HR Analytics

Based on the IBM HR Analytics Employee Attrition and Performance dataset retrieved from Kaggle.

Employees are arguably a company's most valuable resource. A good employee leaving can be a huge loss for the company. In his book on Predictive Analytics, Eric Siegel (2016) discusses a flight risk model developed by HP to predict how likely an employee is to quit. The model assigns employees a flight risk score. In this case study, your team will attempt something similar.

You have access to a fictional dataset created by IBM that contains information (35 attributes) on 1470 employees including whether they left their position (Attrition = Yes if they left, 0 otherwise). Most of the attributes (columns) in the data set are self-explanatory. Some columns represent categorical variables that have been numerically coded. The excel file data_dictionary.xlsx contains additional information on these variables. The data is in the file, EmployeeAttrition_data.xlsx. Analyze this data to develop an employee attrition prediction model.

Descriptive Statistics

1. Start your analysis by describing the data (generate descriptive statistics for the variables in the dataset).

Data Cleaning and Preparation

2. Some of the variables can be removed from the dataset because it is apparent that they have no predictive value. Identify these variables and remove them from the dataset. Provide a justification for removing each variable.
3. Some variables need to be transformed or imputed (or rows deleted) before they can be used in a regression model. Identify and transform these variables. Provide a summary of the identification and transformation strategy.

Model generation

Develop a model that predicts the probability that an employee will quit and classifies employees into two classes – those that will quit and those that will stay. Notice that you must not only choose the right variables but also the right cutoff level for classification.

1. Probability prediction – Start by identifying a parsimonious model that predicts the probability that an employee will quit. Explain the analyses used to select the variables in the final model. Use the entire dataset for this part of your analysis. Provide a summary of your results including insights on factors that affect attrition.

2. Classification – Use the model to classify employees using the predicted probabilities and a cutoff probability.

   Choose the cutoff level carefully and justify your choice. Include the *average* performance measures for your model (accuracy, misclassification rate, sensitivity, precision and specificity) on testing data in your report. Use 5-fold cross validation to measure the average out-of-sample performance of the model and cutoff value. Explain which performance measure(s) were used to select the cutoff level and why. Your analysis should explore at least two cutoff levels. (see details on measuring out-of-sample performance and selecting cutoff levels below).

3. Prediction and classification equations – Clearly identify the equation that can be used to predict attrition probabilities for new observations and the rule that should be used to classify the observations. The regression coefficients for prediction equation should come from step 1 of model generation and the cutoff level should come from the result of step 2.

Measuring out-of-sample performance and selecting cutoff levels

You will select the variables (inputs) for your model in step 1 of model generation using all available data. In step 2, you must choose an appropriate cutoff level. Generate appropriate performance measures for the model chosen in step 1 and compare these measures for several cutoff levels to choose a subset of candidate cutoff levels for further analysis. To choose a final cutoff level, you should look at the out-of-sample performance of the model for the candidate cutoff levels. To do this, you should use a 5-fold cross validation technique.

1. For 5-fold cross validation you should generate 5 subsets from the available data. Normally, you would randomly select the observations to include in each subset. However, in this case you have an additional concern. The dataset is not balanced on the outcome variable (i.e., the target and the nontarget classes are not equally represented). Verify this and report the imbalance in your write-up. Suppose target class cases make up p% of all observations in your data. If you randomly select observations for each subset, you can end up with subsets that don’t have any observations from the target class or where the target class is not adequately represented. Instead of simple random sampling, you should sample observations such that the percentage of target class cases in each subset to be the same as in the entire dataset. You should still select observations randomly but make sure you are sampling appropriately from each group or strata. This will allow you to perform a stratified 5-fold cross validation which is better suited for imbalanced data. Once you have generated 5 stratified samples, you can proceed to the next step.

2. For this step, you should first select two or more candidate cutoff levels using considerations mentioned before. Next, follow the 5-fold cross-validation procedure. In each fold, run (train) the model on a training set comprised of 4 of the subsets. Then, calculate and tabulate the appropriate performance measures for each candidate cutoff level using the remaining subset (testing data). After completing the 5 folds you should have a table similar to the one shown below (Please include this table in your report).

Fold	Cutoff level	Performance measure 1	Performance measure 2	…
1	<value 1>	<value>	<value>	
1	<value 2>	<value>	<value>	
1	…	…	…	
2	<value 1>	<value>	<value>	
2	<value 2>	<value>	<value>	
2	…	…	…	
3	<value 1>	<value>	<value>	
3	<value 2>	<value>	<value>	
3	…	…	…	
4	<value 1>	<value>	<value>	
4	<value 2>	<value>	<value>	
4	…	…	…	
5	<value 1>	<value>	<value>	
5	<value 2>	<value>	<value>	
5	…	…	…	

Use the above data to calculate the average performance of each cutoff level (across the 5 folds) on the relevant performance measure(s). Use the averages to select a final cutoff value.

Cutoff level	Performance measure 1	Performance measure 2	…
<value 1>	<Average value>	<Average value>	
<value 2>	<Average value>	<Average value>	
…	…	…	

References:

Siegel, Eric. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die, John Wiley & Sons, Incorporated, 2016.

Submission guidelines and Evaluation of case reports

Submission guidelines

Each case group should submit a case report and an excel workbook, along with any other material containing the detailed analysis.

1. The **case report** should include an executive summary of the analysis (1 -2 pages and should avoid excessive technical jargon), followed by a description of the analysis (2-4 pages) and an appendix. All tables and figures should be restricted to the appendix and appropriately referenced in the executive summary and analysis sections.

2. An **excel workbook** showing all your work.

Evaluation

The case submissions will be evaluated on the following criteria:

1. **Methodology and Analysis (60 points):** Your work will be evaluated on the quality of the analysis done. Criteria will include
   - Selection of appropriate performance metrics Appropriate application of techniques and principles to model development and evaluation (model and variable selection, model evaluation and comparison, etc.)
   - Correct execution of the above.
   - The extent to which case questions/requirements have been addressed.

   Breakdown of points based on case requirements -
   - Descriptive analytics (10 points)
   - Data cleaning and preparation (15 points)
   - Variable selection (10 points)
   - Cross-validation (10 points)
   - Selection of appropriate performance criteria and cutoff level (15 points)

2. **Writing and Insights (25 points):** A sound analysis is critical to answering questions presented in the case. Effective communication of the methods, results, insights and recommendations is equally important. The quality of your case report will be assessed on criteria such as clarity and precision in writing, clear articulation of the rationale for specific analyses and modeling choices, the degree to which conclusions are supported by the analysis.

3. **Peer evaluation (15 points):** After you turn in the case report, I will ask you to fill a peer evaluation form for the members of your case group. Your final score for the case will be based on the quality of the deliverable as well as the evaluation from your peers.

Tools and software

Please use Excel and RegressItLogistic for this case study.

Use of external resources

Graduate students sometimes encounter a need to refer to external sources (outside of assigned course material) for help with assignments. That is a part of the learning process and I encourage you to research any external sources that you might find useful. Be sure to cite the sources that you use. However, you should know that looking up solutions to the assigned problems on the internet and other sources is not appropriate. This case study uses a dataset from Kaggle and as such you can find webpages that provide partial analysis of this data. Using such webpages (and other resources) to inform your own analysis is a violation of the academic integrity policy. You can research concepts and techniques such as logistic regression and cross validation, but you may not look up information on the application that is the focus of this case study. Inappropriate use of external sources will result in a score of zero for the group. If in doubt as to the appropriateness of a resource, please check with me first.