Examining how classification models using Python in industrial/organizational psychology could be more accurate than a linear or logistic regression model in predicting employee productivity and success in the hiring process.

Beyza Ceylan

Abstract

Companies, industries, and places of business use artificial intelligence and statistics to predict the characteristics of their employees and staff. Data collected from these individuals is also used to make decisions about them regarding their work life, such as promotions, salaries, or within the hiring process. Two models that are commonly used throughout the field of psychology and specifically in industrial/organizational psychology are the linear regression and the logistic regression. Examining different classification models using Python shows the potential that there may be different models that are more accurate in their predictions of employee success, including a Random Forest model and a LightGBM model, which is short for gradient boosting model. The comparison of these models provides evidence suggesting that the Random Forest model and the LightGBM model predict employee productivity more accurately than a traditional linear regression model or a logistic regression model.

Keywords: Python, Random Forest, LightGBM, Linear Regression, Model, Prediction, Employee Success

Introduction

Artificial intelligence and statistical models are used every day to improve the workplace and to improve the functions of institutions. In the field of industrial/organizational psychology, a significant amount of data that helps improve the workplace and the employee's overall workplace experience comes from data on the employees at a company. In order to ensure that companies are getting good employees to help run their company smoothly, companies are becoming more interested in the use of artificial intelligence in the hiring process. Landers et al. (2023) explain that machine learning and computer algorithms are being used more and more to make predictions from new data and artificial intelligence is being used to make predictions about job applicants in addition to being used to help make hiring decisions. Gonzalez et al. (2019) add that although artificial intelligence and machine learning are becoming more common in different organizations that wish to hire high-quality employees, the amount of professional involvement and evaluation of applications of I-O psychologists in this new methodology is still limited in the selection process.

In the present research project, the research question I examined is whether a classification model is more accurate than a classic logistic regression and linear regression model applied in the field of industrial/organizational psychology regarding worker success. The goal and objective of this project were to write two classification models using Python and then compare them with a linear regression and logistic

regression model in terms of the accuracy of the results from the models. The results of this research project could aid and improve companies in how they select individuals in the hiring process.

This research project hypothesizes that the two classification models in Python will outperform the logistic and linear regression in the accuracy of their predictions. All four models will predict the productivity of employees using the number of projects they completed in the past year. The models will each predict the number of projects each employee will complete and then determine whether or not they are classified as productive depending on the number of projects. The motivation for this work is my interest in the field of industrial/organizational psychology and improving the selection process of employees. In industrial/organizational psychology, I/O psychologists currently use traditional linear regression and logistic regression models to predict employee success and productivity. In the selection process, a top-down process is used for employee selection, which essentially picks the highest predicted success employees to be hired. The benefits of this project are that the field of industrial/organizational psychology could have a more accurate and reliable model to replace the logistic regression and linear regression models it currently uses in predicting productivity, employee success, and more.

Methods

Participants

A sample of employees was used from the Kaggle website. Due to the goal of comparing the accuracies of both models, a large data set of approximately 15,000 people was used. This data set was treated as a "made up" data set since the validity and reliability of the data set could not be confirmed. Having said this, this did not harm or hinder the project in any way since the ultimate goal was to compare the accuracy of the two methodologies.

Procedure

Two classification models were completed using the Python language to predict the productivity of employees using the number of projects that they have completed. The data set was taken from the Kaggle website and contains data on 15,000 employees. There were 10 (x) variables or features in this data set that described the employees. This data set was collected from a survey that the employees in this business answered in the form of questions. These included satisfaction level, last evaluation, the number of projects completed, the average monthly hours the employee worked, the time spent at the company in years, the number of work accidents, whether or not they have been promoted or not in the past 5 years, the employee's salary, sales, and whether the employee has left the company or not. The x variable satisfaction level asked employees to rate themselves from a 0 to 1 in

terms of satisfaction at work, the x variable last evaluation was a score between 1 and 0 for that employee, the x variable number of projects was the number of projects an employee completed in a year, the x variable the average monthly hours the employee worked was the number of hours each employee worked in a month, the x variable time spent at the company recorded how long an employee had stayed at that given company measured in years, the x variable number of work accidents recorded the number of work accidents they had in a given year, the x variable of promotion last 5 years recorded a 1 or a 0 to indicate whether or not the employee had received a promotion in the last 5 years, the x variable employee salary was labeled as low, medium or high, the x variable sales described the occupation that each employee was in, such as located in sales or accounting, and the x variable whether the employee has left the company or not gives a 1 or a 0. All x variables were kept in order to have more data but the x variable number of projects became the y variable. These independent variables or features were used to predict the productivity of an employee. The models used two layers to predict information about each employee. The first prediction they made for each employee was the number of projects it believed that particular employee would complete within one year. The project range was from 1-7 projects in a year, however, no employee in the dataset completed 1 project per year therefore the range was 2-7 projects per year. The second layer was a dichotomous classification where the model classified an employee as being productive or not productive based on its prediction of the number of projects completed. The cutoff was found by finding the mean number of projects in the data set, which was 4 projects. The data set was a normal curve so

the mean was determined to be the best measure of the average. Any employee who completed 4 projects and above in the actual data set was considered to be productive and was considered to have performed at an effective performance level. Any employee in the actual data set who completed below 4 projects in a year was considered as not productive.

The first classification model built in Python was the Random Forest model. (See Figure 1) This model uses decision trees to make a prediction about a given employee. Each employee's information was given to the model and the Random Forest model predicted the number of projects that the employee will complete and also classified them as either productive or not productive based on its prediction. 70 percent of the actual data was used to train and familiarize the Random Forest model with the actual data. Once the model had been trained, the other 30 percent of the data was used to act as a testing data set. Since the number of projects was already available to the researcher in the 30 percent of the actual data set, the accuracy could be determined based on the predictions of the model itself. During training, the Random Forest model uses a technique known as bootstrapping to split the data into sections and uses those subsets of data to make predictions about the employees in them. Bootstrapping ensures randomness because it ensures that not every tree is using the same data which helps the model be less sensitive to the original training data. Random Forest also uses sampling with replacement while bootstrapping in order to ensure that each tree will have a different number of employees. This helps to manipulate the data without changing it. While doing the process of bootstrapping, Random Forest also uses random feature selection which increases

the randomness and protects the model against errors from other trees. Since each tree looks different due to the randomness of the sub-data sets, the trees do not act similarly so the model produces different predictions which helps to increase the variance. The number of decision trees used in this model was 1000. This number was determined by the model itself to give the most accurate results. After it splits the data into smaller subsets, the model asks questions throughout each tree and continues to do this until there are no further questions to ask and only one answer is possible, which is referred to as reaching purity. After the training data is complete, the model makes predictions for each employee using the test data where all the trees make a decision and the Random Forest model uses a majority vote to determine what the prediction is. The Random Forest model can also be re-run during its training to produce different trees every time. Once the training was complete, the test data was saved and the model was ready for any new information from new employees that could be placed in the future.

The second classification model used and compared with the linear regression and the logistic regression was LightGBM. (See Figure 2) This classification model also uses decision trees to make its predictions about the employees like the Random Forest model, however, unlike the Random Forest Model, each tree learns from the other's predictions and decisions. More specifically, each tree learns from the mistakes of the previous tree which helps it improve its prediction. Unlike the Random Forest model, there is not a majority vote for the prediction where all the trees predict simultaneously and independently, but the trees learn from one another. Exactly like the Random Forest model, 70 percent of the data was used as training

data for the model and 30 percent was used as testing data to determine the accuracy of the model.

A linear regression and a logistic regression were also modeled in Python to be able to compare the accuracies between these 4 models. The linear regression was also done using SPSS and the results of that linear regression were also compared to the 4 models made in Python. The predictions for all of the models were rounded up to the nearest whole number if it was 0.5 and above or rounded down to the nearest whole number if it was below 0.5. This rounding was kept consistent for all the models to ensure a fair comparison and the predictions were turned into integers to match the original data and make it easier to compare the models. The goals and objectives of the research question were met by comparing the different methodologies and testing for their accuracy. This work did build on existing research and efforts since the Python language and classification models are used a lot in data science, but are not as used in the field of industrial/organizational psychology since it relies more on regression models.

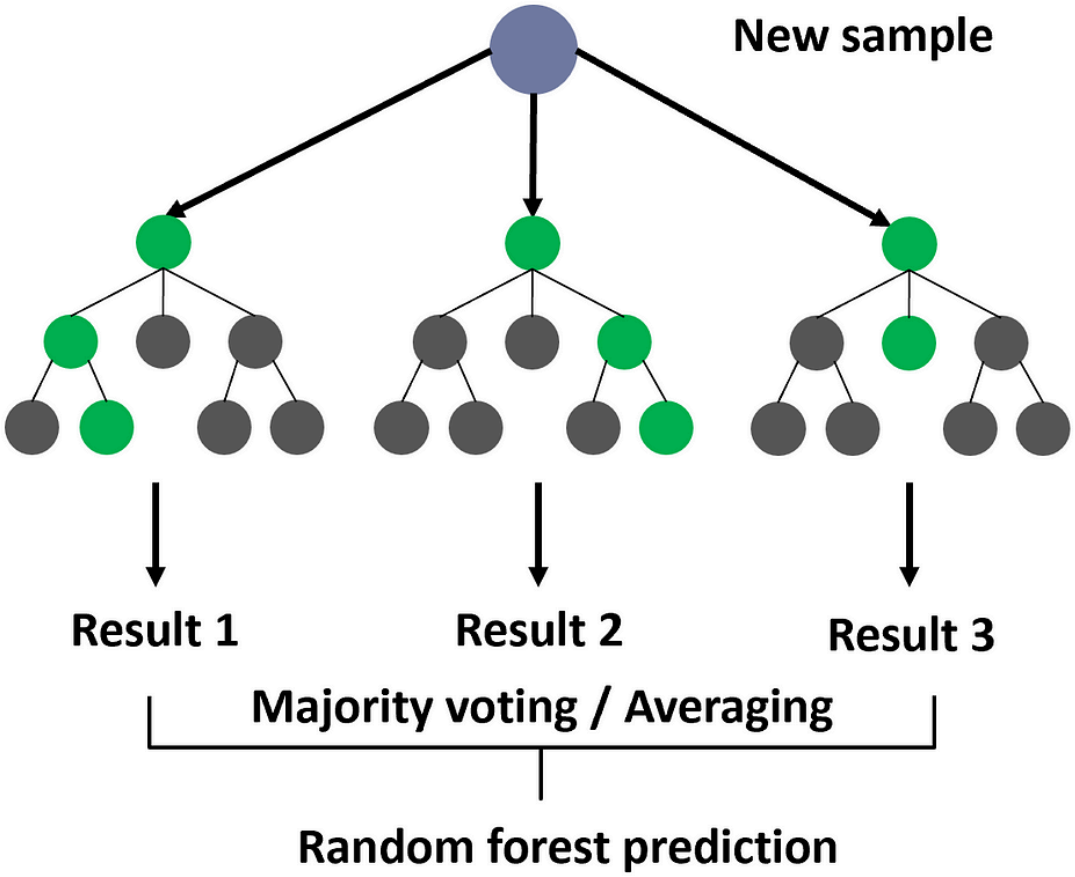

Figure 1. Random Forest Model Visual. This shows how the model receives a new sample, makes decision trees, each comes up with a result, and Random Forest takes the majority vote as its prediction.

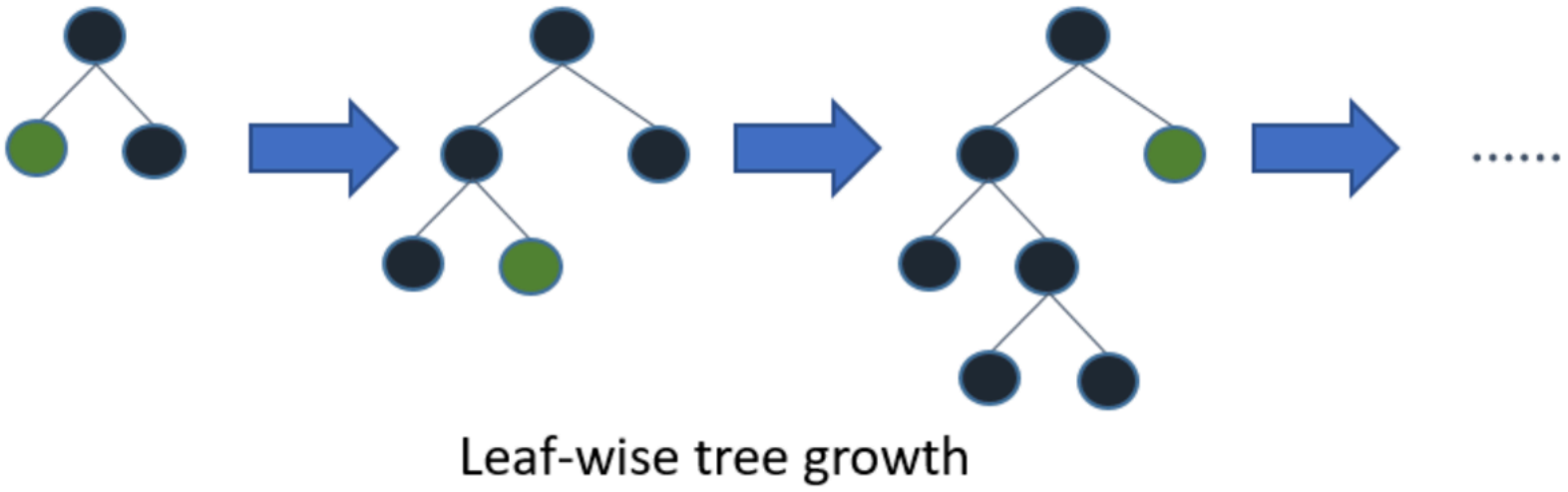

Figure 2. LightGBM Model Visual. This shows how each tree learns from the previous tree and grows from it to make its predictions.

Results

In order to compare the accuracies between the different models, the number of projects predicted correctly for each category was determined. These were displayed in the output of the Python model for each model including the linear regression model, the logistic regression model, the Random Forest model, and the LightGBM model. The categories were 2 projects, 3 projects, 4 projects, 5 projects, 6 projects, and 7 projects. There was no employee in the data set who completed 1

project in a year so that was not a category included. The number of predictions the models got correct for each category was put in the numerator and the denominator was the total number of employees who completed that category of projects in the actual data set. (See Figure 3) For the linear regression, the mean square error was also reported which was 1.09 along with the R_2 value which was 0.28. For all the models, the productivity accuracies were also measured. For every project that was 4 and above that the model had predicted correctly regardless of the number of projects, the model labeled that person as a 1 which meant they were productive. The numerator is the number of projects that are 4 and above predicted correctly with the condition that the model had to predict the number of projects correctly regardless if it was above or below 4. The denominator is the number of projects that are 4 and above for each category in the actual data. Lastly, the overall accuracy of the models was reported. The linear regression had an overall accuracy of 28 percent, the logistic regression had an overall accuracy of 45 percent, the Random Forest model had an overall accuracy of 56 percent, and the LightGMB model had an overall accuracy of 55 percent accuracy. The model with the highest accuracy was the Random Forest model. The model with the lowest accuracy was the linear regression model. In addition to this, the linear regression had an overall productivity classifying accuracy of about 36 percent, the logistic regression had an overall productivity classifying accuracy of about 37 percent, the Random Forest model had an overall productivity classifying accuracy of about 52 percent, and the LightGMB model had an overall productivity classifying accuracy of about 50 percent accuracy.

The model with the highest productivity classifying accuracy was the Random Forest model. The model with the lowest overall productivity classifying accuracy was the linear regression model.

	MODELS						
	Linear regression		Logistic regression		Random forest classifier		LightLGBM (Gredient Boosting Machines)
Number of projects that predicted correctly. Also it shows the accuracy of each number of projects.	Number of projects: 2 Accuracy: 1/716 Number of projects: 3 Accuracy: 428/1217 Number of projects: 4 Accuracy: 826/1310 Number of projects: 5 Accuracy: 98/828 Number of projects: 6 Accuracy: 0/352 Number of projects: 7 Accuracy: 0/77		Number of projects: 2 Accuracy: 495/716 Number of projects: 3 Accuracy: 594/1217 Number of projects: 4 Accuracy: 578/1310 Number of projects: 5 Accuracy: 165/828 Number of projects: 6 Accuracy: 213/352 Number of projects: 7 Accuracy: 0/77		Number of projects: 2 Accuracy: 523/716 Number of projects: 3 Accuracy: 676/1217 Number of projects: 4 Accuracy: 758/1310 Number of projects: 5 Accuracy: 308/828 Number of projects: 6 Accuracy: 212/352 Number of projects: 7 Accuracy: 50/77		Number of projects: 2 Accuracy: 532/716 Number of projects: 3 Accuracy: 651/1217 Number of projects: 4 Accuracy: 709/1310 Number of projects: 5 Accuracy: 324/828 Number of projects: 6 Accuracy: 206/352 Number of projects: 7 Accuracy: 51/77
MSE	Mean squared error (MSE) : 1.09						
R2_Score	Coefficient of determination 1 is perfect prediction - R2_score : 0.28						
Productivity accuracy	Productivity accuracy ratio: 924 / 2567 Ratio: 0.3599532528243085		Productivity accuracy ratio: 956 / 2567 Ratio: 0.3724191663420335		Productivity accuracy ratio: 1328 / 2567 Ratio: 0.5173354109855863		Productivity accuracy ratio: 1290 / 2567 Ratio: 0.5025321386832878
Overall accuracy	0.28		0.45		0.56		0.55

Figure 3. Accuracy results for the linear regression model, the logistic regression model, the Random Forest model, and the LightGBM model.

In addition to this, as part of the results reported, a permutation importance was done for the Random Forest model and the LightGMB model. The permutation importance shows how each x-value or each feature impacts the y-value or the outcome variable, which is the number of projects. For the Random Forest model, the x-value that impacted the number of projects the most was satisfaction level and the x-value that impacted the number of projects the least was whether or not an employee had received a promotion in the last 5 years. (See Figure 4) For the LightGBM model, the x-value that impacted the number of projects the most was satisfaction level and the x-value that impacted the number of projects the least was whether or not an employee had received a promotion in the last 5 years, so both models were the same in this outcome. (See Figure 5)

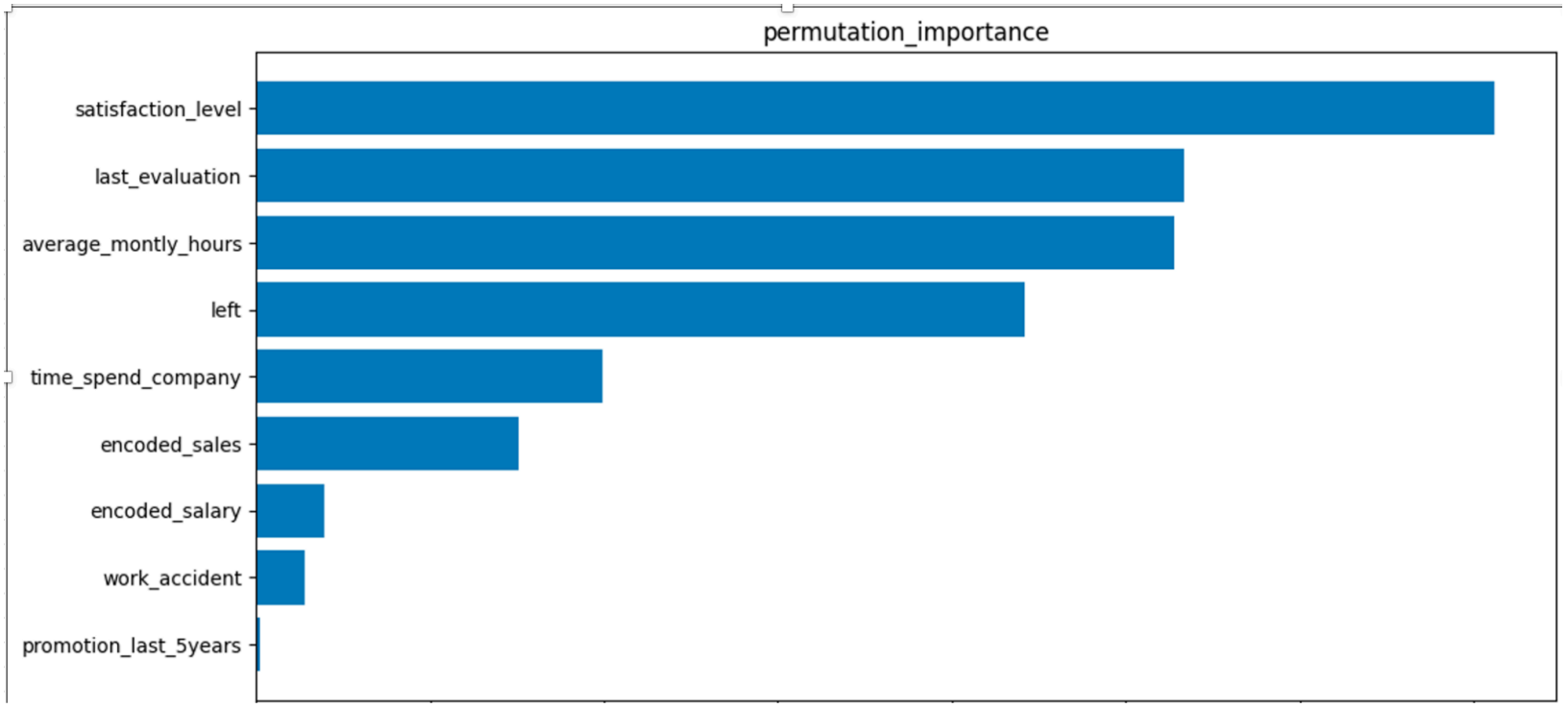

Figure 4. Permutation Importance For The Random Forest Model. Satisfaction level impacted the y-variable the most, promotion in the last 5 years impacted it the least.

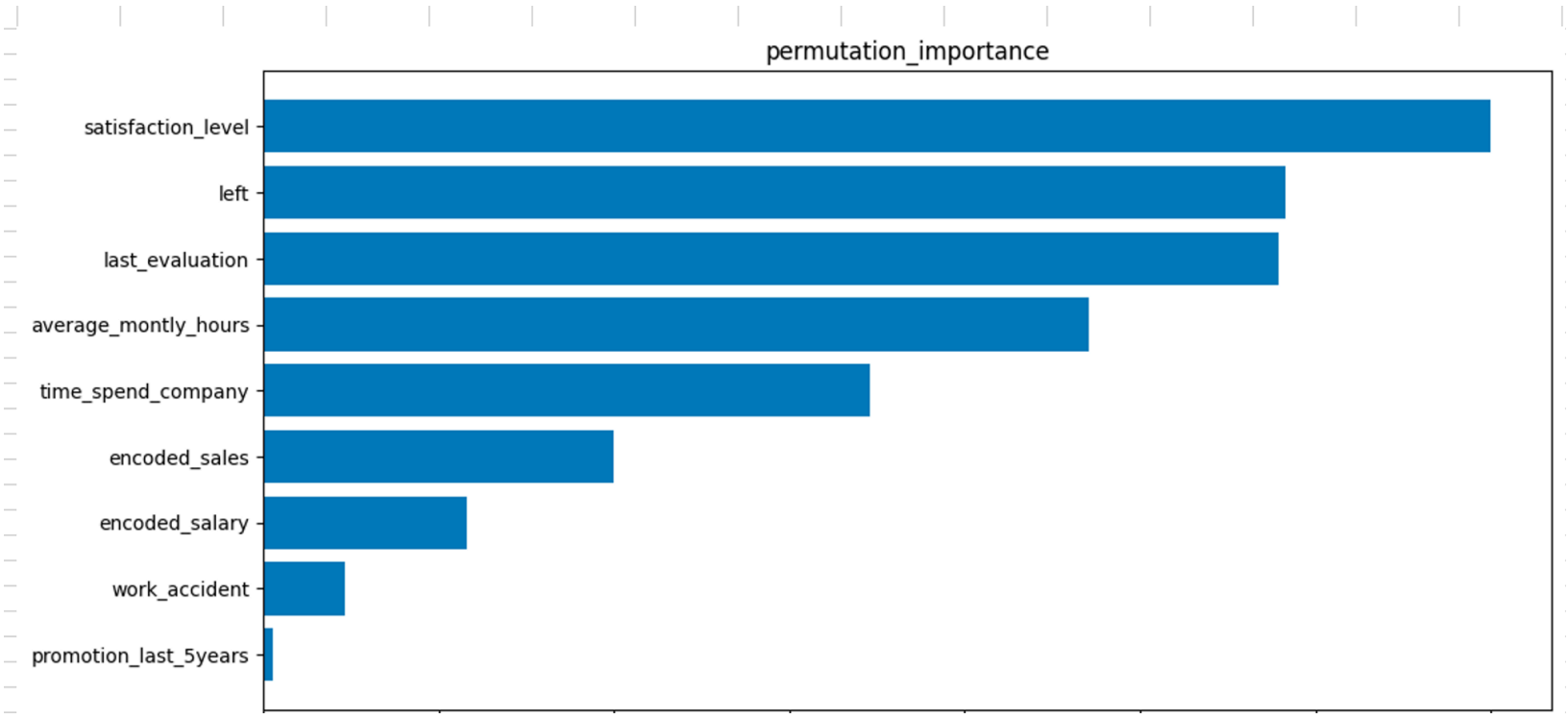

Figure 5. Permutation Importance For The LightGBM Model. Satisfaction level impacted the y-variable the most, promotion in the last 5 years impacted it the least.

Regarding the linear regression model run in SPSS, an Excel file was used to compare the accuracy of that model with the linear regression model done in Python and the other three models. (See Figure 6) First, in column A, the number of projects from the actual data was recorded. In column B, the predictions from the linear regression model in SPSS were recorded. In column C, the predictions from column B were rounded to whole numbers from decimals. In column D, the number of predictions that the linear regression model predicted correctly was recorded. If columns A and C matched up, column D labeled it as 1 which meant a correct prediction. If columns A and C did not match up, column D was labeled as 0 which meant an incorrect prediction. In column E, Excel divided the correctly predicted values in column C by the total number of rows or the total number of employees. Excel found the overall model accuracy for the linear regression done in SPSS to be about 30 percent. Next, in column F, Excel checked the productivity accuracy. If the model did not predict the number of projects for that employee correctly or if it was below 4 projects, then it labeled it as 0, which meant the employee was not productive. If the model did predict the number of projects correctly and the number of projects for that employee was above 4, Excel labeled it as 1 which meant that the employee was productive. Lastly, to measure the productivity accuracy, in column G, Excel divided the number of correctly predicted values in column F by the total number of projects that were 4 or

above in a year located in column A to find that the overall productivity classifier had an accuracy of about 38 percent.

	A	B	C	D	E	F	G
1	**num_projects**	**probs**	**rounded**	**correct_preds**	**Overall model accuracy**	**Productivity (model has to predict corrrectly and no of projects has to be 4 and above**	**Productivity accuracy**
2	5	3.38	3	0	0.302888889	0	0.38410596
3	5	3.71	4	0	Correctly Predicted Ones/ Total # Of Rows	0	Correctly Predicted ones/ Total # Of projects 4 and above
4	3	3.8	4	0		0	
5	5	4.26	4	0		0	
6	4	4.28	4	1		1	
7	3	3.85	4	0		0	
8	5	3.39	3	0		0	
9	4	4.35	4	1		1	
10	4	3.88	4	1		1	
11	5	3.59	4	0		0	
12	4	3.01	3	0		0	
13	3	3.1	3	1		0	
14	5	5.06	5	1		1	
15	4	3.76	4	1		1	
16	2	2.99	3	0		0	
17	4	3.43	3	0		0	
18	5	3.75	4	0		0	
19	4	3.48	3	0		0	
20	4	3.82	4	1		1	
21	5	4.59	5	1		1	
22	7	5.02	5	0		0	
23	3	3.65	4	0		0	
24	4	3.95	4	1		1	
25	4	3.91	4	1		1	
26	2	2.93	3	0		0	
27	3	3.49	3	1		0	
28	4	3.91	4	1		1	
29	4	4.37	4	1		1	
30	4	3.94	4	1		1	
31	7	5.38	5	0		0	
32	5	4.34	4	0		0	
33	4	4.01	4	1		1	
34	5	4.39	4	0		0	
35	5	4.08	4	0		0	
36	4	4.13	4	1		1	
37	5	3.84	4	0		0	
38	4	4.35	4	1		1	
39	3	3.87	4	0		0	
40	3	3.34	3	1		0	
41	3	4.32	4	0		0	
42	3	3.88	4	0		0	
43	3	3.5	4	0		0	
44	3	4.54	5	0		0	
45	4	3.19	3	0		0	

Figure 6. The Accuracy Of The Linear Regression Model Done In SPSS.

Discussion

Overall, the Random Forest model and the LightGBM model performed better than the linear regression model and the logistic regression model. There was a considerable difference in the accuracy and predictions of the linear regression model and logistic regression model compared to the Random Forest model and the LightGBM model. This also brings up the issue of ethics used in industrial/organizational psychology. Since the two classification models in Python performed better and predicted employee success more accurately, it begs the question of whether the linear or logistic regression should always be used in employee selection or other parts of industrial/organizational psychology where predictions from a linear or logistic regression would normally be used. When deciding whether or not an individual gets hired, when deciding if an employee gets promoted, when predicting the salary of a new employee, or other predictions regarding employees and human behavior, it is crucial to evaluate and predict information about individuals and employees as accurately as possible. The results of this research project strongly suggest the possibility that moving forward in the future of industrial/organizational psychology, I/O psychologists need to consider this difference in accuracy while being a consultant for companies knowing that the linear regression came out to around a 28 percent accuracy, which was the worst predictor,

and the Random Forest model came out to having around a 56 percent accuracy, which was the best predictor. Landers et al. (2023) conclude that professionals must take a wider and newer perspective on machine learning and artificial intelligence and must consider what is now possible that has never been possible before in order to understand the new kinds of predictors and predictions that can now be done with new technology. Gonzalez et al. (2019) also concludes that trained I-O psychologists need to be much more involved in the development and assessment of machine learning, artificial intelligence models, and technology while also highly recommending that artificial intelligence and machine learning corporations involve industrial/organizational psychologists who develop and implement their employment-based technologies for more future success.

References

Landers, R. N., Auer, E. M., Dunk, L., Langer, M., & Tran, K. N. (2023). A simulation of the impacts of machine learning to combine psychometric employee selection system predictors on performance prediction, adverse impact, and number of dropped predictors. *Personnel Psychology*, 1–24. Advance online publication. https://doi.org/10.1111/peps.12587

Gonzalez, Manuel F.; Capman, John F.; Oswald, Frederick L.; Theys, Evan R.; and Tomczak, David L. (2019) ""Where's the I-O?" Artificial Intelligence and Machine Learning in Talent Management Systems," Personnel Assessment and Decisions: Number 5 : Iss. 3 , Article 5. DOI: https://doi.org/10.25035/pad.2019.03.005 Available at: https://scholarworks.bgsu.edu/pad/vol5/iss3/5