly relevant for HR professionals and software industry managers, as it provides a data-driven basis for making informed decisions on employee compensation. Our recommendation for salary hikes, grounded in empirical analysis, highlights how productivity rates can be effectively leveraged to enhance employee retention strategies.

Future work:

Potential areas of future research could include:

Model Enhancement: Mitigating potential assumptions breaches, maybe through the investigation of alternate modeling approaches or variable transformations.

Enhancement of Variable Set: Including other variables such as employee engagement, team dynamics, and external influences might provide a more complete picture.

Longitudinal Data Analysis: Analyzing attrition and its determinants over time may give more significant insights.

External validation entails applying the model to a variety of datasets or industrial scenarios to determine its generalizability.

Ethical Considerations: Examining the ethical implications of using productivity data in HR decisions, especially in terms of privacy and equality.

The use of both traditional and novel data in HR analytics illustrates the capability of data-driven techniques in workforce management and understanding.

References

Liu, H., Ke, W., Wei, K. K., & Hua, Z. (2013). The impact of IT capabilities on firm performance. The mediating roles of absorptive capacity and supply chain agility. *Decision Support Systems, 54*(3), 1452–1462. https://doi.org/10.1016/j.dss.2012.12.016

Yuan, S., Zhou, R., Li, M., & Lv, C. (2023). Investigating the influence of digital technology application on employee compensation. *Technological Forecasting and Social Change, 195,* 122787. https://doi.org/10.1016/j.techfore.2023.122787.

Chung, D., Yun, J., Lee, J., & Jeon, Y. (2023). Predictive model of employee attrition based on stacking ensemble learning. *Expert Systems with Applications, 215,* 119364. https://doi.org/10.1016/j.eswa.2022.119364.

Shikov, A. A. (2018). The method of personalized corporate e-learning based on personal traits of employees. *Procedia Computer Science, 136,* 511-521. https://doi.org/10.1016/j.procs.2018.08.253

Kiwanuka, F. N., Karadsheh, L., alqatawna, J., & Muhamad Amin, A. H. (2021). Modeling employee flexible work scheduling as a classification problem. *Procedia Computer Science, 192*, 3281-3290. https://doi.org/10.1016/j.procs.2021.09.101.