Detection of Potential Customer Attrition Via Machine Learning

Kevin Tupin

Western Governors University

Data Management/Data Analysis

Capstone Project

Task 3:Project Report

Table of Contents

Project Overview 4

Research Question or Organizational Need 4

Scope of Project 4

Overview of solution 4

Tools 4

Methodology 4

Project Plan 5

Timeline and Milestones 6

Methodology 6

Data Collection 6

Advantages and Limitations of the Dataset 7

Advantages 7

Limitations 7

Data Extraction and Preparation 7

Data analysis 8

Advantages and Limitations of Tools and Techniques 9

Analysis Steps 9

Results 10

Statistical 10

Practical 12

Overall Success 12

Key Takeaways 13

Summary of conclusions 13

Storytelling 13

Recommendations 13

Panopto Presentation 14

Evidence of Completion 14

References 14

Project Overview

Research Question or Organizational Need

This project endeavored to develop a machine learning model that a bank could use to predict possible customer attrition based on demographic and credit card usage data. The goal upon completion of the project was to produce a machine learning model that could identify attrited customers versus existing customers better than a baseline model. The success criteria were that the resulting model would obtain scores at least .10 points higher than the baseline in no less than seven of the nine scoring metrics used.

Scope of Project

The scope of this project involved producing a learning model that predicts churn customers at a better rate than a baseline classifier. To produce this result, seven classifiers were tested to determine which was the model to be tuned to provide the final product. A limited selection of the available parameters was tested on the final model in order to stay within the scope of the project.

Overview of solution

Tools

This project was completed in JupyterLab using Python code. Extensive use was made of available Python libraries, particularly NumPy, Pandas and ScikitLearn. Matplotlib and Seaborn were used for visual representations.

Methodology

The SEMMA (Sample, Explore, Modify, Model, Access) methodology was utilized for this project. The process was used iteratively throughout the process.

Project Plan

The project plan outlined in the project proposal was followed. The steps, in accordance with the SEMMA methodology, were as follows:

Sample: The complete dataset was large enough to extract significant information and small enough to manipulate easily, so the entire dataset was used instead of taking a subsample.

Explore: Visual and mathematical analysis was used to identify trends in the data and check for missing, irrelevant, or redundant data.

Modify: The data was cleaned, and missing values were replaced. Irrelevant and redundant features were removed, and the data was transformed into a form that can be interpreted by machine learning models.

Model: A "dummy" classifier was utilized with varying parameters in order to produce the highest scores possible to use as a baseline.

Access: The scoring data from the baseline classifier was charted, and the highest score for each metric was set as a baseline for that metric. The absolute highest scores obtained for each metric were used, even though those high scores did not all come from the same parameter test.

With the baseline set, it was now time to iterate through the SEMMA steps.

Sample: There were no new datasets to sample.

Explore: The previous data exploration was sufficient, so no new analysis was conducted at this point.

Modify: The data was then split into a training and testing group in preparation for the model testing (the "dummy" classifier utilized the entire dataset, without it being split into groups).

Model: The seven chosen models were tested to determine which was highest scoring without being tuned. The highest scoring model was then tuned to further improve the accuracy.

Access: The final model scores were compared to the baseline scores.

Timeline and Milestones

Table 1

Projected Milestone Schedule

Milestone	Start Date	End Date	Duration
Install all needed tools and software	2/1/2022	2/1/2022	1 Day
Exploratory Data Analysis	2/2/2022	2/2/2022	1 Day
Clean and preprocess data	2/3/2022	2/5/2022	3 Days
Establish baseline metrics	2/6/2022	2/6/2022	1 Day
Compare classifier models	2/7/2022	2/7/2022	1 Day
Tune most accurate classifier	2/8/2022	2/8/2022	1 Day
Assess results and quality check code	2/9/2022	2/9/2022	1 Day

The majority of the project progressed as planned. The classifier models being tested processed the data faster than expected. The time to compare the classifier models was reduced from two days to one, bringing the project to completion one day ahead of schedule.

Methodology

Data Collection

Data collection was simple as the data was already in csv form. There were no obstacles or data governance issues. The data had been uploaded under a public domain license.

Advantages and Limitations of the Dataset

Advantages

The dataset was relatively clean. There was some missing data, but there were no major outliers. The size of the dataset (just over 10,000 records) was a good size for the machine learning model testing. This allowed more models and parameters to be tested than would have been feasible with a larger dataset.

Limitations

The size of the dataset could potentially have also impacted the study in a negative way. It is possible that a larger dataset would have produced even more accurate results. The two data classes were imbalanced, with the attrition class only comprising approximately 16% of the total records. The dataset was not large enough to utilize oversampling or undersampling comfortably. Also, though there was little missing data, the missing data all occurred in three of the categorical variables. Since it was not possible to obtain a mean for those columns, the mode was used. This may have skewed the end result somewhat since the modes did not appear in significantly larger numbers than the second most common values in their respective columns.

Data Extraction and Preparation

Because the data was contained in a csv file, the Pandas read_csv method was used to extract the data. Data preparation was done first to perform basic analysis of the dataset, and then to convert the data into a form usable by the machine learning models. The main preparations steps were as follows:

1. Two of the columns in the dataset contained data based on previous analysis. These two columns were removed so as not to influence the results of this project.

2. The column labeled ‘CLIENTNUM’ was dropped from the dataset as it contained client ID numbers that were not relevant to the analysis or machine learning processes.
3. The data was divided into numeric and categorical data frames to facilitate exploratory data analysis.
4. A column that contained redundant data was dropped.
5. Missing data points were converted to ‘NaN’ for further exploration.
6. Missing data was replaced with the modes of their respective columns.
7. All categorical data was converted to numerical data for use by the machine learning models via replacement for categories with a natural hierarchy, and the use of “dummy” variables for all others.
8. After the baseline model was tested the data was split into training and testing groups to test the machine learning models.

These methods were appropriate for the purpose of facilitating the machine learning process.

Data analysis

Both descriptive and predictive analysis methods were utilized. Descriptive methods were used to analyze the data to best determine the preparation steps used to ready the data for the machine learning models. In addition to inspecting the data itself in the data frames, Matplotlib and Seaborn tools were used to visualize patterns in the data. The analysis and comparison of the results of the machine learning models was also descriptive. This involved both charting the metrics in data frames as well as creating confusion matrices to assess

performance. The machine learning models and their output involve predictive methods. These utilized ScikitLearn algorithms and metrics.

Advantages and Limitations of Tools and Techniques

The primary tools used were JupyterLab and ScikitLearn machine learning models. The main advantage of JupyterLab is the grouping of tools that have been brought together. In addition to the improved functionality of Jupyter Notebook, there is a csv viewer that parses the file into a more readable format and several new viewing options. Limitations of the tool are that the csv viewer does not have the sort functions that are available in tools such as Excel, and there are not as many stable extensions available as in the standard Jupyter Notebook. ScikitLearn has the advantage of having several different models and parameters, but the potential drawback is that the overwhelming possibilities mean that the best solution will often be unused. The technique used for the final model was Random Forest classification. Random Forest has the advantage of generally providing high accuracy but is not as customizable as many of the other models.

Analysis Steps

The main analysis steps were as follows:

1. The data was examined for null values.
2. The data types of all columns were examined
3. The distributions of all numerical columns were examined via histograms.
4. The individual labels in the categorical columns were compared to the customer attrition status column via use of a series of count plots.
5. A correlation heatmap was utilized to search for potentially redundant columns.

6. Once the data was prepared for the machine learning baseline model it was examined to ensure that all values were numerical and that no null values existed.
7. After the baseline model was run, the results were assessed using ScikitLearn metrics. The results were examined using a data frame with a gradient color palette to easily visualize performance rankings for the various metrics.
8. The results of the models that were being compared were examined via the same process that the baseline model was assessed in order to select the best performing model for further tuning. A series of bar charts was also created to visualize the differences.
9. After the final tuned model was run, the results were visually compared to the baseline model using confusion matrices and a bar chart.
10. The final model metric scores were numerically compared to the baseline metrics to assess whether the success criteria had been met.

Results

Statistical

The model was to be considered successful if it achieved an improvement over baseline of .10 or greater on at least seven of the nine scoring metrics that were used:

- Accuracy
- Precision(binary)
- Precision(micro)
- Precision(macro)
- Precision(weighted)
- Recall

- f1
- fbeta (beta = 2)
- Matthews correlation coefficient

The final model showed improvement over baseline on all nine of the metrics, with the lowest improvement being 0.125 points, and the highest being 0.851 points.

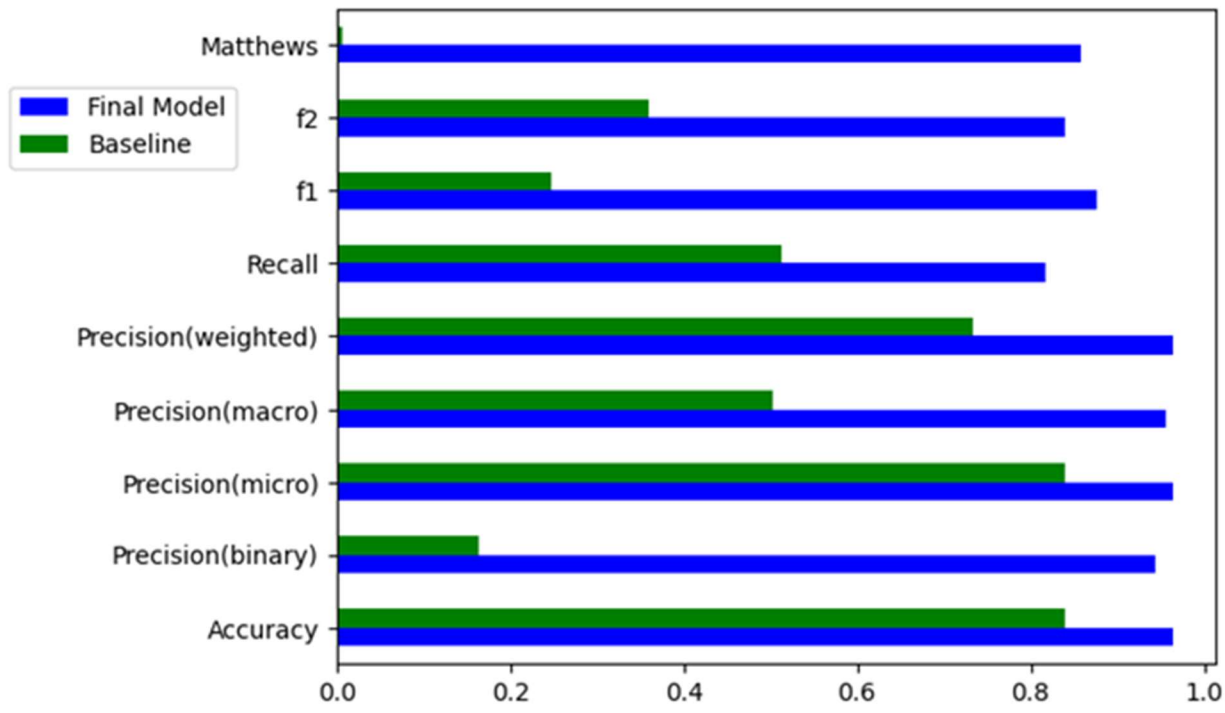
Figure 1

Baseline Comparison Data frame

	Accuracy	Precision(binary)	Precision(micro)	Precision(macro)	Precision(weighted)	Recall	f1	f2	Matthews
Final Model	0.964462	0.944444	0.964462	0.955992	0.963983	0.817308	0.876289	0.839921	0.858715
Baseline	0.83934	0.163894	0.83934	0.502676	0.732997	0.513215	0.247774	0.359263	0.007286
Difference	0.125122	0.78055	0.125122	0.453316	0.230986	0.304093	0.628514	0.480658	0.851428
Meets Success Criteria	YES	YES	YES	YES	YES	YES	YES	YES	YES

Figure 2

Baseline Comparison Chart

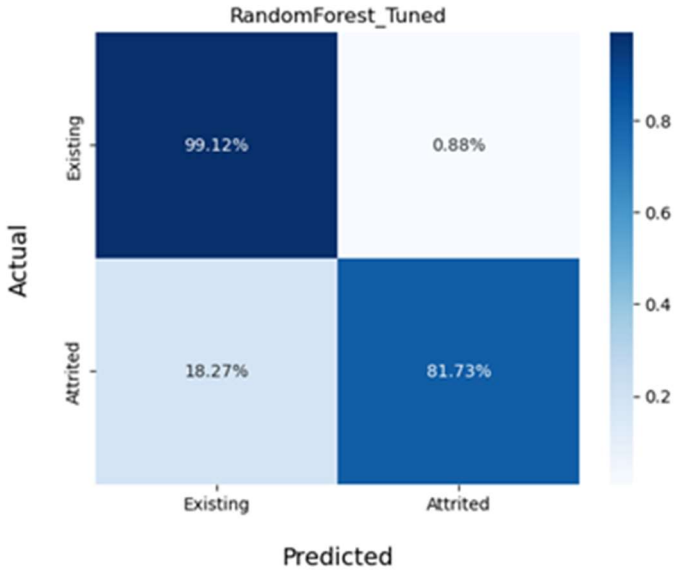


Practical

The practical significance of machine learning models that can improve the detection and mitigation of customer churn is immense. As noted in a study published by the Harvard Business Review, reducing customer attrition by 5% can boost profits 25-85%. (Reichheld & Sasser, 1990) The final model produced by this project was able to accurately identify nearly 82% of the attrited customers while only incorrectly identifying less than 1% of existing customers as attrited, despite the attrited customers only accounting for approximately 16% of the total. This means that the company can focus retention efforts more efficiently without expending resources unnecessarily on customers that are unlikely to leave.

Figure 3

Final Model Confusion Matrix



Overall Success

This project exceeded the stated goal of exceeding seven of the nine metrics by 0.10 or more. The final model outperformed the baseline model by more than the desired amount on all

nine of the goals. Seven of the nine metrics showed an improvement of 0.23 or greater, and three of the nine metrics indicated improvements of 0.62 or greater. The outcome is therefore considered to be a success.

Key Takeaways

Summary of conclusions

This project set out to determine whether a machine learning algorithm could be found and tuned to perform better than a baseline model in the detection of customer attrition. The final model was not only able to do so, but by a significant margin and across several different metric categories.

Storytelling

The tools and graphical representations used were selected in order to show pertinent information in a simple yet effective manner. The graphs and charts provide quick references to the pertinent features of the data without cluttering the graphics with unneeded information.

Recommendations

Two recommendations to provide further value to the company are as follows:

1. Expanded and/or more granular data collection. For example, the usage change data only reveals the change from Q1 to Q4, and only for one given year. Data showing the changes month to month, quarter to quarter and year to year could provide insight into trends that shed more light on patterns leading up to customers leaving, possibly allowing these customers to be identified even earlier or more accurately. It could also be beneficial to compare the customer data against similar data from other companies, or the industry as a whole, if that data can be sourced legally and ethically.

2. Expanded investigation into the data that has been gathered and the ability of machine learning algorithms to provide even deeper insights. The testing of the final machine learning model determined that utilizing only five features provided the most accurate results. Further research into what those five features are and what it is about those features that enhance the prediction power could further improve understanding of the factors that reveal customer attrition, and possibly even give clues as to why customers leave.

Panopto Presentation

A Panopto presentation regarding this project can be found at https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=2863c4f9-1431-431e-887b-afa8010a1811

Evidence of Completion

The following will be submitted with this project as evidence of completion:

- The original csv file containing the data used for the project
- An ipynb file containing the Python code written for the project and all outputs
- An HTML file of the same

References

Reichheld, F. F., & Sasser, W. E. (1990, September). Zero defections: Quality comes to services. *Harvard Business Review*. https://hbr.org/1990/09/zero-defections-quality-comes-to-services