DS 501 Case Study 4 Report

Wednesday December 6, 2023

Employee's Work Life Balance Data Analysis

Table of Contents:

Overview

Case Study 4 required the students to perform a data analysis task on their preferred dataset as the final project. The lifestyle and wellbeing dataset from Kaggle was selected as our dataset. This dataset contained about 16,000 survey responses with 24 attributes that described the factors affecting how people live. The work-life balance survey assesses how each individual thrives in both their personal and professional life and as a result, how well your lifestyle and life habits are shaped in accordance to maximizing overall life satisfaction. This report outlines the work done by our group to comprehend, assess, and analyze the dataset.

Problem 1) The Business Part

The key element for this dataset was the work-life balance score (WLBS). This score was calculated based on 24 attributes like fruits and veggies consumption, daily stress level, how often you shout at someone, how many close friends do you have, etc. As a data science group we proposed a project that we deemed valuable to our company to do well in the market. As a company, we were looking to examine our employees' work life balances. We believe that maintaining a good work life balance is important to the work environment and would make employees more productive, increasing our profit. The task in this data analysis is to examine the work-life balance of the employees which plays an important role in the employee's satisfaction and as a result, the quality of their life.

Motivations

We believe this problem is important to be solved as maintaining work-life balance at the optimum level has a positive correlation with higher levels of employee engagement, which leads to an increase in profit and productivity. Also, a good work-life balance affects the well-being and satisfaction of the workers and can also improve their mental health. By solving this problem, organizations can make informed decisions about their employees' support and well being. Nowadays, offering a healthy and happy work environment is a crucial factor in the competitive job market that the organizations consider for attracting the top talents. All in all, addressing work-life balance issues is not just about meeting the needs of employees, it is a crucial factor in the overall success of an organization.

Regression models were considered as we were attempting to predict their life balance score. We experimented with different regression approaches like Lasso Regression and Regression Trees. Lasso regression is useful in handling collinearity among attributes and regression trees help us with the interpretation of the relationships between various factors affecting work-life balance score. We also wanted to consider clustering to see how the clusters differ based on attributes. By implementing a clustering method, we hoped to identify distinct groups of employees with similar patterns. With these models, our business will be able to decide if they need to take action to improve their employees' lives.

Since there are many factors affecting the work-life balance score, we first categorized each attribute. Our approach for making the division could make a difference but we ultimately divided the attributes into 5 categories: productivity, social life, financial security, wellbeing (physical health and mental health), and demographics. Then we allocated the attributes we had into the categories below:

Category	Attribute	Range	Meaning
Social Life	Places visited	0 - 10	Number of new places visited
	Core circle	0 - 10	Number of close friends
	Supporting others	0 - 10	Number of people you help to achieve a better life

	Social network	0 - 10	Number of people you interact with during the day
Social Life/financial Security	Donation	0 - 5	How often do you donate your time or money to a good cause?
Financial Security	Sufficient income	1 - 2	How sufficient is your income to cover life expenses?
Productivity	To-Do completed	0 - 10	How well do you complete weekly to-do lists?
	Flow	0 - 10	How many hours in a typical day do you experience flow?
	Personal Awards	0 - 10	How many recognitions have you received in your life?
	Achievement	0 - 10	Number of remarkable achievements are you proud of
Wellbeing (Physical)	BMI	1 - 2	Body mass index range
	Fruits and Veggies	0 - 5	Number of fruits and veggies eaten per day
	Daily Steps	1 -10	Number of step (in thousands) taken per day
	Sleep hours	1 - 10	Number of hours you sleep
	Lost vacation	0 - 10	Number of days of vacation typically lost per year
Wellbeing (Mental)	Daily Shouting	0 - 10	How often do you shout or sulk at someone?
	Life Vision	0 -10	How many years ahead is your life vision clear for?
	Daily Stress	0 - 5	How much stress do you experience per day?
	Time for Passion	0 - 10	How many hours per day do

			you spend doing something you are passionate about?
	Weekly Meditation	0 - 10	How many times per week do you meditate?
Demographics	Age	Age range	
	Gender Score	Male/Female	

We grouped together the features and explored the groups' effects on work life balance. We divided the features by categorizing them into 3 groups which were: personal life, work life, and both personal and work life. We then added each feature that affects these categories:

- Personal Life: Fruits and Veggies, Places Visited, Core Circle, Donation, BMI, Daily Steps, Life Vision, Sleep Hours, Lost Vacation, Personal Awards, Weekly Meditation
- Work Life: Flow, Sufficient Income
- Both Personal and Work Life: Daily Stress, Supporting Others, Social Network, Achievement, To-Do Completed, Daily Shouting, Time for Passion

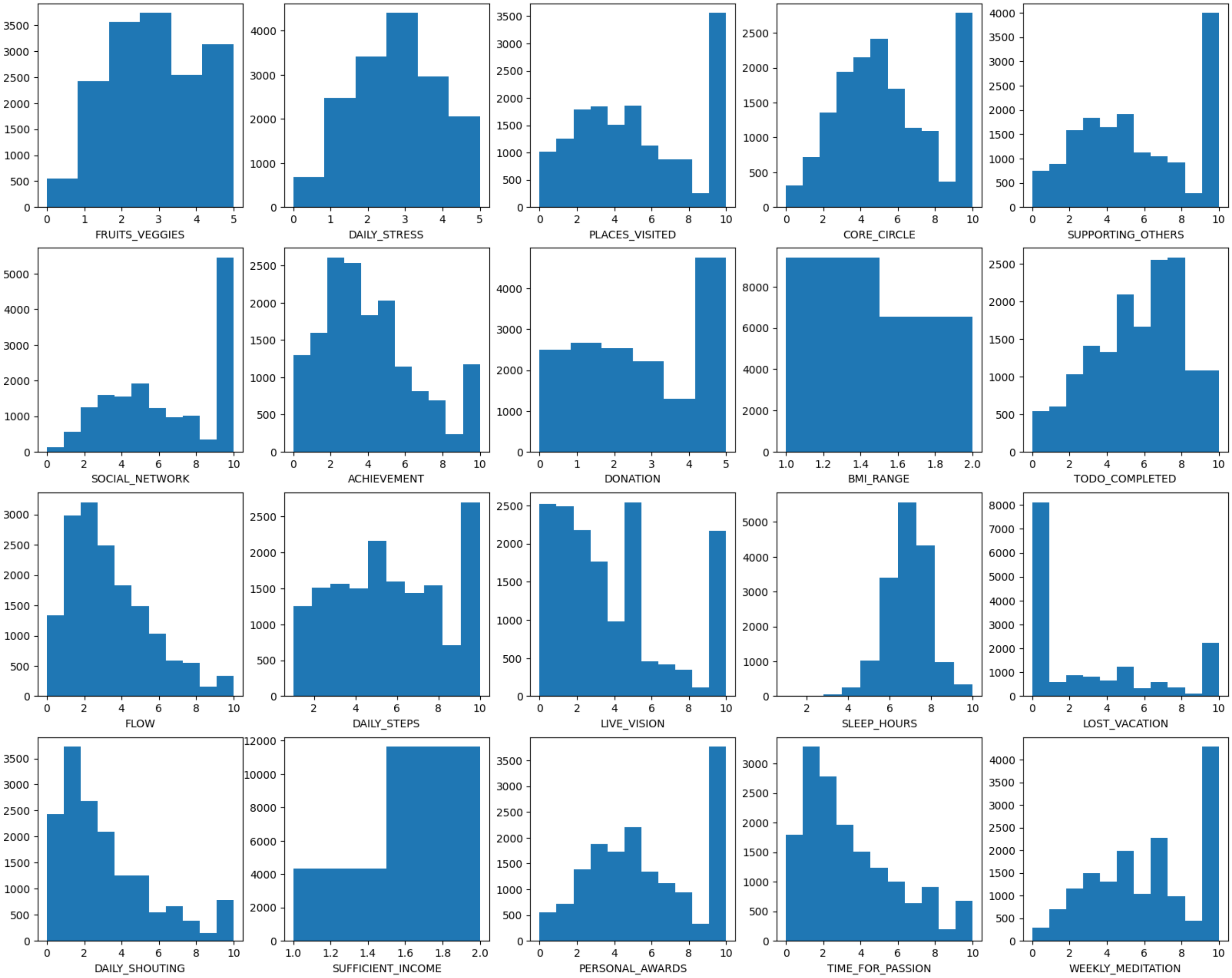
Also, we wanted to explore demographics of people with higher versus lower work life balance scores.

By implementing different categories we could define a structured way to explore our data. We think the proposed idea deserves the financial resources of the company based on different considerations. Firstly, determining the work-life balance score and trying to improve it leads to more productivity and participation from the employees. By having satisfied employees, we can have more positive contributions in the path of company improvement. We believe prioritizing the well-being of employees has a direct impact on the organization's success. This investment can place our company as a potential choice of employers in the competitive job market. Overall, allocating financial resources to this idea shows the commitment of a company to employee satisfaction, success, and sustainability.

Problem 2) The Data Analysis

For the data analysis part, we started with data cleaning and preprocessing. First, the ‘Timestamp’ column was dropped from the dataframe, then the values of ‘Daily Stress’ column was replaced and converted to ‘int64’ data type. Then we tried to collect a list of all the numeric columns in order to plot the histograms below.

Figure 1: Attribute Distributions



In the histograms we notice that the counts spike toward the end of the scales. This makes sense as people responding to surveys usually give more extreme reactions.

Then we created boxplots for the work-life balance score grouped by other features in the dataframe. Overlap can be seen between different groups but a general trend can be observed.

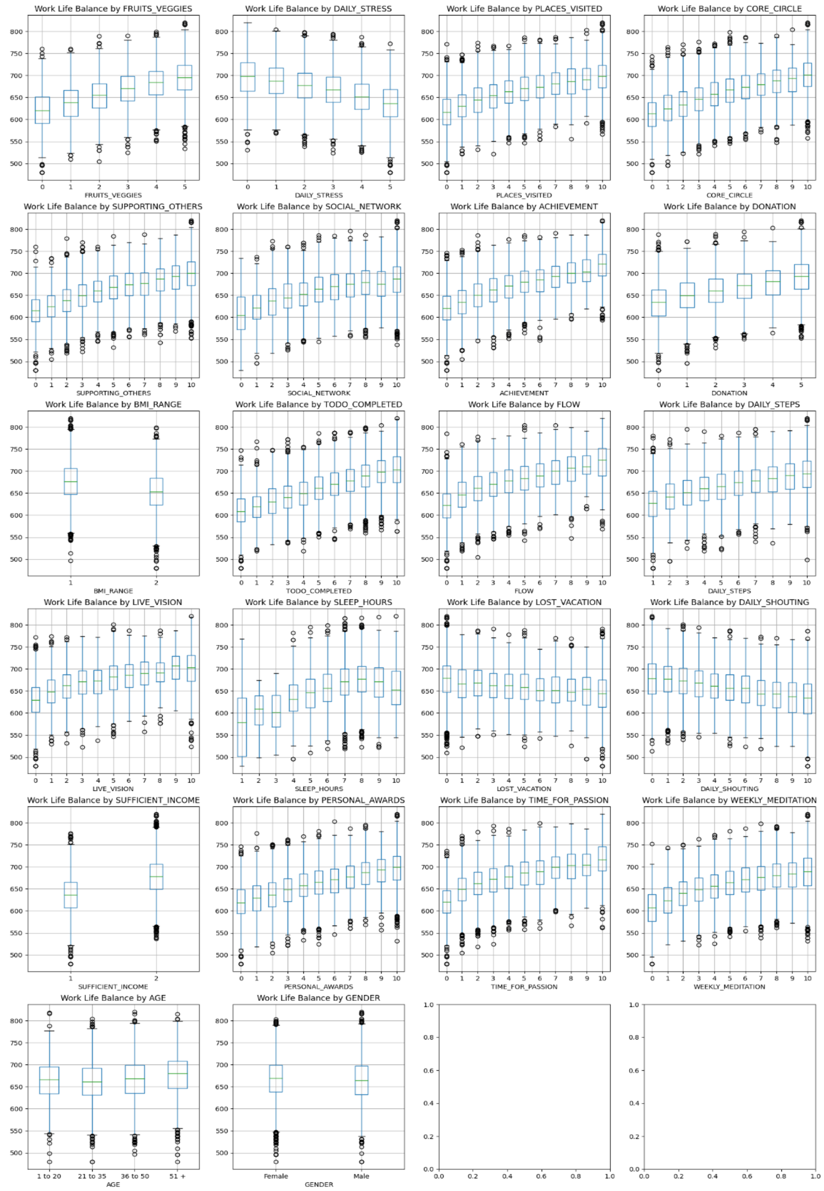


Figure 2: Box Plots

The histogram of the work-life balance score was also created and showed a normal distribution with the center around 650. In this context, existence of normal distribution provides insights into how the work-life balance scores are distributed across the population or sample.

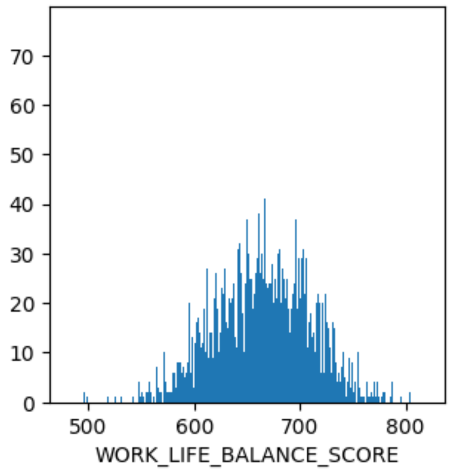


Figure 3: Work-life Balance Score Histogram

Then, we furthered our investigation by plotting bar graphs of the work-life balance score across different features based on age and gender. As the number of the attributes affecting the work-life balance score was high, only a part of the graphs have been shown in the report. Some conjectures that were made in this part are as follows:

- Increased social activity leads to higher work-life balance scores across both genders and all age groups
- In general higher levels of productivity leads to higher work-life balance scores across both genders and all age groups. This could also be a measure of a person's purpose (willingness to complete personal or corporate responsibilities)

These mentioned conjectures can be justified by the following plots:

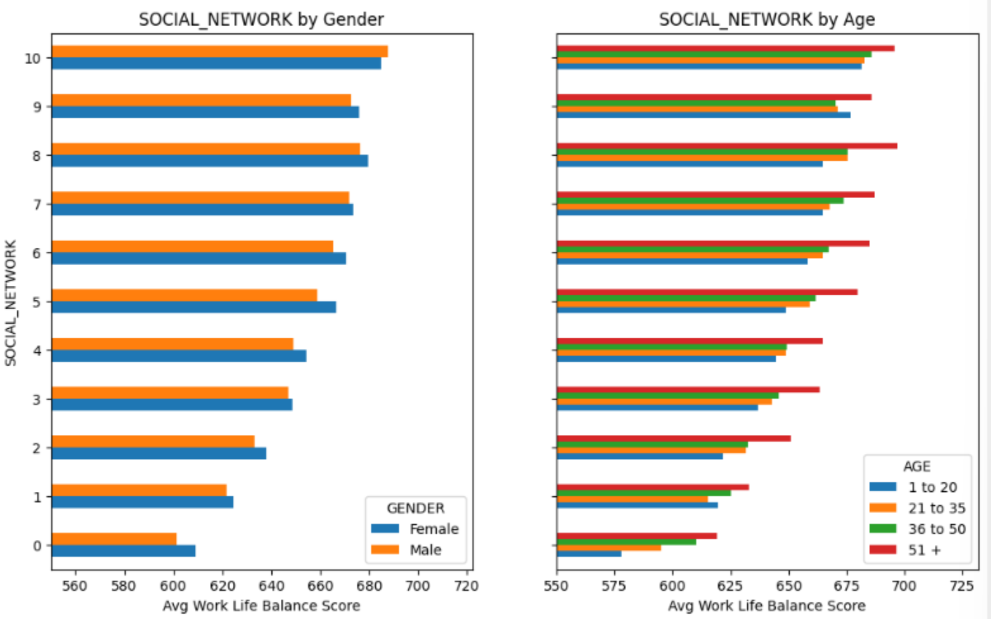


Figure 4: Social Network Breakdown

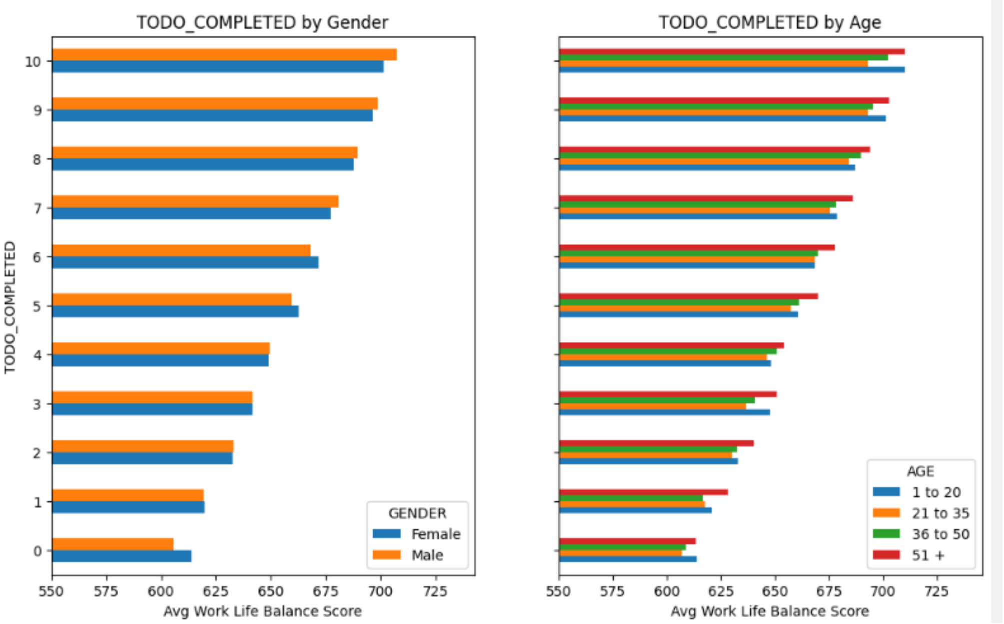


Figure 5: TO-DO Completed Breakdown

As personal life attributes had the major impact on the work-life balance score, we selected the personal life attributes with a strong correlation and plotted their corresponding bar graphs. Below is the result of some of the graphs plotted:

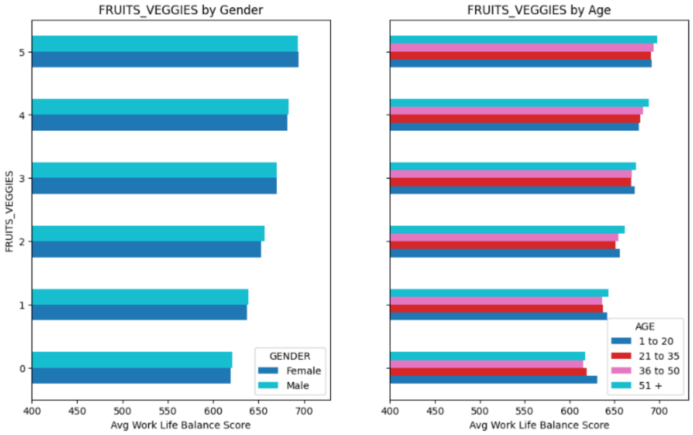


Figure 6: Fruits and Veggies Breakdown

Fruits and Veggies

- No large gender difference, although a clear upward trend for WLBS for both.
- Upward trend generally across all ages, although younger ages may see a bigger impact.

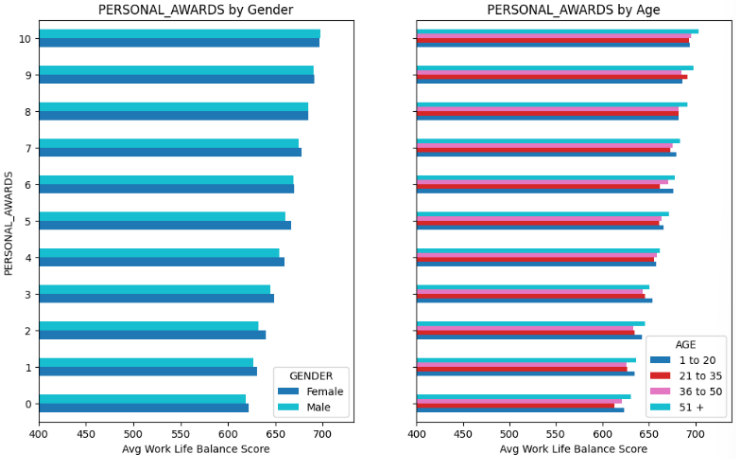


Figure 7: Personal Awards Breakdown

Personal Awards

- At fewer awards, females seem to outscore males, but this trend disappears at a higher number of personal awards.
- A lower number of awards may matter more for people 21-35 and 36-50.

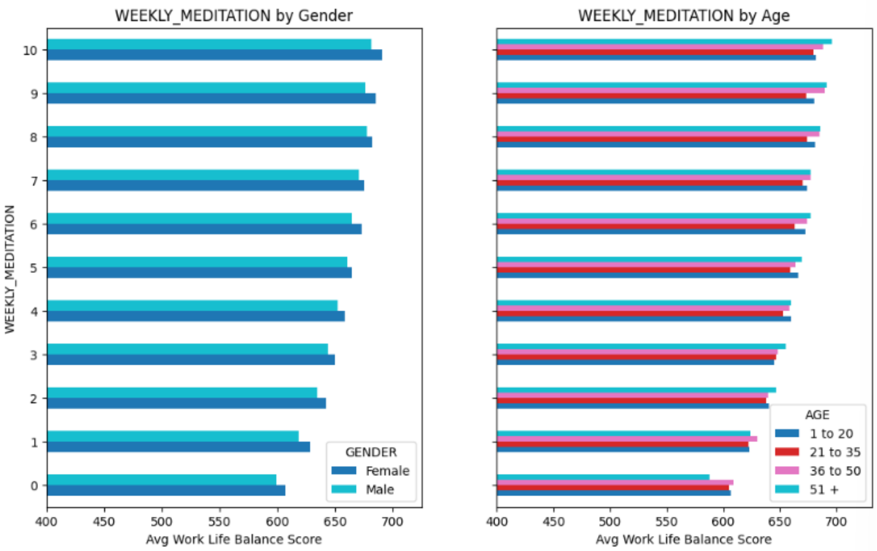


Figure 8: Weekly meditation Breakdown

Weekly Meditation

- Females generally score higher than males no matter how much weekly meditation they do.
- Lower amounts of weekly mediation may negatively impact people 51 + more than other groups.

Then the correlation matrix was calculated for the numerical columns and created a heatmap plot of the correlation matrix. Based on the following plot, among personal life attributes, BMI range, lost vacation and daily shouting have negative correlation on the work-life balance score and among work life attributes, both daily stress and daily shouting have a weak negative effect on the work-life balance score. For positive correlation, weekly meditation, fruits and veggies, daily steps, daily shouting, places visited, core circle, donation, daily steps, life vision, sleep hours and personal awards are among the personal life attributes and supporting others, social network, achievement,to do completed, daily flow, sufficient income and time for passion as work life attributes.

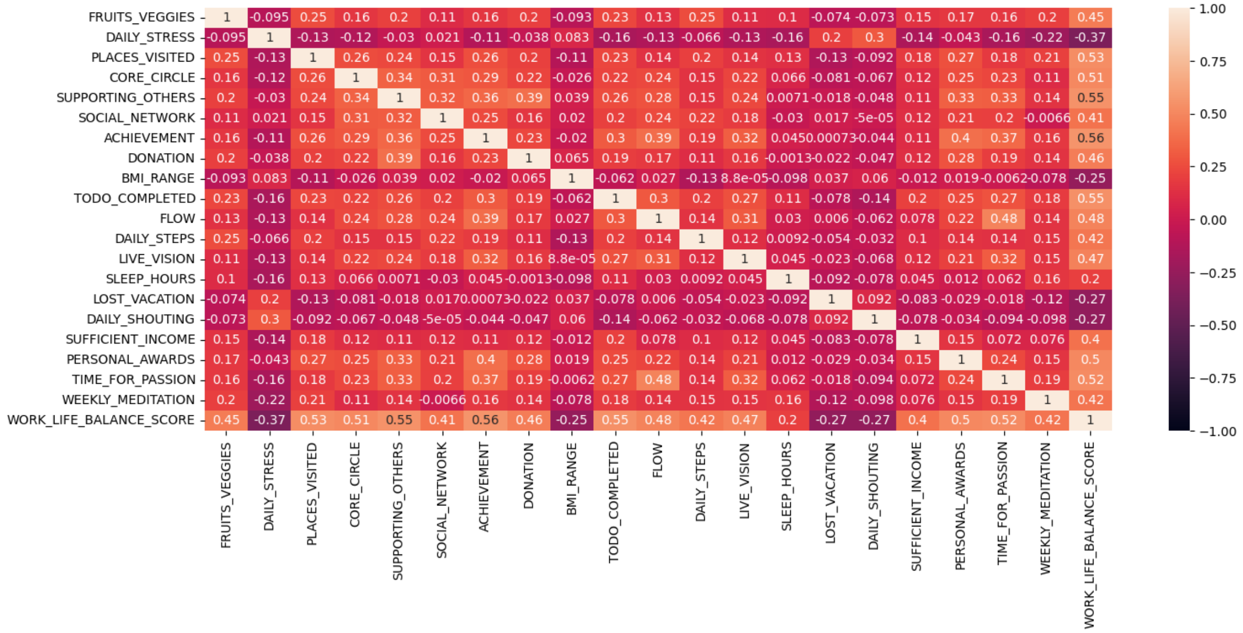


Figure 9: Correlation Matrix

Problem 3) Model Development

The basis of our model development was that some features, such as social aspects of work, have a larger impact on the WLBS than others. We also found that the demographics of our employees do not impact their score, so it would be appropriate to make models that apply to all our employees. Using these two conjectures, our model would be able to accurately predict the score for an employee, and provide the features that are most important.

Regression Tree

As the WLBS is a continuous variable that the team wanted to predict, a Decision Tree Regression model was selected. This type of model was chosen due to its ease of implementation and its interpretability. This model was trained using all available features as the model would recursively divide the feature space based on the most influential features.

As a prerequisite to training, the dataset was divided into a training set used to train the model and a test set for evaluation. The “max_depth” hyperparameter was initially set to 5 to gain an understanding of how well this model could perform. With an initial model created, the team decided to tune the model and improve its performance. To accomplish this, the model was tuned using “GridSearchCV” and 5-fold cross validation. The team performed the grid search for tree depths that ranged from 3 to 20 and plotted the cross validation error. The model that produced the lowest cross validation error was then selected and trained.

Lasso Regression

Another model trained by the team was Lasso regression, which was also appropriate for predicting WLBS. This model is able to reduce the dimension by performing feature selection. Increasing the regularization term, alpha, leads some of the coefficients to become zero. In other words, Lasso completely eliminates some features, leading to the reduction in dimension. In addition to reducing the model complexity, which reduces the chance of overfitting, simplicity and interpretability are the other factors that make Lasso an appropriate candidate for the model selection.

The process of training this model can be described as follows: First, the features were scaled using StandardScaler. Hence, the features are standardized by removing the mean and scaling to unit variance. Then, the Lasso model was created and trained using the scaled training data. Afterward, hyperparameter tuning was done using

GridSearchCV for finding the optimal value of alpha from a defined range of values. Finally, the performance of the tuned model is evaluated. The metrics that are used for the purpose of model evaluation are Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared values. MSE measures the average squared difference between the estimated values and the actual value. MAE measures the average absolute error. R-squared indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

Clustering

The team attempted to perform clustering on the data in order to better understand groups of employees which may have low, average, and high work-life balance scores; however, the results indicated that the dataset does not contain clearly distinguishable clusters.

To visualize the dataset and to get a notion of how many clusters there could be, TSNE was used. The results of the TSNE visualization are shown in the graph below. These results indicate that while there may be small clusters in the data, there may not be great separation between the clusters. This is also supported by the overlapping box plots of the dataset in the data analysis section of this report, where the relationship between the WLBS and the feature values does not create distinct ranges.

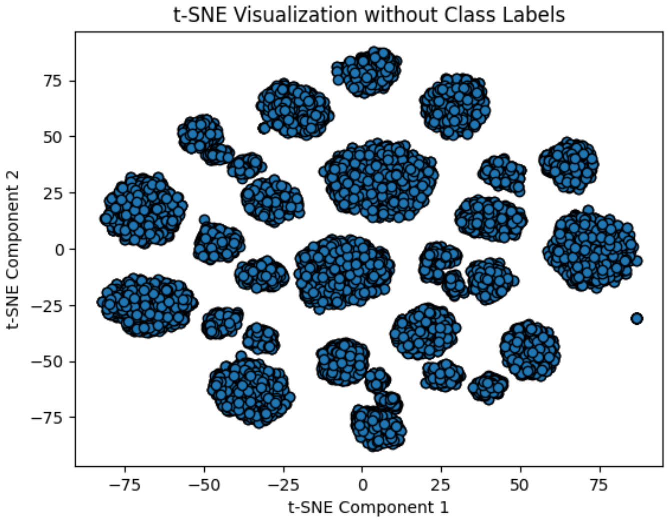


Figure 10: T-SNE Visualization

Additionally, we performed PCA on the dataset to explore potential feature engineering to improve clustering results. The table below shows the explained variance ratios for five components.

1st	2nd	3rd	4th	5th
0.18947886	0.08215014	0.06469823	0.06032714	0.05202449

The first principal component shows an explained variance value of 0.18, which is low for a first component. The explained variance decreases drastically with subsequent components. This indicates that dimensionality reduction through the use of principal components may not be a viable strategy for improving clustering results.

For clustering, we used a range of cluster values and evaluated the results by calculating the silhouette score. First, K-Means clustering was performed on the full dataset. The figure below shows the results of K-means clustering on the full dataset.

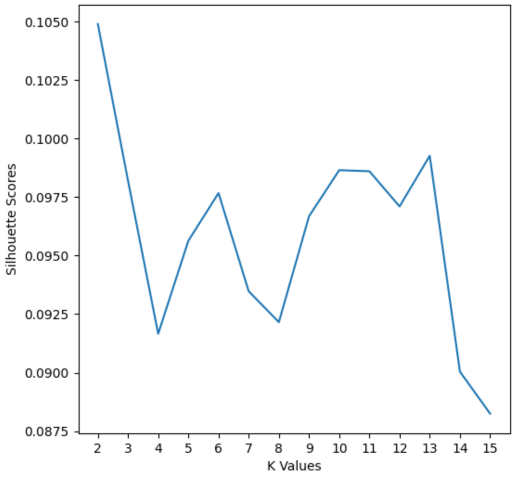


Figure 11: Silhouette score for K-means

With the number of clusters ranging from 2 to 15, the highest silhouette score achieved by K-means was approximately 0.1050 with two clusters. As a silhouette score close to 1 indicates well-formed clusters, these results indicate that the clustering was ineffective.

The second approach used agglomerative clustering and reduced the number of features being clustered in order to improve the results. Only features related to work were kept: supporting others, social network, achievement, to-do completed, time for passion, and sufficient income. Flow, although a work attribute, was removed after early results showed that removing it improved the silhouette score. With this feature set, agglomerative clustering was tuned with different linkage values (ward, single, complete, and average). Average and single linkage showed to perform similarly with maximum recorded silhouette scores of 0.36 and 0.35 respectively. The graph below shows the silhouette scores for agglomerative clustering with average linkage.

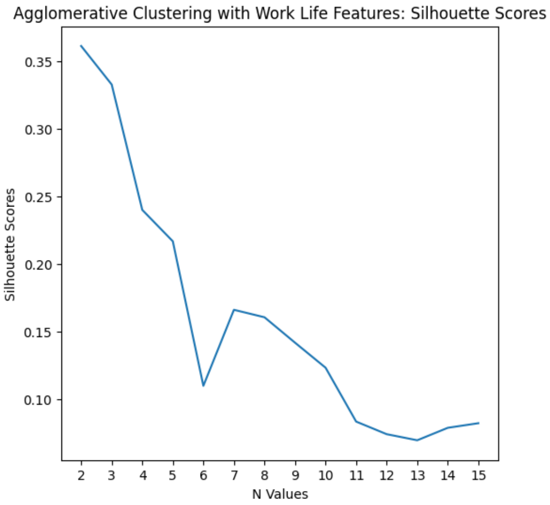


Figure 12: Silhouette score for Agglomerative Clustering

Although a silhouette score of 0.361 is the best result from the clustering experiments, it is still a poor score. It indicates that there are likely many points which may be incorrectly clustered, i.e. that they may be better suited to their nearest neighboring cluster or that they are not similar to the points in their own cluster. Thus, the results from agglomerative clustering were not used in the final recommendations to the company.

Classification

An alternative to the company is using classification instead of regression. This can be done by creating groups of the scores. In our implementation we created 5 groups, each containing 20 percent of the data. We used training, validation, and testing data sets to train, validate, and get a final test of the models. The classification models we created were support vector machine (SVM), random forest, and logistic regression. Each of these models have hyperparameters to train before deciding on the final model. For SVM, we explored the linear, polynomial, and radial basis function kernels to see which one would work best for our data. The linear kernel model performed the best out of these three approaches. For the random forest approach we explored the number of trees and found 300 trees produced the highest accuracy. Moving to logistic regression we tuned the solver hyper parameter and found the saga solver to produce the highest accuracy. With the model hyperparameters set, we trained our final models and applied them in a business setting.

Model Business Application

Regression Tree

After training several regression trees, the initial model with its depth set 5 achieved a mean absolute error of 22.38. On tuning the model with “GridSearchCV”, the cross validation error was plotted, and is displayed below.

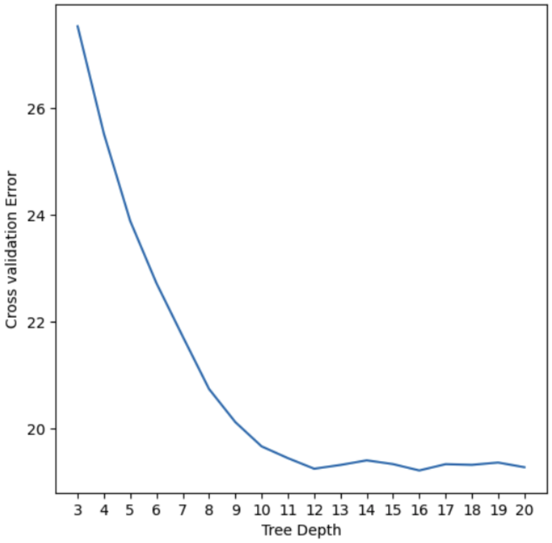
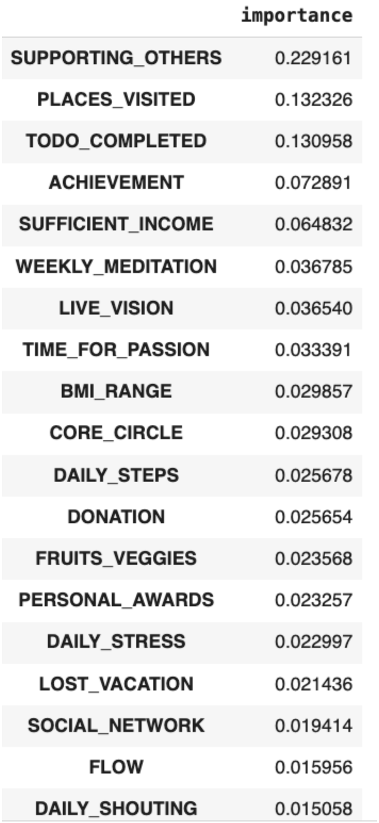


Figure 13: Regression Tree Cross Validation error

The model that achieved the lowest cross validation error occurred at a depth of 16. This model was then retrained achieving an improved mean absolute error of 17.77.

To assist with interpreting the results, the feature importance was also calculated using the "feature_importances_" attribute of sci-kit learn's "DecisionTreeRegressor" class. This attribute returns an array of the normalized total reduction of the criterion brought by each feature.

As mentioned in the Business Proposition section of this report, the team sought to predict an employee's WLBS and improving this score would lead to increased productivity and more profit. The Regression Tree model predicts the WLBS and given the feature importance calculation, it also indicates which features are most important. The company can leverage these more important features and develop policies aimed at improving these aspects of an employee's life. The most important features are listed below:



	importance
SUPPORTING_OTHERS	0.229161
PLACES_VISITED	0.132326
TODO_COMPLETED	0.130958
ACHIEVEMENT	0.072891
SUFFICIENT_INCOME	0.064832
WEEKLY_MEDITATION	0.036785
LIVE_VISION	0.036540
TIME_FOR_PASSION	0.033391
BMI_RANGE	0.029857
CORE_CIRCLE	0.029308
DAILY_STEPS	0.025678
DONATION	0.025654
FRUITS_VEGGIES	0.023568
PERSONAL_AWARDS	0.023257
DAILY_STRESS	0.022997
LOST_VACATION	0.021436
SOCIAL_NETWORK	0.019414
FLOW	0.015956
DAILY_SHOUTING	0.015058

Figure 14: Regression Tree Feature Importance

Supporting Others was shown to be an important feature, so a company could implement a mentorship program or team building exercises. The company could also offer rewards for completing tasks since “TODO_COMPLETED” and “ACHIEVEMENT” were also shown to be important features.

Lasso Regression

After training the Lasso model, the best achieved alpha value is 0.001. Consequently, the obtained values for MSE, MAE, and R-squared are 7.69e-06, 0.002, and 0.99, respectively. Comparing the MAE obtained for Lasso regression to that of decision tree, we can observe that Lasso achieves a lower MAE indicating that this model is a more accurate model. We show the obtained coefficients for different values of alpha and visualize them in Figure 15.

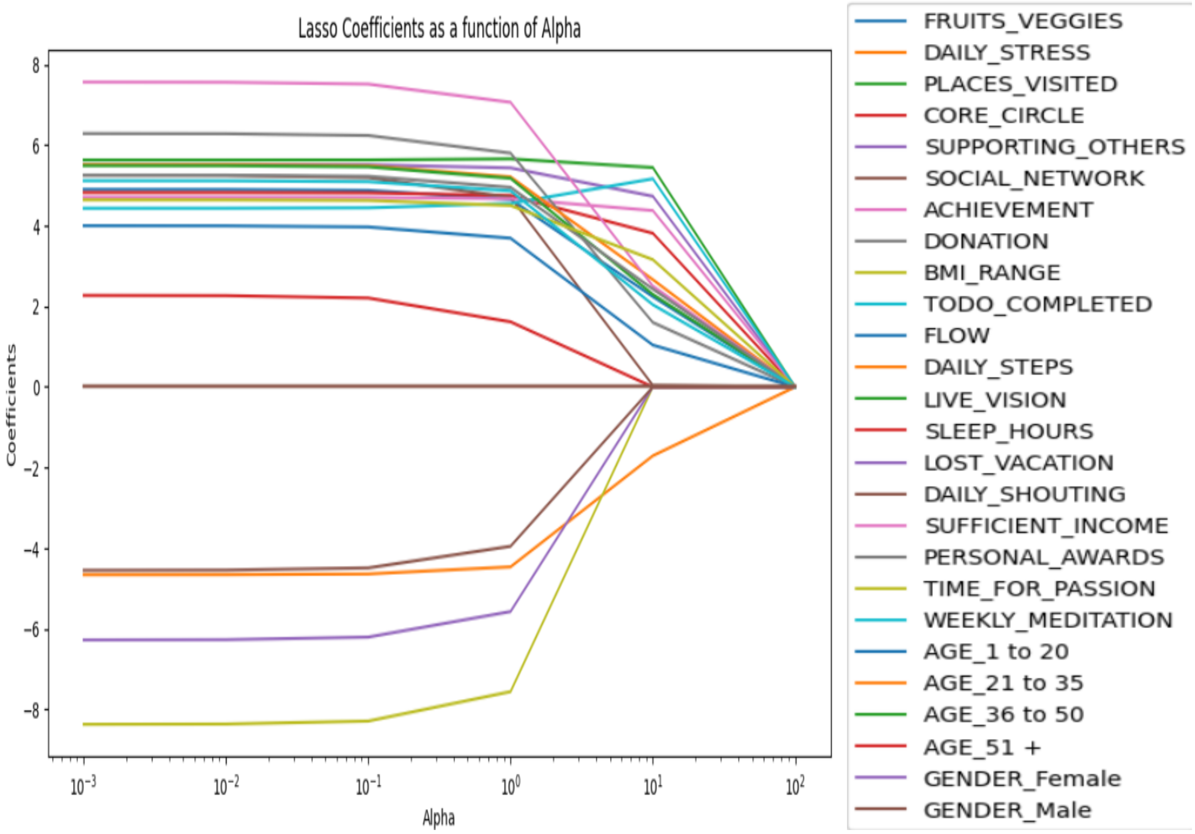
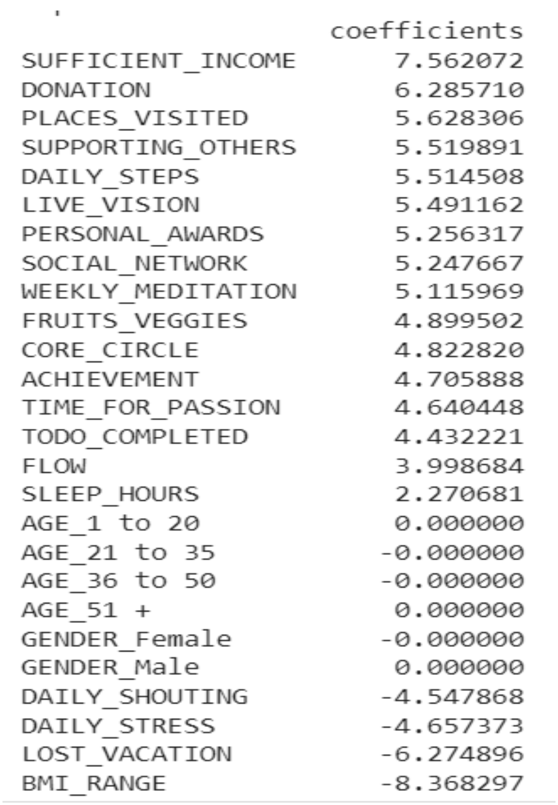


Figure 15: Lasso Coefficients as a Function of Alpha Graph

Figure 15 shows the impact of increasing alpha on dimensional reduction. To be specific, as alpha increases, the number of features decreases and the most important features can be extracted. For example at alpha equal to 100, all the coefficients reach zero. The coefficients of the Lasso with alpha set to 0.001 are shown below.



```
  '
                         coefficients
SUFFICIENT_INCOME            7.562072
DONATION                     6.285710
PLACES_VISITED               5.628306
SUPPORTING_OTHERS            5.519891
DAILY_STEPS                  5.514508
LIVE_VISION                  5.491162
PERSONAL_AWARDS              5.256317
SOCIAL_NETWORK               5.247667
WEEKLY_MEDITATION            5.115969
FRUITS_VEGGIES               4.899502
CORE_CIRCLE                  4.822820
ACHIEVEMENT                  4.705888
TIME_FOR_PASSION             4.640448
TODO_COMPLETED               4.432221
FLOW                         3.998684
SLEEP_HOURS                  2.270681
AGE_1 to 20                  0.000000
AGE_21 to 35                -0.000000
AGE_36 to 50                -0.000000
AGE_51 +                     0.000000
GENDER_Female               -0.000000
GENDER_Male                  0.000000
DAILY_SHOUTING              -4.547868
DAILY_STRESS                -4.657373
LOST_VACATION               -6.274896
BMI_RANGE                   -8.368297
```

Figure 16: Lasso Coefficients at alpha = 0.001

These coefficients quantify the extent to which an employee's WLBS can be improved or reduced. For instance, the coefficient associated with "SUFFICIENT_INCOME" is 7.56. This means that if a company can increase an employee's perception of a sufficient income by 1 point while keeping all other attributes the same, that employee's WLBS will increase by 7.56. We can also see that the weights for age and gender are both 0 meaning that age and gender do not affect an employee's WLBS proving one of our conjectures. A company can leverage this information by not only implementing policies aimed at improving the WLBS, but also quantify any net improvement or reduction of that score.

Classification

Using the hyperparameters determined best fit when using the training and validation data, we produced three different classification models.

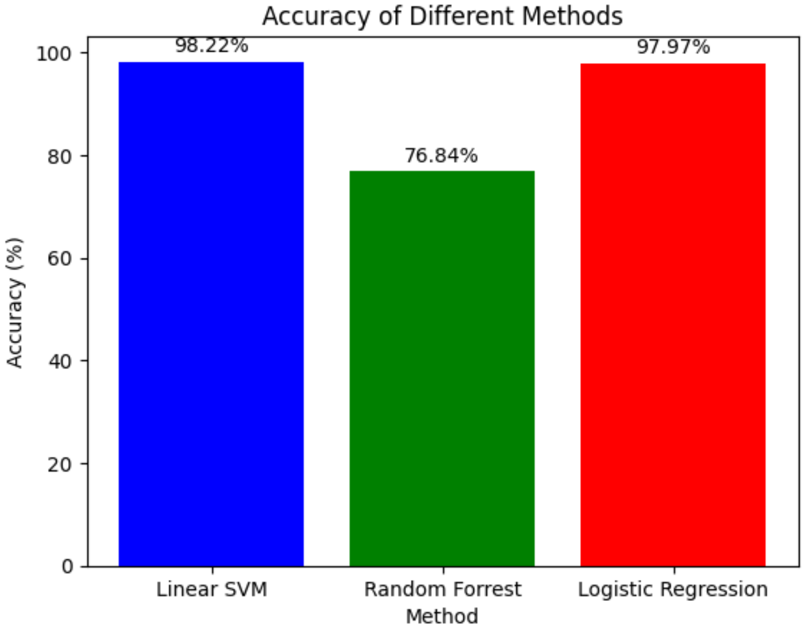
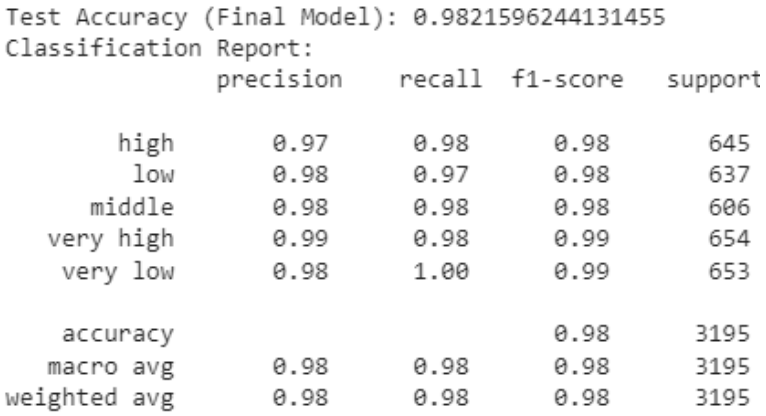


Figure 17: Model Test Accuracies

In this graph we can see random forest performed significantly worse than the other two models.



```
Test Accuracy (Final Model): 0.9821596244131455
Classification Report:
              precision    recall  f1-score   support

        high       0.97      0.98      0.98       645
         low       0.98      0.97      0.98       637
      middle       0.98      0.98      0.98       606
   very high       0.99      0.98      0.99       654
    very low       0.98      1.00      0.99       653

    accuracy                           0.98      3195
   macro avg       0.98      0.98      0.98      3195
weighted avg       0.98      0.98      0.98      3195
```

Figure 18: Linear SVM Classification Report

Taking a closer look at the classification report we see that linear SVM has a better accuracy, and better recall and precision for all the groups by about one percent compared to logistic regression. For this reason we would suggest you to use the linear SVM model to predict what class your employee will fall in. In doing so you can easily identify those employees that are in the lowest class, and in a similar way to the regression tree, create programs and training to assist those employees in creating a better work life balance. While these programs will not directly generate profit, creating a better environment for your employees generates better performance, and hence more profit. If the company is more interested in model interpretation, we would suggest logistic regression as you would be able to examine the coefficients and p-values. Then this model could be applied in a similar context to the Lasso Regression model.

Conclusion

The objective of this case study was to create a model for a company that could be used to predict WLBSs for their employees and to understand the features most important to achieving a high work-life balance. With such a model, a company could implement changes to promote greater work-life balance.

We used regression, clustering, and classification methods to find a model that could be trusted. Although clustering was unsuccessful, the regression and classification models showed strong results.

The decision tree model showed that supporting others, places visited, and to-do completed were the top three most important features. A company can use these results to implement policies such as a mentorship program, which can help increase how much employees support others and are supported by others. Additionally, more productivity tools can be put in place for employees to make it easier for them to complete their to-do lists. With the Lasso regression model, a company can tune their approach to improving employees' WLBSs in a more fine-grained manner. For instance, for an increase in a particular feature, the increase in the employee's work-life balance score can be directly quantified. The important features that influence employee scores can then be paid extra attention.

The classification models also achieved high F1 and accuracy scores, particularly SVM and logistic regression, which indicates they may be helpful in predicting which work-life balance score range an employee belongs to.

In conclusion, the team produced several promising models that a company could use to improve the work life balance of their employees, leading to increased productivity and profits for the company.

Challenges and How We Worked Together

We spent a lot of time brainstorming on how we could apply our ideas into a business setting to generate money. We had a lot of different ideas about how to model the data, but kept the idea of how this can make money in mind. Chad helped the group out by creating histograms and boxplots before our second meeting to discuss our analysis. This aided us in speeding up the process of taking our ideas into implementation. We could visualize what features are interacting with the WLBS and deduce what to do from there. From there we all did our own analysis and conjectures on specific features with a focus on gender and age. We found similar ideas that served as a basis of our models. One of our largest problems that we faced was implementing a successful clustering model. We envisioned it so that it would create distinct groups that we could analyze separately. What we envisioned happening was possibly one group had a low WLBS and other features. Then another group with a low WLBS and different features. This way we would have two groups that have low scores for two different reasons. The company could then make more specific decisions on how to address these low scores. In reality we faced a lot of difficulty in creating distinct groups and had a silhouette score of .3. We had our regression models done, but we thought we could go further, so we transitioned to proposing a classification model so that the company can set its own groups and make decisions off of that. Other than our troubles with clustering, our ideas in regards to regression went smoothly. We envisioned our regression tree to provide feature importance, which it did, and lasso to provide an accurate prediction of an employee's score. Relating these models back to the business was simple as we had decided as a group to pick these models because of their business applications. The user of the models can use them to make programs and decisions to aid their employees' work life balance. Overall, it was a group effort in ensuring we stayed on track from our complex and scattered analytical ideas, into implementation in a business culture to produce profit.

Authorship Table and Contribution of the Members

Problem 1	Chad Hucey, Pegah Emdad
Problem 2	Chad Hucey, Talia Andrews, Noushin Khosravi Largani, Aaron Brady
Problem 3 (Model Development and Conjectures)	Chad Hucey, Pegah Emdad,Talia Andrews, Noushin Khosravi Largani, Aaron Brady
Rough Draft of Report	Chad Hucey, Pegah Emdad,Talia Andrews, Noushin Khosravi Largani, Aaron Brady
Final Report (Proofreading and Editing)	Chad Hucey, Pegah Emdad,Talia Andrews, Noushin Khosravi Largani, Aaron Brady
Presentation Slides	Chad Hucey, Pegah Emdad,Talia Andrews, Noushin Khosravi Largani, Aaron Brady
Group Meeting Arrangements	Talia Andrews, Aaron Brady
Creating and Sharing Required Files and Documents (Google Collab, Google Docs, Google Slides)	Pegah Emdad,Talia Andrews, Aaron Brady