*Forecasting Imminent Employee Attrition Using Machine Learning*

Allen Gustrowsky   
School of Computing, Southern Adventist University  
Collegedale, TN, USA  
agustrowsky@southern.edu

Caeden Scott  
School of Computing, Southern Adventist University  
Collegedalee, TN,   
cdscott@southern.edu

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (*Abstract*)

Keywords—component, formatting, style, styling, insert (key words)

# Description

The dataset for this project came from Kaggle. It was a fairly clean dataset in terms of missing and consistent values; null values were not an issue, and each row seemed to have data for all features. There were thirty-four features not including the class field. Basic data points such as age and gender were captured, but more nuanced features like “StockOptionLevel,” “YearsSinceLastPromotion,” and “WorkLifeBalance” were included. Fortunately, many of these fields were in number format already, but some had to be transformed. This brings us to the data engineering process.

Data engineering was the most time-consuming part of this project. A few different techniques were employed to prepare and optimize the data for the models. First, any textual features (such as “Department”) were transformed into numerical data with a label encoder. Since this textual data did not require sentiment analysis but were rather like categories, we though label encoders would work fine.

The second step was data resampling, which was perhaps the longest parts of the data engineering process. We tried a few different approaches. For the test data, we decided to use up-sampling. More than one algorithm was experimented with, and we chose the SMOTE algorithm by Imbalanced-Learn. This algorithm produces synthetic data from the minority set until the number of records in that minority set equal the number in the larger set. With testing data, we tried both up-sampling and down-sampling. The former seemed to yield better results, but it seemed like the test result would be less credible because we were introducing synthetic data that was not actually real. Because of this, we chose to use down-sampling of the testing data.

The third step was feature selection. We used the chi squared algorithm and selected the top 14 most useful features. Different approaches were researched, but ultimately this seemed like one of the best balance of results, time to implement, and technical compatibility. Lastly, we normalized the data.

The order was important for all of these steps because some steps are required for others to take place. For example, numerical data is needed for normalizing. Also, it is best to resample data and make the training set complete before selecting the most relevant features. Without resampling first, the dataset being used to select the best features would be biased.

# Results

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this file and download the file “MSW\_A4\_format”.

With the approach we chose, the final result with our best model was fairly good. Before resampling, the results yielded excellent outcomes for predicting “no” for employee attrition (in the 90s), but predicting “yes” was often far below 50% - sometimes in the 20s! Taking advice provided by Dr. Alferez, we resampled our data. While the highest results were not in the 90s anymore, the results were overall much more balanced.

There were a few experiments we ran in the process of optimizing our models. One was on algorithms to use for up-sampling the training data. There were two algorithms provided by Imbalanced-Learn: SMOTE and ADASYN. These both up-sample with synthetic data to equalize numbers of each class. Testing with these two algorithms was not extensive, and while they yielded very similar results, ADASYN performed slightly more poorly than SMOTE, so we chose the latter.

Another experiment we carried out involved adjusting how many features were selected in the feature selection process. We used Scikit-Learn’s SelectKBest algorithm, which allowed us to specify how many of the top features we wanted to pick. Out dataset initially had 34 features, and by experimenting with the number of top features selected, we were able to optimize our result. Taking the top twenty yielded a good result, but taking the top 14 seemed to be the best. With this number, our model accuracy f-1 was in the mid 60s.

Our final results are as follows. The best model used Support Vector Machines and had an accuracy f-1 of 70%. The f-1 scores of both classes were also only 4% apart, which was fairly good compared to other models. The second-best model

used Linear Regression and had an accuracy f-1 of 67%. This model may have done so well because we only had two classes to predict: “yes” and “no.”

1. Decision Tree

| **Metrics** | **Score** | | | |
| --- | --- | --- | --- | --- |
| ***Precision*** | ***Recall*** | ***F1-Score*** | ***Support*** |
| No | 0.60 | 0.79 | 0.68 | 38 |
| Yes | 0.69 | 0.47 | 0.56 | 38 |
| Accuracy |  |  | 0.63 | 76 |
| Macro Average | 0.65 | 0.63 | 0.62 | 76 |
| Weighted Average | 0.65 | 0.63 | 0.62 | 76 |

1. Linear Regression

| **Metrics** | **Score** | | | |
| --- | --- | --- | --- | --- |
| ***Precision*** | ***Recall*** | ***F1-Score*** | ***Support*** |
| No | 0.65 | 0.74 | 0.69 | 38 |
| Yes | 0.70 | 0.61 | 0.65 | 38 |
| Accuracy |  |  | 0.67 | 76 |
| Macro Average | 0.67 | 0.67 | 0.67 | 76 |
| Weighted Average | 0.67 | 0.67 | 0.67 | 76 |

1. Support Vector Machines

| **Metrics** | **Score** | | | |
| --- | --- | --- | --- | --- |
| ***Precision*** | ***Recall*** | ***F1-Score*** | ***Support*** |
| No | 0.67 | 0.76 | 0.72 | 38 |
| Yes | 0.73 | 0.63 | 0.68 | 38 |
| Accuracy |  |  | 0.70 | 76 |
| Macro Average | 0.70 | 0.70 | 0.70 | 76 |
| Weighted Average | 0.70 | 0.70 | 0.70 | 76 |

1. K Nearest Neighbors

| **Metrics** | **Score** | | | |
| --- | --- | --- | --- | --- |
| ***Precision*** | ***Recall*** | ***F1-Score*** | ***Support*** |
| No | 0.59 | 0.71 | 0.64 | 38 |
| Yes | 0.63 | 0.50 | 0.56 | 38 |
| Accuracy |  |  | 0.61 | 76 |
| Macro Average | 0.61 | 0.61 | 0.60 | 76 |
| Weighted Average | 0.61 | 0.61 | 0.60 | 76 |

1. Artificial Neural Networks

| **Metrics** | **Score** | | | |
| --- | --- | --- | --- | --- |
| ***Precision*** | ***Recall*** | ***F1-Score*** | ***Support*** |
| No | 0.59 | 0.79 | 0.67 | 38 |
| Yes | 0.68 | 0.45 | 0.54 | 38 |
| Accuracy |  |  | 0.62 | 76 |
| Macro Average | 0.63 | 0.62 | 0.61 | 76 |
| Weighted Average | 0.63 | 0.62 | 0.61 | 76 |

# Conclusions

Before you begin to format your paper, first write and save the content as a separate

text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**