Using Machine Learning to Predict Employee Turnover

Assignment 3: CSCI 3141: Foundations of Data Science Using R

Jamshid Zar

2022-11-27

Contents

1 Background **2**

2 Objective and Problem Definition **2**

3 Summary of Results **2**

4 Data **2**
4.1 Variables Description . . . 2
4.2 Exploratory Data Analysis . . . 3
4.2.1 Missing Values . . . 3
4.2.2 Duplicate Observations . . . 3
4.2.3 Correlation Analysis . . . 3
4.2.4 Distribution of the Target Variable, Left . . . 3
4.2.5 Turnover by Department . . . 3
4.2.6 Turnover by Salary Bands . . . 4
4.2.7 Turnover and Promotion in the Last 24 Months . . . 4
4.2.8 Summary Statistics for Numeric Variables . . . 6
4.2.9 Summary Statistics for Non-numeric Variables . . . 6

5 Choice of Variables **6**

6 Building the Models **6**
6.1 Experimental Setup and Model Evaluation . . . 6
6.2 Model Evaluation . . . 6
6.3 Training and Testing Sets . . . 7
6.4 Logistic Regression Model . . . 7
6.5 Decision Tree Model . . . 7
6.5.1 Model Evaluation . . . 8
6.6 Random Forest Model . . . 10

7 Conclusion **10**

8 References **10**

9 Appendix **11**

1 Background

In this analysis, we use data regarding employee turnover in an organisation to build machine learning models that will help the management to predict employee churn. Employee turnover is a major cost for organisation - moreso the replacement and training new employees.

As Skelton, Nattress, and Dwyer (2019) aptly note;

> Employee turnover expenses can cost businesses more than 100 per cent of a single employee's annual wages and negatively affection an organization's production and profits. High employee turnover also could affect community tax collections, social programs and physical and mental health issues. Therefore, understanding contributors to higher employee turnover remains essential for organizational managers from both a corporate and societal standpoint (p.1).

This project aims to predict employee turnover using machine learning. If managers could reasonably pinpoint the employees that are likely to leave, then they could initiate mitigation measures, saving the organisation valuable financial resources.

2 Objective and Problem Definition

The objective of this project is to build a machine learning model that predicts the probability of an employee leaving the organisation.

3 Summary of Results

4 Data

In this section, I explore the data starting with a description of the variables and exploratory data analysis (EDA) - data visualization and computing summary statistics.

4.1 Variables Description

I begin by loading in the data.

The data `employees` has 9540 observations of 10 variables. The data consists of the following variables.

Table 1: Table 1: Variables Description

Variable	Description
department	The department the employee belongs.
promoted	1 if the employee was promoted in the previous 24 months, 0 otherwise.
review	The composite score the employee received in their last evaluation.
projects	How many projects the employee is involved.
salary	For confidentiality reasons, salary comes in three tiers: low, medium, and high.
tenure	How many years the employee has been at the company.
satisfaction	A measure of employee satisfaction from surveys.
avg_hrs_month	The average hours the employee worked in a month.
left	Target variable: 'yes' if the employee ended up leaving, 'no' otherwise.

Note that the outcome of interest is `left`, indicating whether or not the employee left the organisation.

4.2 Exploratory Data Analysis

In exploring and cleaning the data, I examine the folloing matters.

- Missing values.
- Duplicate observations.
- Unusually high correlations between the independent variables.

4.2.1 Missing Values

As Figure () below shows, the data has no missing values.

Table 2: Missing values

variable	missing
department	0
promoted	0
review	0
projects	0
salary	0
tenure	0
satisfaction	0
bonus	0
avg_hrs_month	0
left	0

4.2.2 Duplicate Observations

Again the data set has no duplicate observations.

```
## # A tibble: 0 x 11
## # ... with 11 variables: department <chr>, promoted <fct>, review <dbl>,
## #   projects <dbl>, salary <chr>, tenure <dbl>, satisfaction <dbl>,
## #   bonus <dbl>, avg_hrs_month <dbl>, left <fct>, dupe_count <int>
```

4.2.3 Correlation Analysis

Figure () captures the correlation analysis showing an especially high correlation () between employee tenure and average hours worked per month. Given that this is a nearly perfect correlation, we drop one of these variables from the analysis.

4.2.4 Distribution of the Target Variable, Left

Figure () Panel A below shows the distribution of the target variable, `left`. The graph illustrates the lower prevalence of employees who left compared to those that remained with the organisation. In the modeling stage, this class imbalance can cause problems. To address this matter, we up-sample the data so that the proportion of employees who left is roughly equal to those that remained in the training set. We shall revisit this issue later in the modeling section.

4.2.5 Turnover by Department

Next, we examine the turnover by department. The analysis in Figure () Panel B below shows that staff in the IT and Logistics departments have a higher incidence of leaving the organisation staff in other departments.

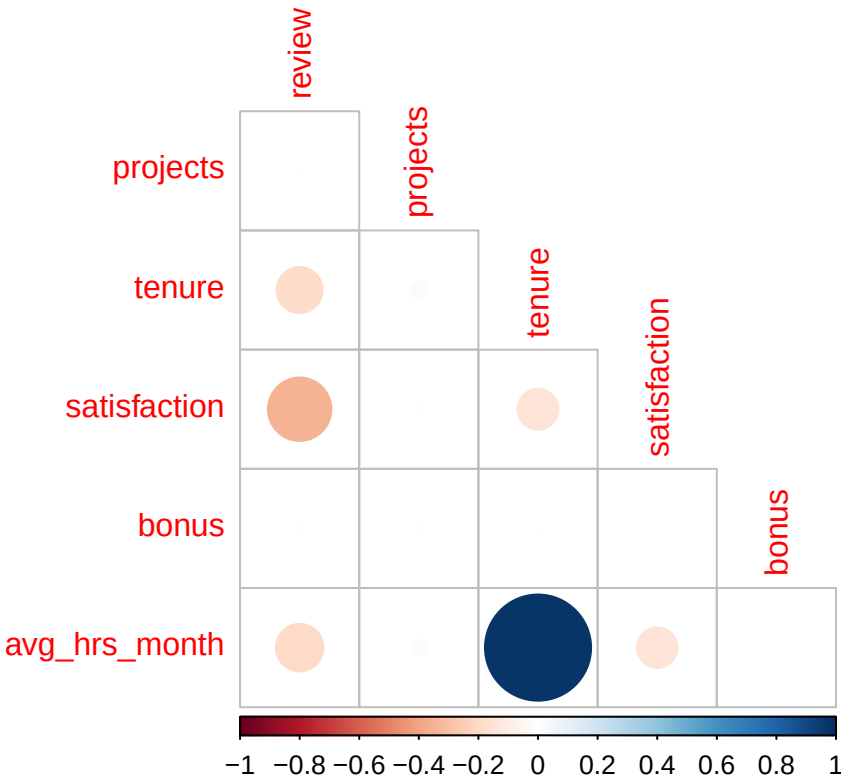


Figure 1: Correlation Matrix for Dependent Variables

4.2.6 Turnover by Salary Bands

Figure () Panel C below shows no notable differences in staff turnover by salary. While pay is an important element of an employee motivation to work, it does not appear to make a difference in the decision for the staff to stay with or to leave the organisation.

4.2.7 Turnover and Promotion in the Last 24 Months

Promotion seems to have a notable relationship with the likelihood of an employee leaving the organisation (see Figure () panel D).

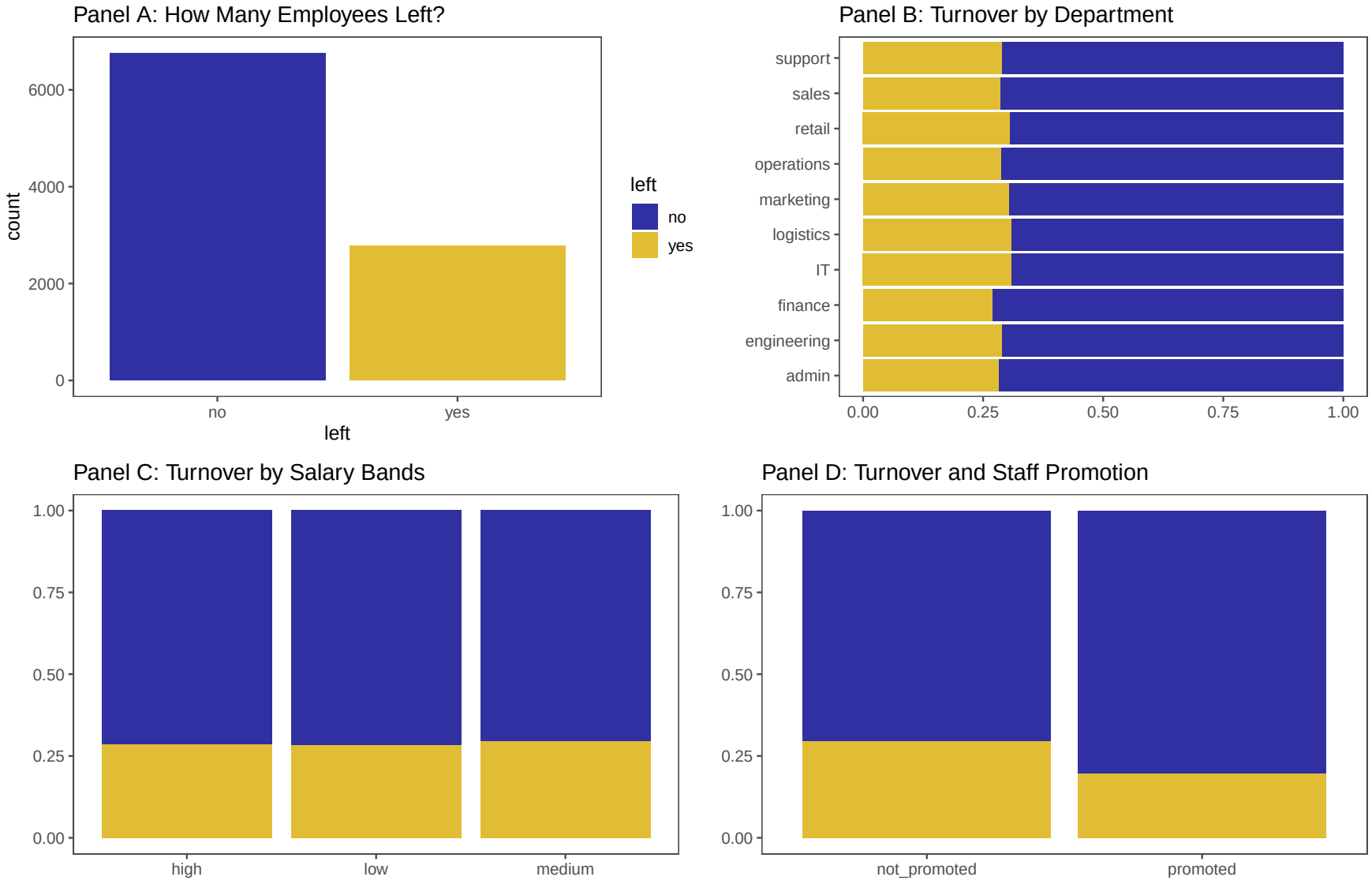


Figure 2: Target Variable Distribution

4.2.8 Summary Statistics for Numeric Variables

In this section, I compute summary statistics for the numeric variables followed by the non-numeric variables. The summaries complement Figure () in Appendix 1. Overall, the imbalance in the target variable `left` and the skewed distribution of some of the other variables in the dataset, especially `salary` and `bonus`. Te variable `bonus` also has substantial outliers meaning a few employees get very high bonuses relative to other employees.

Table 3: Summary Statistics

skim_variable	complete_rate	Mean	SD	Min	Q1	Median	Q3	Max
review	1	0.65	0.09	0.31	0.59	0.65	0.71	1
projects	1	3.27	0.58	2.00	3.00	3.00	4.00	5
tenure	1	6.56	1.42	2.00	5.00	7.00	8.00	12
satisfaction	1	0.50	0.16	0.00	0.39	0.50	0.62	1
bonus	1	0.21	0.41	0.00	0.00	0.00	0.00	1
avg_hrs_month	1	184.66	4.14	171.37	181.47	184.63	187.73	201

4.2.9 Summary Statistics for Non-numeric Variables

Table 4: Summary Statistics

skim_variable	complete_rate	character.min	character.max	character.n_unique	character.whitespace
department	1	2	11	10	0
salary	1	3	6	3	0

5 Choice of Variables

I drop the variable `avg_hrs_month` given that it has a very high correlation with `tenure`. However, I still retain `salary` and `bonus`. Although the two variables exhibit very little contribution to employees leaving the organisation, the little difference that exists could still contribute to the model performance.

6 Building the Models

6.1 Experimental Setup and Model Evaluation

For this analysis, I will run 3 predictive models.

1. Logistic Regression Model.
2. The Decision Tree Model.
3. The Random Forest Model.

In all cases, I tune the parameters and use a distinct training and testing set with cross validation in the training set.

6.2 Model Evaluation

I use `specificity`, the area under the curve (`roc-auc`), and the `balanced accuracy`. The justification for using specificity is that it is better to predict an employee will leave the organisation when, in reality, they do not leave. However, it is not as good for a model to predict that an employee will not leave the organisation

when they in reality do leave. In other words, the cost of not detecting employee churn is much higher. The Area Under Curve (AUC) supplements the specificity by evaluating how well the models discriminate the positive (those employees who leave) and negative (employees that remain with the organisation) at all thresholds (Narkhede 2018). Like the AUC, the balanced accuracy, the arithmetic mean of `sensitivity` and `specificity` evaluates how well the models discriminate between employees who leave versus the employees who stay (Gorzałczany and Rudziński 2016).

6.3 Training and Testing Sets

I choose an 80-20 training-testing set split. To split the data, I use the initial_split function from `tidymodels`, a streamlined set of packages for machine learning in R.

The training set has 7632 observations while the testing set has 1908. Henceforth, we only use the testing set for model evaluation. However, I implement cross validation using the training set in each of the models that follow.

Next, I up-sample the data to cater for the imbalance in the target variable, `left`. I also drop one of the independent variables with a correlation beyond an absolute value of 0.85.

I also set up folds for cross validation. We use these folds in all the models that follow.

6.4 Logistic Regression Model

6.5 Decision Tree Model

The decision tree model uses a tree like model of decisions and their consequences. Although it is simple to fit and interpret, it is prone to over fitting. In this model, we shall tune two parameters.

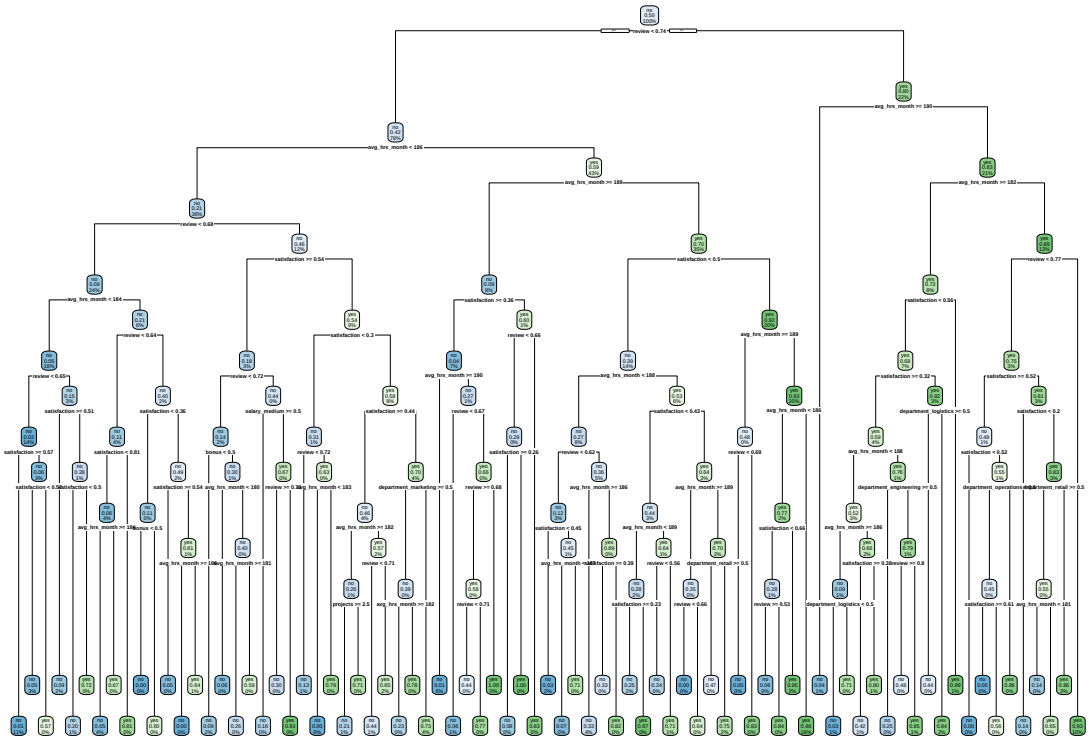
- Tree depth: How many splits a tree can make before coming to a prediction.
- Cost complexity: This parameter allows us to control the size of the tree and hence manage over-fitting by adding a penalty for every new branch created.

Next, we fit the model allowing for hyper parameter tuning;

I then set up a workflow

Next, I generate a grid of hyper-parameters.

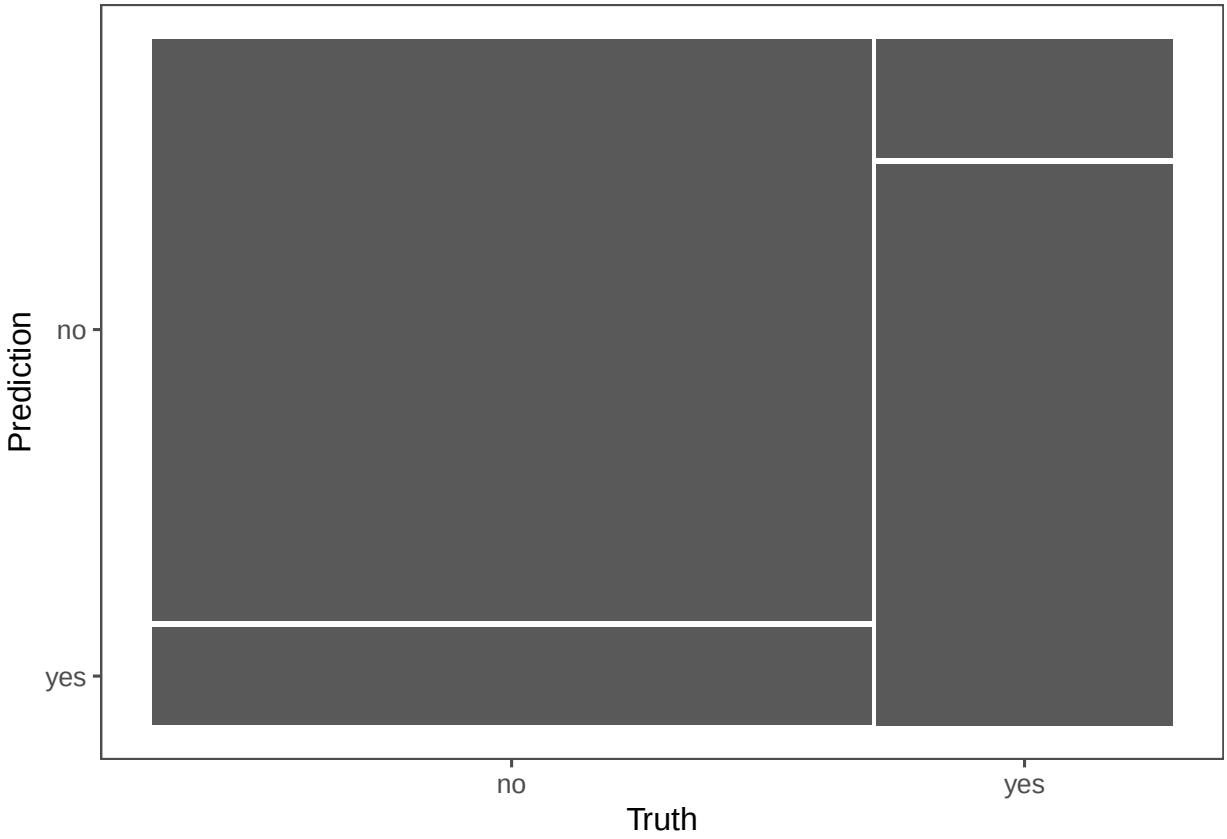
I use the workflow to run models over different parameters and choose the best combination.



6.5.1 Model Evaluation

In this section, I evaluate the decision tree model on the test set. As noted earlier, we focus on the `specificity` and `AUC`.

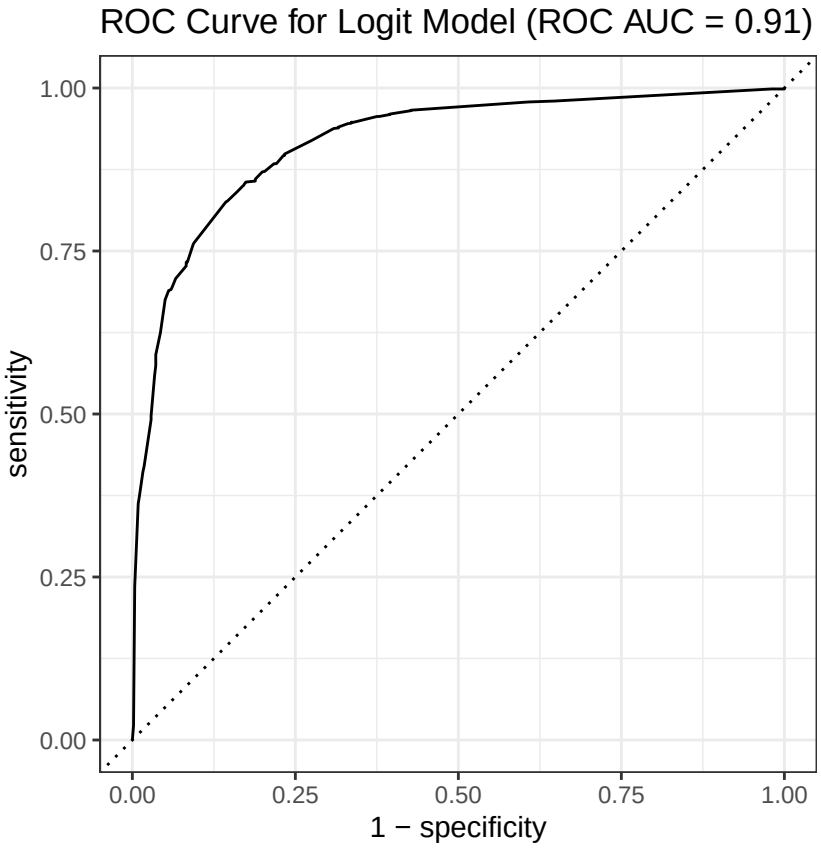
Next I compute the metrics.



The Table () summary of the metrics of the models. Note that the `specificity` and `sensitivity` and are almost equal at 0.83. Consequently the balanced accuracy is also 0.83.

```
## # A tibble: 13 x 3
##    .metric              .estimator .estimate
##    <chr>                <chr>          <dbl>
##  1 accuracy             binary         0.847
##  2 kap                  binary         0.648
##  3 sens                 binary         0.856
##  4 spec                 binary         0.826
##  5 ppv                  binary         0.923
##  6 npv                  binary         0.702
##  7 mcc                  binary         0.653
##  8 j_index              binary         0.682
##  9 bal_accuracy         binary         0.841
## 10 detection_prevalence binary         0.657
## 11 precision            binary         0.923
## 12 recall               binary         0.856
## 13 f_meas               binary         0.888
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.917
```



6.6 Random Forest Model

The Random Forest Model is an ensemble technique that works by creating many decision trees at run time. It is less prone to over-fitting and even performs well using the default metrics without tuning the parameters. Due to limits in computation power, it is hard for us to fine tune this model.

Ideally, we need to tune the following parameters;

- Mtry: This is the number of variables to sample randomly in each split.
- Min_n: The minimum number of data points required in a node for the node to be split further.

I then set up a workflow consisting of the model, the recipe.

7 Conclusion

8 References

Gorzałczany, Marian B, and Filip Rudziński. 2016. "A Multi-Objective Genetic Optimization for Fast, Fuzzy Rule-Based Credit Classification with Balanced Accuracy and Interpretability." *Applied Soft Computing* 40: 206–20.

Narkhede, Sarang. 2018. "Understanding Auc-Roc Curve." *Towards Data Science* 26 (1): 220–27.

Skelton, Angie R, Deborah Nattress, and Rocky J Dwyer. 2019. "Predicting Manufacturing Employee Turnover Intentions." *Journal of Economics, Finance and Administrative Science*.

9 Appendix

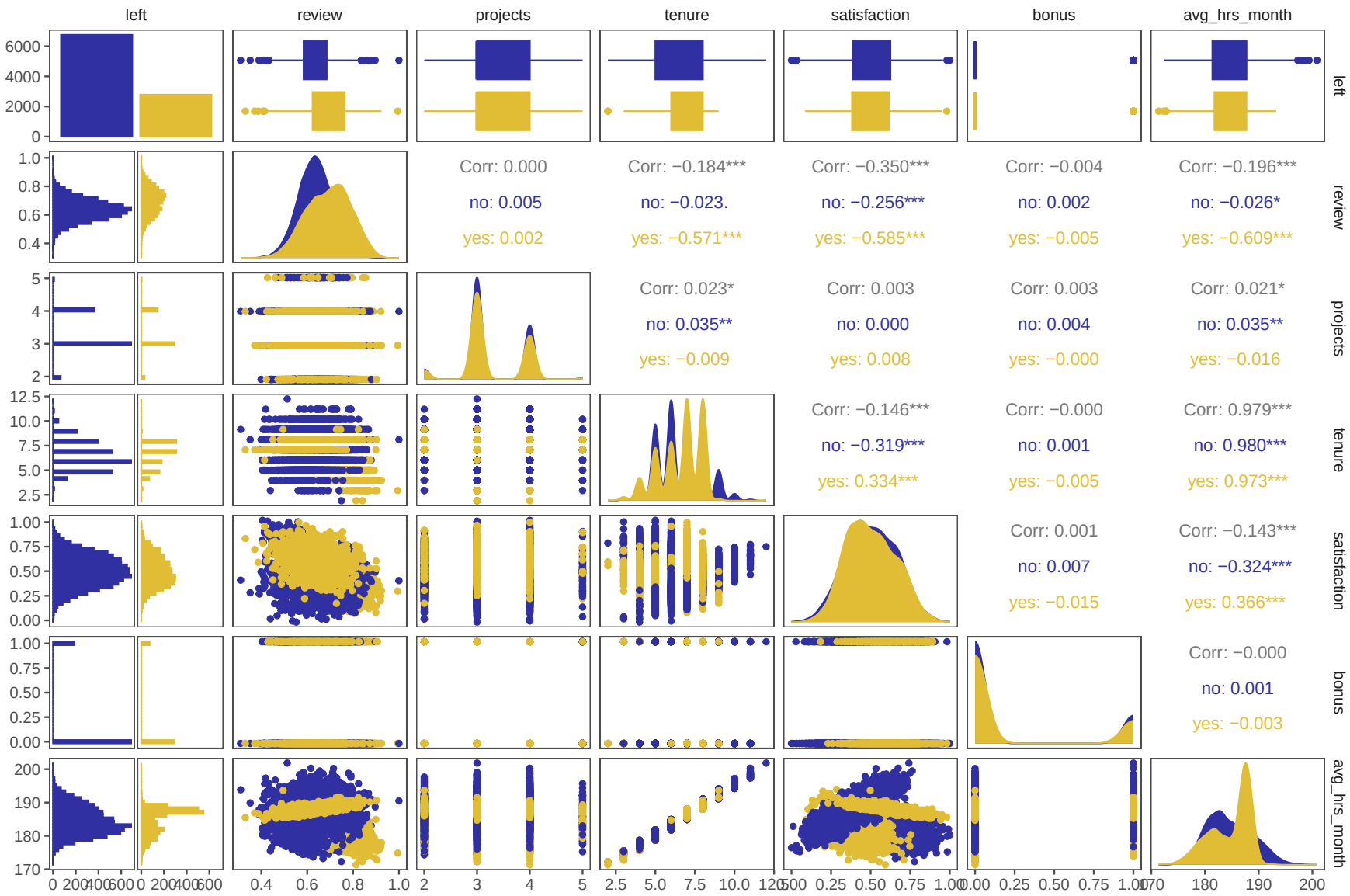


Figure 3: Visualising Numeric Variables Against the Target Variable, Left