CPSC 66 Final Report: Analysis of Feature Selection Across Different Machine Learning Algorithms

Scout Clark ACLARK2@SWARTHMORE.EDU
Swarthmore College, 500 College Avenue, Swarthmore, PA 19081 USA

Henry Han HHAN3@SWARTHMORE.EDU
Swarthmore College, 500 College Avenue, Swarthmore, PA 19081 USA

Abstract

Feature importance and dimensionality reduction are important for effectively visualizing and interpreting real-world datasets, as well as improving prediction accuracy. Using the Students Academic Performance Dataset[1,2] from Kaggle, we implemented support vector machines, Bayesian Networks, and logistic regression with L1 and L2 norms. These algorithms with various hyperparameters were implemented to determine feature importance and predictive power of the models. These results were then compared to the important features determined from the preprocessing techniques: Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Using R^2 values, which explain the percentage of the response variable explained by the variation in the model, it was found that L1 regularized logistic regression and Bayesian networks best fit the data. The higher R^2 values identify that these algorithms had the most predictive power, allowing them to identify the most important features in the dataset. SVM, LDA, and L2 regularized logistic regression least accurately fit the data with SVM being the next highest R^2 value and L1 being the lowest. The important features between Bayesian networks, PCA, and LDA were compared, while the irrelevant features determined by SVM and L1/L2 logistic regression were also evaluated.

1. Introduction

For any given dataset, data scientists aim to build the simplest model with the most predictive power. This predictive power depends on the influence of each feature in the dataset on a given label. However, most datasets pulled from real-world activities are often too complex and contain too much noise to easily determine a feature's significance in the model, or they are filled with irrelevant features. For example, in a dataset with 300 features, maybe only 50 of them actually affect the label outcome, which would make for a much simpler model to visualize and interpret than a model with 300 features. Similarly, many datasets suffer from the notion of the curse of dimensionality (COD), which is a phenomenon that occurs when data becomes sparse in a high dimensional space because as more dimensions are added, the amount of data needed to define the feature space increases exponentially. As such, it is important to perform feature selection and/or dimensionality reduction as a preprocessing or implicit algorithmic technique on complex datasets to obtain meaningful predictive information.

Feature selection generally consists of identifying the features that are most important in predicting an example label through some statistical method like Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or the model's R^2 value. It can also improve the accuracy of the resulting model by removing redundant, irrelevant, or noisy data.[3] Dimensionality reduction involves the mapping of a dataset into a lower dimensional feature space. The variance in the data is mostly represented through the reduction of the data to only the most relevant features or discriminative components.[3] There is much overlap between the two methods, as both encourage a simpler model that is in a lower dimension. Similarly, some algorithms do feature selection and dimensionality reduction at the same time.

Using a Students Academic Performance Dataset[1,2] from Kaggle, we explore how machine learning algorithms such as support vector machines, Bayesian networks, L1 logistic regression, and L2 logistic regression, han-

CPSC 68 Bioinformatics Final Report Proceedings, Swarthmore College, Spring 2017.

dle feature importance and dimensionality reduction to determine the predictive power of features in a dataset. Additionally, we use preprocessing techniques such as Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to extract feature importance. We are interested in how the preprocessing and modeling approaches compare to each other in terms of the top four important features chosen from each process. Specifically, we compare which features were chosen and the R^2 values among the algorithms to determine fit of the model to the data and predictive power. Moreover, the relative weights, coefficients, and/or r-values among predicted important features from each method are also compared across the models.

Based on the R^2 values from each model, we found that L2 regularized logistic regression and Bayesian networks were the best fit for the data; i.e. as models, they had the most predictive power and could identify the most important features in the dataset. SVM was performed for each 1-vs-all label and was the second best fit model. LDA and L1 logistic regression were the least accurately fitted to the data, and had the least predictive power to allow them to perform feature selection. Moreover, we compared the important features from the Bayesian network, LDA, and PCA and found similarities among the features chosen. For SVM, L1 logistic regression, and L2 logistic regression, we found that some of the same features were deemed as irrelevant and not important to the predictive power of the model.

2. Related Work

Hira et al. explored various models that could perform dimensionality reduction on a gene microarray dataset, resulting in accurate classification of new examples in the future. Moreover, they evaluated and compared the most efficient feature selection techniques for simplifying the gene microarray dataset.[4]

Hira et al. splits feature selection into three categories: filtering, wrapping, and embedding techniques. Filtering is a type of feature extraction that does not involve learning. Univariate (features are evaluated separately) and multivariate (dependencies between features are considered) techniques are described. Most notably, information gain ranking (univariate), describes the conditional dependency between the feature and the label and determines feature importance from those rankings. Correlation techniques such as Correlation-based Feature Selection (CFS) classify good features as ones that are highly correlated with the class labels but not one another.[4] We explore the information gain and CFS filters when implementing Bayesian networks in this paper.

For deterministic wrappers, a mix of PCA and SVM approaches were used in a sequential forward selection (SFS), where each feature is evaluated and is only added until the evaluation of the feature does not constitute improvement (i.e. until it converges). The feature with the highest score is permanently selected. This process is repeated until the all of the most important features are selected.[4]

The embedded algorithms discussed include those that have LDA as a preprocessing step followed by recursive SVMs. The most important features are chosen through a hard SVM method, where features are thrown out based on their weight. This notion is further supported using cross-validation.[4] In addition, many classifiers do internal feature selection including logistic regression with regularization. We use aspects of deterministic wrappers and embedded algorithms in our analysis of feature importance and will identify if these algorithms are more or less efficient than the logistic regression regularization algorithms chosen for this project.

Haury et al. evaluates the accuracy of feature selection techniques on breast cancer prognosis datasets. The models they analyze include wrapper methods such as SVM and embedded methods such as Lasso regression (L1 regularized logistic regression) and the elastic net, which is a combination of the L1 and L2 norms of logistic regression. They performed all of these models using an ensemble method in order to obtain the best accuracy per model possible. Accuracy of each model was determined by the area under the ROC curve. The experiment concluded that elastic net and Lasso regression were the most accurate. SVM in comparison to the other models was the most computationally expensive and did not provide the best accuracy.[5] While we did not directly compare the accuracy of feature selection techniques, we analyzed feature selection in our experiments of L2 regularized logistic regression and linear SVM based on fit of the models to the data (R^2 values) and comparison of important features selected.

3. Methods

3.1. L2 Regularized Logistic Regression

In this project, we will look at two regression analysis methods that regularize the logistic regression model: logistic regression with an L1 norm and logistic regression with an L2 norm.

Logistic regression is a classifier that models the probability of distribution of labels in a dataset given their feature

values by assigning feature vectors regression weights.[6] The L2 regularizer for logistic regression (similar to Ridge regression) penalizes terms that could overfit the model by summing the squares of the regression coefficients (Eq. 1).[7]

$$(1)\ \mathbf{w}^* = arg\min_{\mathbf{w}} \sum_j (t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j))^2 + \lambda \sum_{i=1}^{k} w_i^2$$

It is also known as the Least Squares Error (LSE) function because as it minimizes the sum of the squared errors of the weights, it also decreases variance within the dataset (Fig. 1). By doing so, the L2 regularizer also alleviates multicollinearity, the phenomenon in which one predictor variable in a regression model can be linearly predicted from the others with a substantial degree of accuracy.[7] By reducing multicollinearity, the L1 regularized model allows for more accurate analysis on individual features and their importance in the model. While the L1 norm sums sparse coefficients (i.e. some coefficients become zero), which is ideal for feature selection, the L2 norm does not do this. Instead, the L2 norm still affects the weights, but the coefficients are simply large or small, not zero. As such, feature selection can still be performed by looking at the larger coefficient values.[7]

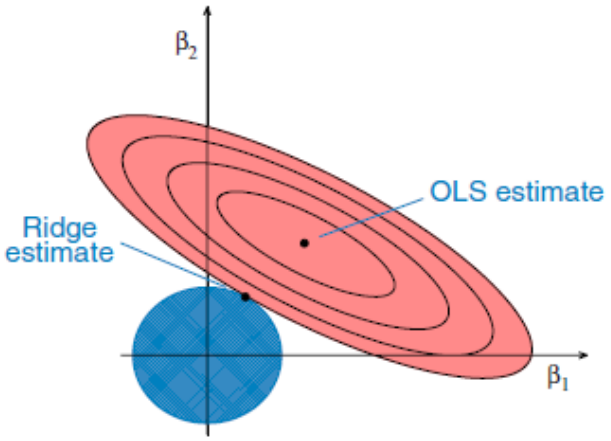


Figure 1.[1] The red ellipses represent the residual sum of squares, with the minimization of the data occurring at the ordinal least squares (OLS) estimate (the smallest ellipse). The blue circle represents the L2 penalty (also known as Ridge). The minimal circle and ellipse size, or where the circle and ellipse meet is the L2 logistic regression estimate. As such, L2 regularized logistic regression minimizes variance in the model.[7]

[1]Taken from https://onlinecourses.science.psu.edu/stat857/node/155

3.2. L1 Regularized Logistic Regression

The L1 regularizer for logistic regression (which is very similar to Lasso regression) converges to the global maximum for a Lagrangian constraint optimization problem.[8] Logistic regression is a model of the probability distribution of the label given the feature vector. For L1, there is a penalty (i.e. the LaPlacian prior) applied to the MAP estimation of θ to prevent overfitting (Eq. 2).[9]

$$(2)\ \mathbf{w}^* = arg\min_{\mathbf{w}} \sum_j (t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j))^2 + \lambda \sum_{i=1}^{k} |w_i|$$

L1 norm performs both feature selection and regularization to increase predictive accuracy and interpretability of a model.[8] The model achieves this by "minimizing the residual sum of squares with the constraint that the sum of the absolute value of the coefficients is less than a constant" (Fig. 2).[8] This leads to coefficients that are exactly 0, and thus has the same effect as being taken out of the model. Moreover, as the complexity parameter, C, decreases, more features are reduced to zero as more of a penalty is applied.[9]

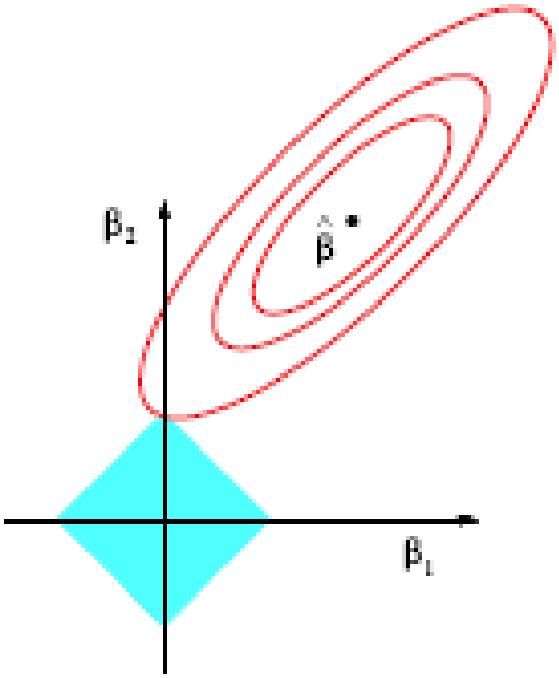


Figure 2.[2] Much like L2 logistic regression, the red ellipses represent the residual sum of squares of the data,

[2]Taken from https://onlinecourses.science.psu.edu/stat857/node/158

with the smallest ellipse being the OLS estimate. The blue square represents the L1 penalization, where θ values may or may not be zero depending on the importance of the feature. As such, L1 logistic regression effectively removes unnecessary features. The value at which the minimal square size and the smallest ellipse meet is the L1 estimate.[9]

3.3. Linear Support Vector Machines

Another model we use in our analysis is linear support vector machines (SVM), which is an optimal hyperplane classifier. The hyperplane represents the maximum margin between two classes of functions, which is created through a constrained quadratic optimization problem.[10] Any training examples that lie on the margin, otherwise known as support vectors, affect the model's classification ability (Fig. 3). The solution to the optimization problem for each example is the coefficient or weight of the training vector and also corresponds to the feature importance of each vector. If a coefficient is high relative to other feature coefficients, the feature is ranked as more important.[11]

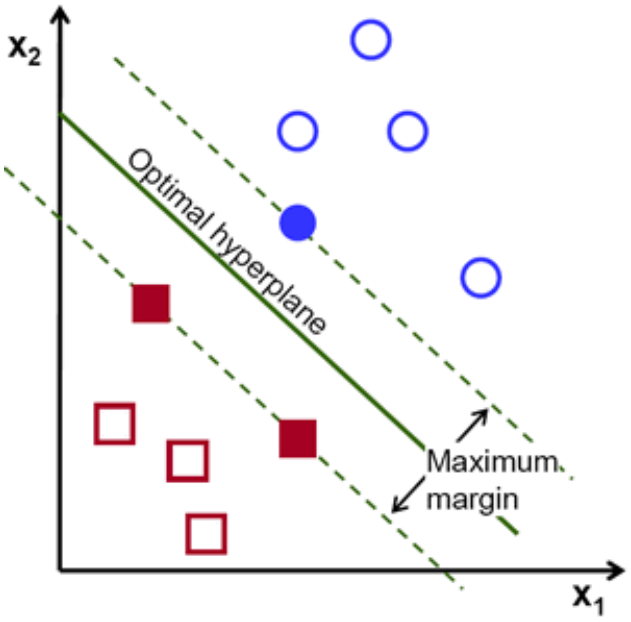


Figure 3. The maximum margin is placed such that it is the farthest from all feature vectors. Those feature vectors that lie on the margin are known as support vectors and are the only features that affect the classification of the two groups on either side of the hyperplane. According to the SVM model, the vectors with the highest coefficients are the more important features in the dataset.[3]

[3]Taken from https://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

3.4. Bayesian Network

The last model we identified as a feasible feature selection method for this project was a Bayesian Network. A Bayesian Network involves the use of a Directed Acyclic Graph (DAG) with nodes that contain features with conditional probabilities and directed edges between features (Fig. 4). The makeup of this model is important in that it illustrates the conditional dependencies of features and allows more in-depth questions to be asked about a dataset, rather than simply questions about predictive power. Important features can be extracted from this model in a number of ways. Some methods that we evaluate are gain ratio and information gain rankings of features, as well as correlation between features and labels to extract the most significant features in the dataset.[12]

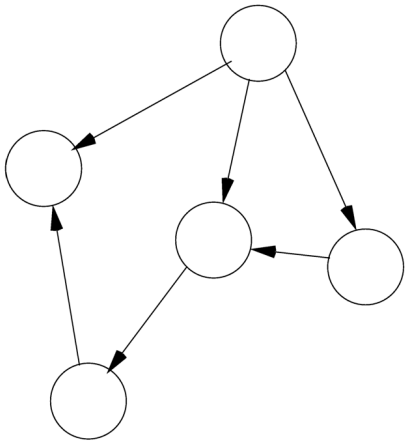


Figure 4.[4] The above DAG is a simple representation of a Bayesian Network. Given the directed nature of the graph, features may be conditionally dependent on the observation of preceding features. For example, the leftmost label is dependent on two features. Additionally, the graph must be acyclic in order for the idea of feature conditional dependencies to hold true. A feature is considered important given the value of the edges (by correlation, information gain, gain ratio etc.) between other features and/or labels.[12]

[4]Taken from http://www.cse.unsw.edu.au/~cs9417ml/Bayes/Pages/Bayesian_Networks_Inference.html

3.5. PCA and LDA

We will be comparing the above four algorithms to pre-processing techniques that have built-in feature selection properties: PCA and LDA. PCA is a linear transformation method that projects a high dimensionality feature set onto a lower dimension space without looking at feature labels (Fig. 5). This is done by calculating eigenvectors that represent linear combinations of the feature space and can be represented by their eigenvalues (i.e. a coefficient of the unit-scaled eigenvector). These eigenvector/value pairs represent the principal components of the new feature space. A simpler/lower dimension feature space is created by taking the highest values eigenvectors, which allows the less important features to be phased out. Feature importance can be obtained by looking at the linear combinations in the first few eigenvectors with the highest eigenvalues. Since the eigenvectors are unit-vectors by nature, the vector can be multiplied by a loading (scaling) factor to obtain the most prominent features in each eigenvector.[13,14]

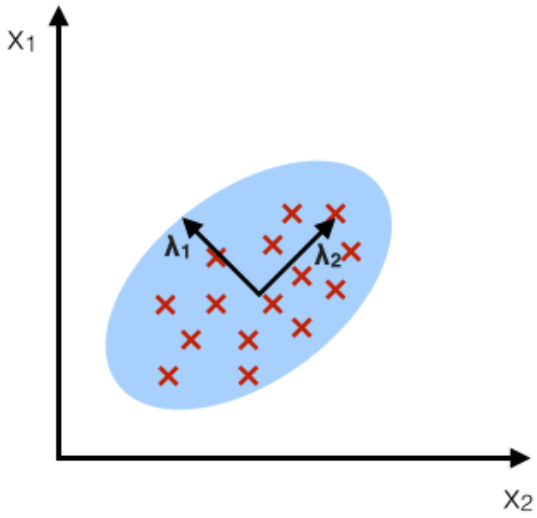


Figure 5.[5] Since PCA does not take labels into account, the highest eigenvalues represent the best projected axes (λ_1 and λ_2) of the feature set from high to low dimension. In other words, λ_1 and λ_2 represent the component axes that best maximize the variance of the lower dimension feature set.[13,14]

LDA also involves linear transformation of high dimensionality features onto a lower dimensional space, but it takes the labels of the features in account. In particular, it maps out the new subspace to be able to distinguish between different class labels (Fig. 6). While the new feature space will look different, LDA still involves taking the linear combination of features and taking the higher eigenvector/value pairs to project onto the new space. As such, feature importance can also be determined by extracting the most prominent features in each linear combination through the loading or scaling multiplier.[13,14]

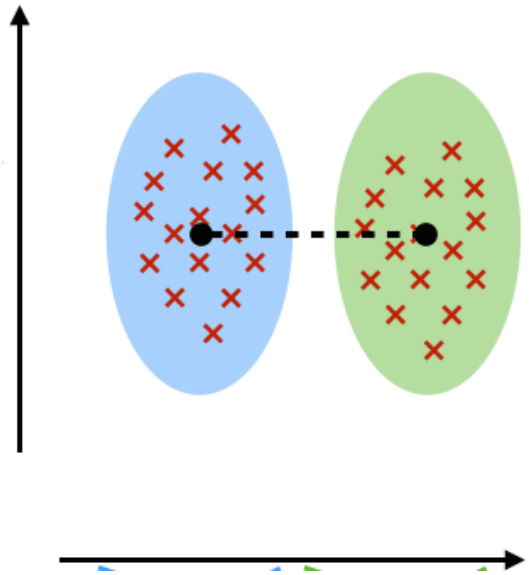


Figure 6.[6] LDA performs feature label analysis when executing dimensionality reduction. The dotted line between the blue and green ovals represents the maximum component axis calculated with the LDA model. This axis maximizes the separation between two different class labels as it projects the feature set into a lower dimensional space.[13,14]

The above algorithms and methods provide different approaches to feature selection and dimensionality reduction. Some perform selection implicitly while others have the ability to have feature selection extracted from already calculated values (i.e. PCA and LDA). We have performed each of these methods on the Students Academic Performance Dataset and compared the results of what features were deemed important with each algorithm and the specific hyperparameters used.

[5]Taken from http://sebastianraschka.com/Articles/2014_python_lda.html

[6]Taken from http://sebastianraschka.com/Articles/2014_python_lda.html

4. Experimental Results

4.1. Experimental Methods

The Students Academic Performance Dataset was taken from Kaggle and consists of 16 features, 480 instances, and three class label values. Since the labels were multinomial with categories of low level student, medium level student, and high level student, the labels were turn into three 1-vs-all columns. This allowed us to run SVM on each class label and obtain the top features for each. The L1 and L2 regularizer of logistic regression and linear SVM were implemented using the SciKit Learn API. The C (complexity) values were tuned. Smaller C values specify stronger regularization/penalty. Additionally, the training algorithm is different for both L1 and L2 penalties, and as such, was tuned for both norms as well. For the L1 penalty, the SAGA solver (A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives) was chosen; for the L2 penalty, the Newton-CG Augmented Lagrangian Method solver was chosen. Weka allowed us to construct a Bayesian network that could be based on information gain, gain ratio, or correlation to learn the edges of the graph. The number of principal components in the PCA model was tuned to be 4 (Fig. 7).

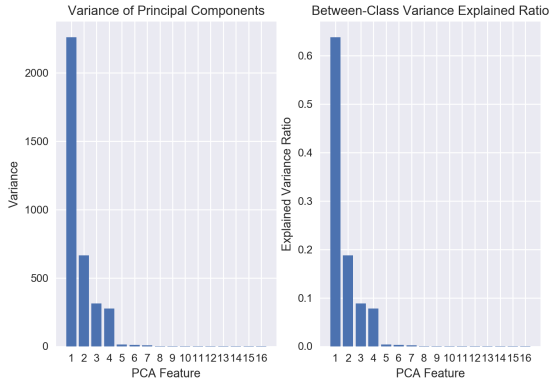


Figure 7. Only the first four principal components from PCA have a significant variance and explain a significant amount of variance in the dataset. Thus, the number of components in the PCA model was tuned to be 4.

For LDA, the number of components is 1 less than the number of class labels in the dataset. Thus, the number of components in the LDA model is 2 (Fig. 8).

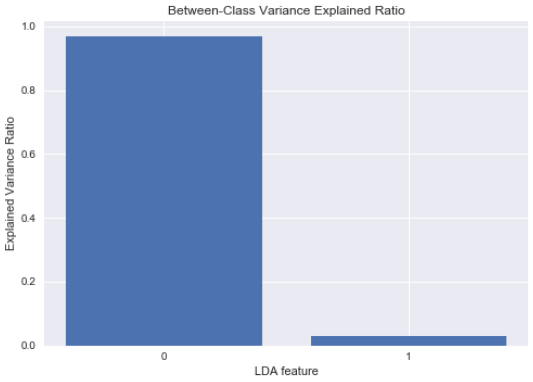


Figure 8. The first LDA feature explains almost all of the variance in the model.

4.1.1. Hypothesis: Comparison of Algorithms

We hypothesize that L1 regularized logistic regression will exhibit the best model fit for the data and will have the most predictive power (i.e. be able to identify important features) as it can make irrelevant feature weights zero and essentially toss them out of the model. Additionally, since, Bayesian networks are generative and describe the feature space, as well as take into account feature dependencies, we would expect it to also fit the data more accurately and to determine important features through its predictive power. We believe PCA and LDA will be the least accurate, since the classification is multinomial. Since LDA takes feature labels into account, it is suspected that LDA will be better than PCA at selecting important features that influence correct label prediction. While we cannot infer which model will be better, we can generalize that SVM and L2 logistic regression will fall somewhere in between the models described above. It is assumed L2 will be worse than L1 because L2 does not implicitly perform feature selection. Since SVM is a discriminative model, we suspect that it will not be able to fit the data as well and will be least accurate at predicting important features because it does not describe the feature space.

4.1.2. Evaluation Techniques

The models will be evaluated on accuracy of predicting important features by comparing their R^2 values and the features that they predict. Moreover, the r-values, coefficients, weights, or Eigenvalues/Eigenvectors of the features depending on the model, will be compared within each model to determine the extent in which a certain feature was predicted as important. Through these qualitative and quantitative methods, we will evaluate the model fit and predictive power of each model and its ability as a feature selection algorithm.

4.2. Results

A Bayesian network was run three times with information gain, gain ratio, and correlation criteria. The top 5 features for each are shown in the following schematics (Table. 1). The numbers next to each feature pertain to the r-value (or the ranking) of the feature in predicting a given label. The overall r^2 value of the model is .8164, meaning that 81.64% of the variation in Class can be explained by variation in the model.

Information Gain Ranking Filter

r-value	Feature
0.45801	VisitedResources
0.39745	StudentAbsenceDays
0.37337	RaisedHands
0.2578	AnnouncementsView
0.1504	ParentAnsweringSurvey

Gain Ratio Feature Evaluator

r-value	Feature
0.40986	StudentAbsenceDays
0.25378	RaisedHands
0.19878	VisitedResources
0.17636	AnnouncementsView
0.15212	ParentAnsweringSurvey

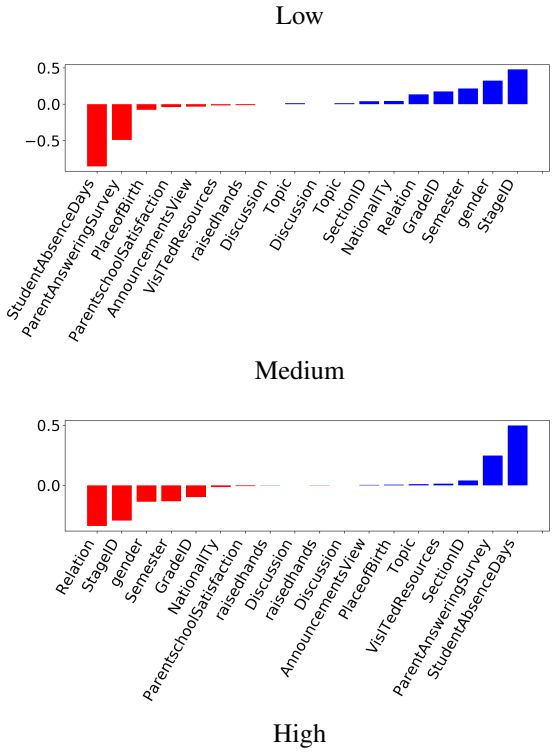
Correlation Ranking Filter

r-value	Feature
0.3829	VisitedResources
0.3608	StudentAbsenceDays
0.3283	RaisedHands
0.2895	AnnouncementsView
0.2369	ParentAnsweringSurvey

Table 1. Information gain, gain ratio, and correlation-based classification of the important features in the Kaggle dataset. The top five features for all three methods were the same, with only the top three differing in order between the information gain and correlation models and the gain ratio model. Information gain has the highest r-values for the first three terms, suggesting that it is strongest indicator of feature importance for this dataset.

In both the correlation and information gain criteria, the VisitedResources label was ranked as the most important feature. Moreover, the information gain run had the highest r-values in the top three features, while the other two parameters spread out the r-values amongst the top five features. Information gain highly correlates the top three features, VisitedResources, StudentAbsenceDays, and RaisedHands, to the values of the labels. Gain Ratio also ranks the top three features as most important (the same three as information gain but in a different order). The correlation ranking behaves differently, however, and ranks the first five features quite evenly in correlating to the feature labels. While these different parameterizations were not run with test set data, based on the r-values of the features, it seems that the information gain parameter makes for the simplest model as the first three features are the most heavily weighted. It is also important to note that the Bayesian network was run only once with all multinomial features included. In other words, the top features for the dataset as a whole were predicted, which is different than SVM and L1/L2 logistic regression where we analyze the top features for each class label (1-vs-all) of high, medium, or low.

SVM was run three times with each multinomial class label (i.e. the three labels were separated into 1-vs-all columns). Since SVM feature importance is based on the feature vector weights that form the positive and negative classifications on either side of the hyperplane, the larger bars on the following graphs represent the most important feature for each label (low, medium, and high) (Fig. 7). The R^2 value of the model for the high level student label is .7375, meaning that about 74% of the variation in the linear SVM model with the "high" label is explained by the variation in the model. The medium level labeled SVM has an R^2 value of .8229 and the low level labeled SVM has an R^2 value of .5583. These values suggest that the SVM model best fit the data when the class was defined as medium since it has the highest R^2 value.



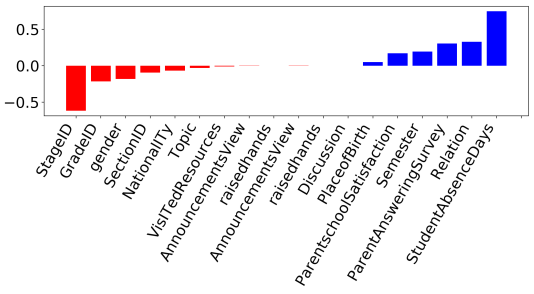
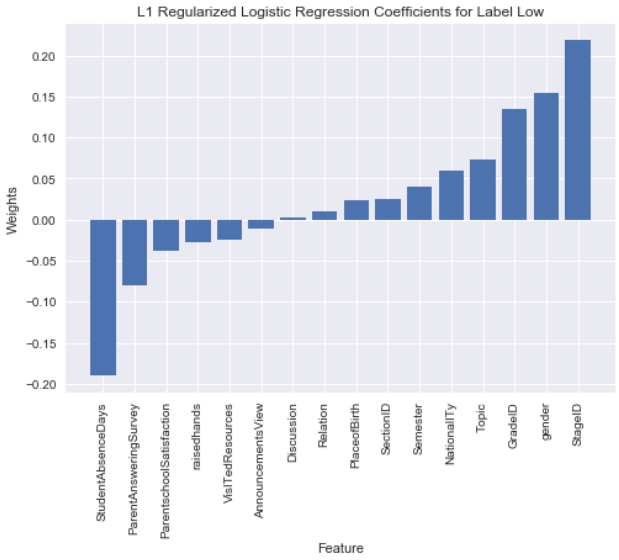
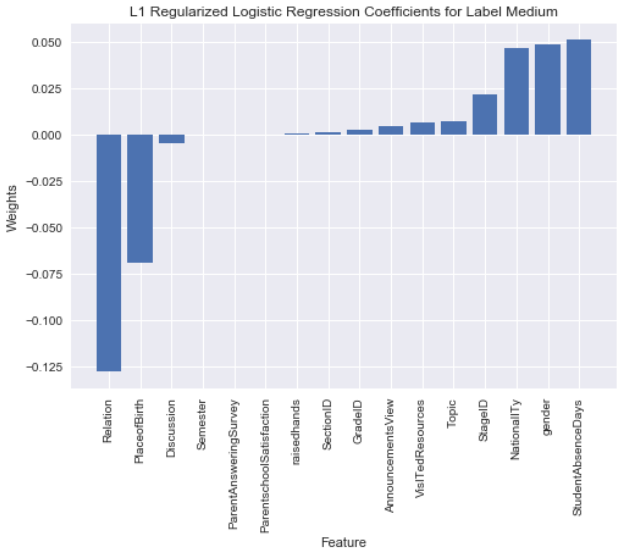


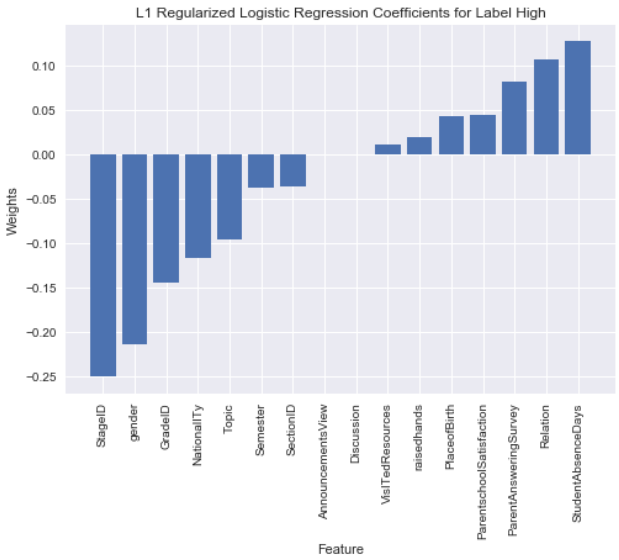
Figure 9. Bar graphs of important features for each label (low, medium, and high level students) from SVM. The highest weighted blue features represent the most important features for a positive label value, while the largest red features are the most important features that suggest a label that is anything but the positive label value.

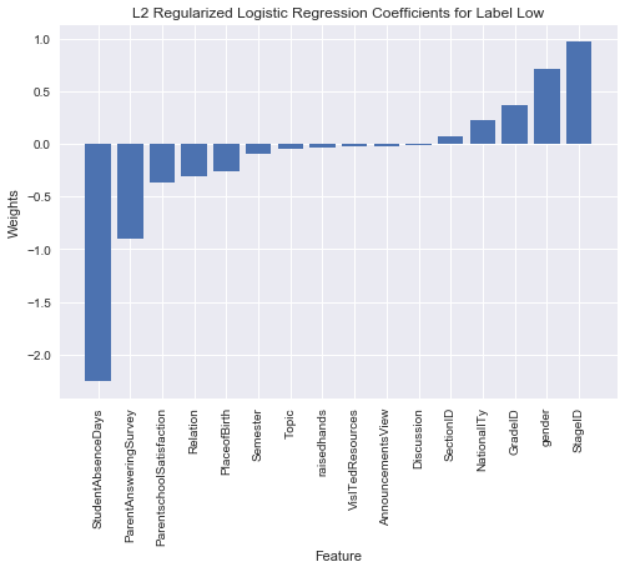
The graph for the SVM run for the low level student indicates that StageID, Gender, and Semester highly correlate with a positive label for a student being low level. StudentAbsenceDays, ParentAnsweringSurvey, and PlaceofBirth are features that correspond with negatively classifying a low level student; i.e. they are the features that pertain to a student not being low level. The features chosen for the medium level are much different, as StudentAbsenceDays, ParentAnsweringSurvey, and SectionID highly pertain to positively classifying the example as medium. It was found that Relation, StageID, and Gender correlated with classifying a student as something other than a medium level student. Lastly, a high level student can be determined by the StudentAbsenceDays, Relation, and ParentAnsweringSurvey features, while other labels besides high can be identified with the StageID, GradeID, and Gender features. This makes sense, because for example, if the high label is negatively classified with Gender and StageID (meaning the student is instead medium or low), the most predictive features for low are StageID and Gender, so the student is likely a low level student. Similarly, the medium level is not predicted with features of Stage ID and Gender, which further confirms that the student is probably low level and that the feature classifications are interconnected across the different labels.

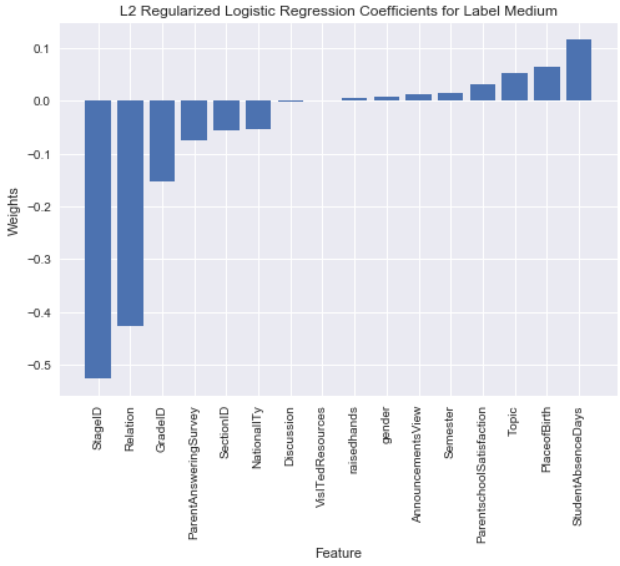
The L1 Regularized Logistic Regression Model was run for each multinomial class label. Since L1 regularization with a SAGA solver "minimizes the residual sum of squares with the constraint that the sum of the absolute value of the coefficients being less than a constant", the nonzero weight vectors that the model calculates are the important features in the model.[8] The R^2 value of the L1 regularized model was 0.61, meaning that 61% of the variation in Class can be explained by the variation in the model. Although the most important features cannot be extracted from the model, we know which features were essentially removed from the model. From figure 10, we can see that for the label "low", the features AnnouncementsView, Discussion, and Relation had coefficients of virtually 0; for the label "medium", Discussion, Semester, ParentAnsweringSurvey, ParentschoolSatisfaction, raisedhands, SectionID, GradeID, and AnnouncementsView all had coefficients of virtually 0; and for the label "high", AnnouncementsView, Discussion, and VisITedResources had coefficients of virtually 0.











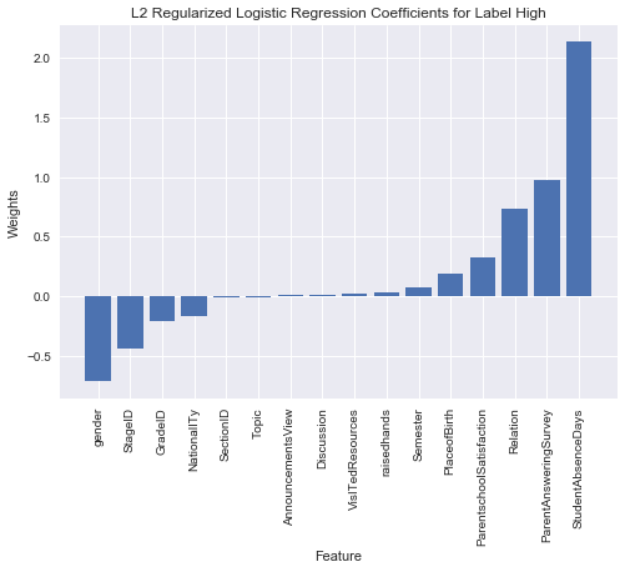


Figure 10. The features in the middle of the bar graphs have weights of virtually 0. These features essentially have no effect on the predictive power of the model and thus the L1 Regularization implicitly performs dimensionality reduction.

The L2 Regularized Logistic Regression Model was also run for each multinomial class label. Since L2 regularization with a Newton-CG Augmented Lagrangian Method Solver introduces a bias factor to minimize the sum of the squared residuals, it alleviates multicollinearity amongst predictor variables in a model. The R^2 value of the L2 regularized model was 0.77, meaning that 77% of the variation in Class can be explained by the variation in the model. Again, although the most important features cannot be extracted from the model, by alleviating multicollinearity, we have weights for each feature that can more accurately fit new data.

Figure 11. Because of the L2 Regularizer, it is unclear whether or not a feature is necessary to be included in the model. However, it can be seen that some features are highly correlated with others and thus have a higher weight to help account for this multicollinearity and eliminate its redundant effect in the model.

For Principal Component Analysis, the PCA model was able to map the 16-dimensional data into just 4 dimensions, although the data is not distinguishable between the different class labels (Fig. 12). The PCA model from the SciKit Learn API didn't return an R^2 value, but rather the average log-likelihood of all samples. The average log-likelihood of our dataset was -40.74. By examining the Eigenvectors and Eigenvalues, PCA found StudentAbsenceDays, VisITedResources, GradeID, and Semester to be the features with the most predictive power.

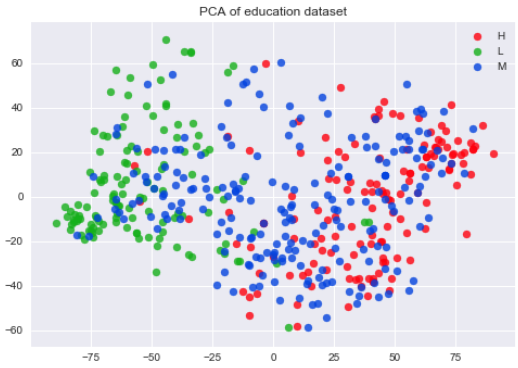


Figure 12. PCA linearly transformed 16-dimensional data into 4 dimensions, but could not distinctly separate each class label.

For Linear Discriminant Analysis, although the model was not able to linearly transform the data to distinguish between different class labels (Fig. 13), the model had a 0.729 R^2 value. The LDA model explains almost 73% of the variability in the response variable Class. By examining the Eigenvectors and Eigenvalues, LDA found that gender, Semester, SectionID, and StudentAbsenceDays were the features with the most predictive power.

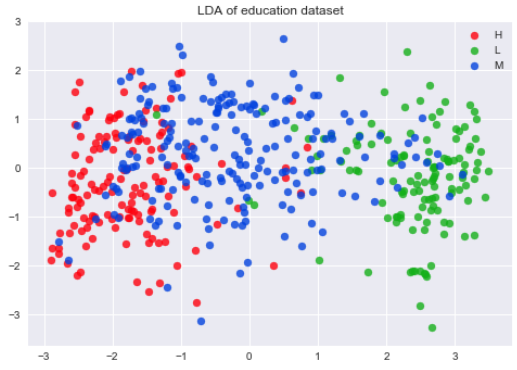


Figure 13. LDA linearly transformed 16-dimensional data into 4 dimensions, but also could not distinctly separate each class label.

4.3. Discussion

Based on the R^2 values of all the models except for PCA (which has a log likelihood value that is not comparable to R^2), the Bayesian network (81.64%) and the L2 regularized logistic regression (77%) had the most predictive power that allowed them to identify feature importance. While Haury et al. did not evaluate Bayesian networks, they found that Lasso regression (L1 logistic regression) was one of the highest accuracy models for feature selection. Although we did not evaluate accuracy of feature selection using precision and recall as did Haury et al., Lasso regression having a high accuracy does correlate to satisfactory model fitting to the data and predictive power, which does not match our result for L1 (we identified L1 with the lowest R^2 value .[5] The following models round out the ranking of R^2 values: SVM (74% for high, 82% for medium, and 56% for low), LDA (73%), and L1 logistic regression (61%). We can analyze model fit and predictive power in order to identify important features based on R^2 values because the higher the value, the more likely the variation in the class label can be explained by the variation in the model.

We hypothesized that the Bayesian network and L1 logistic regression would be the best fitting/most predictive feature selection methods, which was only confirmed for the Bayesian network by the R^2 evaluation method above. However, we believed that LDA would be the least predictive/unable to identify important features, when L1 logistic regression ended up having the smallest R^2 value, most likely due to overfitting. Moreover, we predicted that SVM would neither be the most or least accurate at fitting the data, which was shown through our R^2 analysis as well. Although SVM was run three separate times for each

1-vs-all class label and the other models performed feature selection once on the entire dataset, the average R^2 value (71%) is still higher than that of L1 logistic regression, meaning that SVM was a better fit model for the data.

Analysis can also be performed on what important features were chosen between Bayesian networks, PCA, and LDA. The Bayesian network achieved the highest r-values (ranking values) for the first four important features when using information gain to learn the edges, and as such, we will use these features as the official results for the Bayesian network. VisitedResources, StudentAbsenceDays, RaisedHands, AnnouncementsView were the highest ranking features according to the Bayesian network. PCA found that StudentAbsenceDays, VisitedResources, GradeID and Semester were among the top features in the dataset. LDA predicted Gender, Semester, SectionID, and StudentAbsenceDays for the important features. While there are differences between the features selected, we can say that StudentAbsenceDays has a large impact on whether a student is a low, medium, or high level student because each method selects that feature. The above analysis cannot be done with SVM because the model was run three times for each class label, and thus, the features predicted are based on each class label, not the dataset as a whole. L1 and L2 logistic regression does not allow feature extraction from the model as it performs dimensionality reduction, but can identify which features were essentially removed from the model because they were not important to the predictive power of the model.

Despite the fact the SVM performs feature importance algorithms and L1/L2 logistic regression performs dimensionality reduction, we can compare the least important features from SVM and from L1/L2 logistic regression. For the "low" label, it was found that Discussion, Topic, VisitedResources, and RaisedHands had the lowest weights and were therefore least important in determining the "low" label for SVM. L1 has coefficients of virtually zero (meaning they are not important features for the model) for AnnouncementsView, Discussion, and Relation. L2 has coefficients of around zero for Discussion and AnnouncementsView. From this, we can infer that Discussion is not an important feature for any of the models and must not be important to the predictive power of the model. Similarly, for the "medium" label, RaisedHands, Discussion, AnnouncementView, and ParentSchoolSatisfaction were low weights for SVM. L1 logistic regression also included RaisedHands, AnnouncementView, and ParentSchoolSatisfaction as unimportant features, while L2 listed only Discussion and VisitedResources as unimportant. Lastly, the "high" label had similarities across the three models of AnnouncementView and Discussion being unimportant labels for determining the "high" label for a given student.

5. Conclusion

Feature importance and dimensionality reduction are necessary preprocessing/algorithmic techniques for large datasets that suffer from the curse of dimensionality. By projecting a dataset onto a lower dimensional space or by removing the irrelevant features from a dataset, the interpretability and simplicity of the dataset increases. Additionally, it is easier to identify the predictive power of the dataset without noise and redundancy. We found that L2 regularized logistic regression and Bayesian networks best fit the data, resulting in more predictive power that allowed them to identify important features. L1 logistic regression had the smallest R^2 value, meaning it did not fit the data as well and as such, did not have as much predictive power. This is probably due to the fact that it cannot implicitly perform feature selection.

Future experiments should evaluate feature importance with a wider-breadth of algorithms, such as with ensemble models (i.e. Random Forest) and on more than one dataset to account for variance in noise, feature dependencies, redundancy, etc. Additionally, using datasets with more features may help to better visualize the results of dimensionality reduction and feature selection techniques. Similarly, choosing all feature selection or all dimensionality reduction models may be helpful to better be able to compare the results of the experiments, whether it be by comparing the important features or the removed features.

Acknowledgments

We would like to thank Professor Ameet Soni for graciously answering our questions and providing feedback throughout our project.

References

(1) Amrieh, E. A., Hamtini, T., Aljarah, I. (2016). Mining Educational Data to Predict Students academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*, 9(8), 119-136. https://www.kaggle.com/aljarah/xAPI-Edu-Data.

(2) Amrieh, E. A., Hamtini, T., Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. *In Applied Electrical Engineering and Computing Technologies* (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE. https://www.kaggle.com/aljarah/xAPI-Edu-Data.

(3) Chizi, Barak., Maimon, Oded. Dimensionality and Feature Selection. *Data Mining and Knowledge Discovery Handbook*. 2005. 93-111.

(4) Hira, Zena M., and Duncan F. Gillies. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics*, Hindawi, 11 June 2015.

(5) Haury, Anne-Claire, et al. The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. *PLoS ONE*, vol. 6, no. 12, 2011, doi:10.1371/journal.pone.0028210.

(6) Hoerl, Arthur E., and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, vol. 42, no. 1, 2000, pp. 8086. JSTOR, JSTOR.

(7) Differences between L1 and L2 as Loss Function and Regularization. Choika, 18 Dec. 2013, www.chioka.in/differences-between-l1-and-l2-as-loss-function-and-regularization/.

(8) Tibshirani, Robert. Regression Shrinkage and Selection via the Lasso: a Retrospective. *Journal of the Royal Statistical Society*. Series B (Statistical Methodology), vol. 73, no. 3, 2011, pp. 273282. JSTOR, JSTOR.

(9) Lee, S., Honglak, L., Abbeel, P., and Ng, Andrew Y. Efficient L1 Regularized Logistic Regression. *In the Association for the Advancement of Artificial Intelligence* (AAAI). 2006.

(10) Hearst, Marti A. Support Vector Machines. *IEE Intelligent Systems*. 1998. 18-28.

(11) Bakharia, Aneesha. *Visualizing Top Features in Linear SVM with SciKit Learn and Matplotlib*. Medium. 2016. https://medium.com/@aneesha/visualising-top-features-in-linear-svm-with-scikit-learn-and-matplotlib-3454ab18a14d.

(12) Murphy, Kevin. A Brief Introduction to Graphical Models and Bayesian Networks. 1998.

(13) Raschka, Sebastian. *Principal Component Analysis*. Sebastian Raschka's Website, 27 Jan. 2015, sebastianraschka.com/Articles/2015_pca_in_3_steps.html#pca-vs-lda.

(14) *Loadings vs. Eigenvectors in PCA: When to Use One or Another?* Cross Validated, Stack Exchange Inc., 2017, stats.stackexchange.com/questions/143905/loadings-vs-eigenvectors-in-pca-when-to-use-one-or-another.