| **DATA 430 Technical Report Assignment 2: Bayesian Classification** | **Justin Blain** |
| --- | --- |
| **Analysis of key factors in startup success** | |
| **URL to dataset: https://www.kaggle.com/datasets/manishkc06/startup-success-prediction/data** | |

This template should be used in conjunction with the assignment instructions. The size of the text area below will expand to the length of your response; the area should not be interpreted as a required or suggested length of response. Responses within the text area should be single spaced with Times New Roman 12pt font. The body of the document will likely be 6-9 pages, not including the Appendix; length may vary depending on specifics of the analysis and the dataset. As needed, APA format in-text citations should be included, along with a full references list at the end of the document.

| **Overview** |
| --- |
| **Problem Domain**: give some background and context about the problem domain (application area). For instance, if you are doing the analysis for predicting heart disease, provide some context about the disease and include some interesting statistics about it. Also, discuss how the method is relevant for the chosen problem. |
| When new businesses are first starting up and very dependent on outside investors, it can be very difficult to project whether the entire endeavor will be a flop and a waste of time and money or whether it will create a massive windfall for both founders and investors. Being able to predict which qualities are most important in a startup’s success can be a confounding endeavor for both investors and entrepreneurs. In this case, looking at the records of previous can help to predict which new startups are more likely to succeed. This can inform our investors which startups are likely to be a good investment. Using a logistic regression type classifier allows us to also generate a confidence interval which can be used to infer an appropriate level of risk with which to invest. |
| **Objective**: clearly state the objective of the analysis in relation to the kind of algorithm you are employing. Use specific language as to what question(s) you are trying to answer using the specific analysis/modeling type. |
| A tool that could predict which start-ups will achieve unicorn status (over 1 billion valuation) with high confidence would radically increase an investor’s profit margin. Ideally, this analysis will allow us to see which features are perhaps most important to guide those in the process of creating a startup as well as creating a tool that supports an investor's decision-making.  The data set with which we’re working today contains a lot of categorical data. This means that a Bayesian classifier could potentially make good use of the categorical features which proved difficult to leverage with a logistic regression model. |
| **Analysis** |
| **Exploratory Analysis**: describe the data including the source, the collection method, and variables. Perform exploratory analysis. Also, select a few key variables (including the target variable for supervised learning) and study their distributions using plots such as histograms, box plot, bar chart, etc. |
| This data set contains information regarding the early years of many startups’ context as the independent variables. armed with this data we can attempt to build a classifier capable of predicting the success or failure of such startups.  Found on Kaggle and provided by Ramkishan Panthena at GMO. User Mannish KC on Kaggle’s Data card describes some of the features as follows:  age\_first\_funding\_year – quantitative  age\_last\_funding\_year – quantitative  relationships – quantitative  funding\_rounds – quantitative  funding\_total\_usd – quantitative  milestones – quantitative  age\_first\_milestone\_year – quantitative  age\_last\_milestone\_year – quantitative  state – categorical  industry\_type – categorical  has\_VC – categorical  has\_angel – categorical  has\_roundA – categorical  has\_roundB – categorical  has\_roundC – categorical  has\_roundD – categorical  avg\_participants – quantitative  is\_top500 – categorical  status(acquired/closed) – categorical (the target variable, if a startup is ‘acquired’ by some other organization, means the startup succeeds) |
| **Preprocessing**: armed with the exploratory analysis, perform the necessary preprocessing, both general and specific types appropriate for the modeling type being employed. |
| Bayesian regression should allow us to work with categorical data. We should only need to  df = df.dropna()  (unfortunately this only leaves us with 120 entries to work with)  and then to encode the categorical data:  object\_columns = df.select\_dtypes(include=['object']).columns.tolist()  le = LabelEncoder()  for column in object\_columns:  df[column] = le.fit\_transform(df[column]) |
| **Model Fitting**: explain the key steps and activities you perform to fit the model. Experiment (as appropriate) with parameters tuning. This is key, what separates highly accurate model from a less accurate ones is the amount of performance tuning performed. |
| I found the GaussianNB and BernoulliNB classifiers to each achieve the best results even if on relatively small proportions of data.  tuning the ‘var\_smoothing’ parameter  **params\_NB = {'var\_smoothing': np.logspace(0,-9, num=100)}**  **gs\_NB = GridSearchCV(estimator=clf,**  **param\_grid=params\_NB,**  **cv= ShuffleSplit(n\_splits=5, test\_size=0.2, random\_state=42),**  **verbose=1,**  **scoring='accuracy')**  **gs\_NB.fit(X\_train, y\_train)**  **print(gs\_NB.best\_params\_)**  gives the following result  **Fitting 5 folds for each of 100 candidates, totaling 500 fits**  **{'var\_smoothing': 1.0}**  but using this variable does not increase the score any    oh… well then. |
| **Results** |
| **Model Properties:** explain the components of the fitted model and their characteristics. Leverage functions to summarize the model properties. Also, leverage visualization as required. |
| GausionNB Assumes a normal distribution so outliers must be estimated or smoothed to fit that assumption. However, as shown above, either manipulating this variable does not effect the out come or there is an error in my calculations. |
| **Output Interpretation**: explain the result and interpret the final model output using terms that reflect the application area and in relation to the stated objective. This is where you check whether or not the stated objective is met. |
| The first value in our target variable column is “acquired”. This means that our binary encoding of that column will label that value as zero meaning that a “positive” indication in this model predicts the eventual closure of that start-up.  One could simplify the metric’s output to say that this model has a 96% chance at correctly predicting the outcome of a given start-up using all the available categorical data. |
| **Evaluation**: employ appropriate metrics to quantitatively evaluate the performance of the fitted model. For supervised classification, this includes simple accuracy, precision & recall (or sensitivity & specificity), all of which can be generated from a confusion matrix, or ROC. |
| precision recall f1-score support  0 0.00 0.00 0.00 1  1 0.96 1.00 0.98 23  accuracy 0.96 24  macro avg 0.48 0.50 0.49 24  weighted avg 0.92 0.96 0.94 24 |
| **Conclusion** |
| **Summary**: highlight the main findings in relation to the stated objective. You don’t need to discuss the details of the analysis and the model such as accuracy here, just focus on the key findings. |
| Using all of the categorical data allows us to predict the end status of a given start up very precisely. Our best model achieves the following classification report:  precision recall f1-score support  0 0.00 0.00 0.00 1  1 0.96 1.00 0.98 23  accuracy 0.96 24  macro avg 0.48 0.50 0.49 24  weighted avg 0.92 0.96 0.94 24 |
| **Limitations & Improvement areas**: discuss the limitations of the analysis and identify potential improvement areas for future work. This could be related to the data, algorithm, or a combination of the two. |
| I could attempt to implement smoothing to address the zero frequency problem |

| **Appendix** |
| --- |
|  |

**References**

**https://www.kaggle.com/datasets/manishkc06/startup-success-prediction/data**