Bruce's Pearls to Success

How Auburn wins and loses

Leader: Cole Morris

Members: April Kruse, Aryan Makhija, Chris Choi, Feagin Tracy, and Josh Leadingham

Introduction

- During the 2021-22 basketball season, Auburn University became a basketball school. Auburn Men's Basketball reached #1 in the AP poll for the first time in history.
- They also had a winning streak of 19 games until Arkansas broke it after going into overtime. Bruce Pearl brought more pride and wisdom to the community than ever before.
- There are multiple factors that contribute to successes and challenges during these games. In this project, we dive into teams in which Auburn has played under Bruce Pearl's leadership.
- We want to explore the variables that contribute to the success of Auburn basketball. This will benefit Bruce Pearl and his players to pinpoint strong and weak areas in the program. Adjusting training in said areas will maximize Auburn's strength and improve performance as well as winnings.
- The data we chose to use are games Auburn has played during seasons 2014-2019 under Bruce Pearl's coaching. We are using teams in the SEC as well as some other teams Auburn has played in the Big 12 and other conferences for a total of 194 games.
- Our objective is to find variables that have the biggest impact on whether Auburn's Men's Basketball team wins or loses.

Data

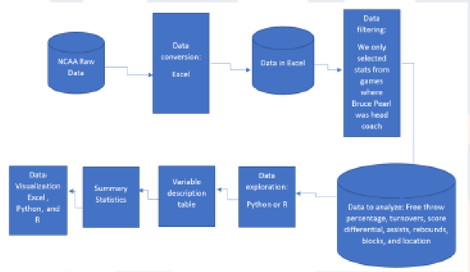
- We decided to have a total of 9 variables.
- Of the variables listed below, Total Score Differential is our Y variable, with all other variables being our X variables.

Variable Name	Variable Definition	Variable Type
Free Throws Percentage	Free throws are known as an unguarded scoring attempt that a referee awards a basketball player after an opposing team member commits a foul against them. Percentage are those made divided by total shots attempted.	Ratio
Free Throws Attempted (FTA)	FTA is the total number of times a team is given the chance to make a free throw.	Ratio
Free Throws Made (FTM)	FTM is the number of times a team has successfully made a free throw attempt.	Ratio
Turnovers	Known as the number of times the offensive team loses the possession of the ball without shooting and scoring.	Ratio
Total Score Differential	Known as the amount scored and amount given up. The location and year could answer whether or not the atmosphere changes team performance.	Interval
Assists	Known as the last pass to a teammate when it directly leads to a basket.	Ratio
Rebounds	Known as gaining possession of the ball following a missed field goal attempt.	Ratio
Blocks	Known as the event in which a player makes contact with an opposition's shot attempt resulting in a missed shot.	Ratio
Location	Where the game is played. Specifically here this means Auburn and Tuscaloosa or Home and Away, respectfully.	Categorical

Methods

- We began with taking the raw NCAA data from the Excel sheets given, and then filtered games and statistics that we needed to keep into a new sheet. We then generated graphs to find comparisons between variables.
- For our methods, we are using Linear Regression and Decision Tree.

- In this project, we will experiment with the score differential, our Y variable. Python is our focus when coding and computing the results for the models, but we also be using some R-Studio.
- We made the decision to keep any outliers since there were very few and there will always be some random moments throughout seasons.
- Shown below is our method flowchart

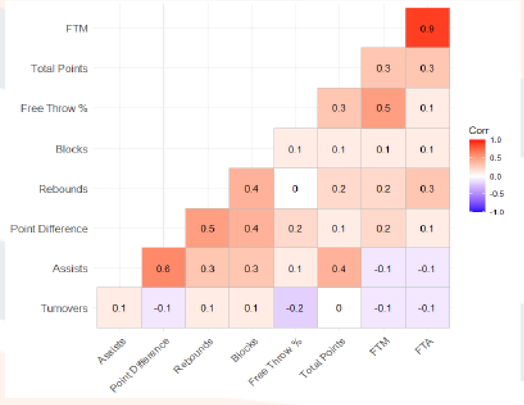


Cross-Validation and Descriptive Analysis:

- Our data was divided 70 vs 30 for training and testing, respectively.
- This was done primarily due to data constraints since each variable needs at least 5-10 data points. We used the same code that we used from developing our models.

```
In [27]: import pandas as pd
         from sklearn.model_selection import train_test_split
         from sklearn.linear_model import LinearRegression
         BBall = pd.read_csv("All Pearl's Games.csv")

In [33]: #split datasets 70/30
         from sklearn.model_selection import train_test_split
         x_names = ["FTM", "Rebounds", "Assists", "Blocks", "Turnovers"]
         X = BBall[x_names]
         y = BBall["Point Difference"]
         X_train, X_test, y_train, y_test = train_test_split\
         (X,y, train_size = 0.7, random_state = 16)
```

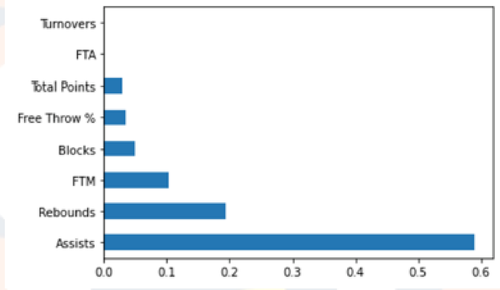


- Above is the heat map we used to look at correlations between each of our variables. As we explored it, we learned that rebounds and assists are the most significant. Therefore, Auburn needs to focus on these areas.

Modeling Results

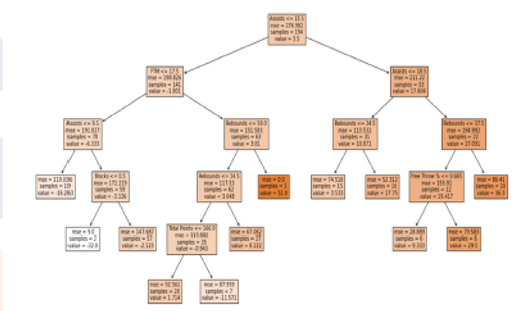
Importance Chart:

- This Chart is a Feature importance chart, which shows how significant the X variables are at accurately predicting the Y value (Score Differential).
- In this model, Assists are shown to be the most important variable given the high importance score followed by rebounds.
- On the other hand, Turnovers and Field Goals Attempted had no impact on score differential. The remainder of variables had very little impact.



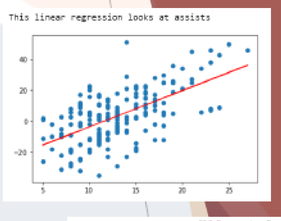
Regression Decision Tree Model:

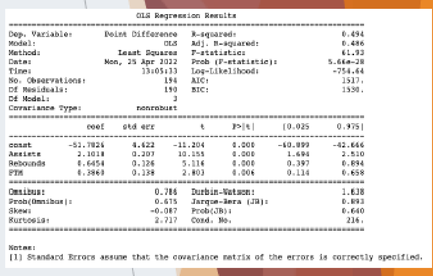
- Pictured below is the decision tree model. In a regression decision tree model, each node contains an X variable and a greater than or equal statement for a variable. Depending on if it is true or false you will navigate to that respected node
- The first node is called a root node, which contains the most significant X variable. Each node that branches off to more nodes are called decision nodes.
- Leaf nodes are the end predictions for the model based on the test data the model is fed. In these leaf nodes the value describes the predicted point differential (a positive value is an Auburn win, and a negative one is a loss).
- Then samples mean how many times in the training set the predicted values were observed. MSE is an estimate of how much error is observed in the model.



More Modeling Results

- According to the linear regressions below, rebounds and assists are the most significant. The correlation is moderately high in both.
- R-squared is relatively high at 0.494 along with the f-statistic of 61.93. Therefore, these variables are proven to be statistically significant to the success of Auburn's performance.





OLS Regression Results

Dep. Variable:	Point Difference	R-squared:	0.494
Model:	OLS	Adj. R-squared:	0.486
Method:	Least Squares	F-statistic:	61.93
Date:	Mon, 25 Apr 2022	Prob (F-statistic):	5.66e-28
Time:	13:05:13	Log-Likelihood:	-754.64
No. Observations:	194	AIC:	1517.
Df Residuals:	190	BIC:	1530.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-51.7826	4.622	-11.204	0.000	-60.899	-42.666
Assists	2.1018	0.207	10.155	0.000	1.694	2.510
Rebounds	0.6454	0.126	5.114	0.000	0.397	0.894
FTM	0.3868	0.138	2.803	0.006	0.114	0.658

Omnibus:	0.786	Durbin-Watson:	1.638
Prob(Omnibus):	0.675	Jarque-Bera (JB):	0.893
Skew:	-0.087	Prob(JB):	0.640
Kurtosis:	2.717	Cond. No.	216.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Conclusions & Implications

- Given our research, the most important continuous variables are assists and rebounds.
- This is true in both our importance chart and within the regression decision tree. Assists and rebounds are also shown to have the best correlation in our heatmap as well.
- In conclusion, the variables that are the most statistically significant are assists and rebounds, respectively.
- Bruce Pearl and his Auburn basketball players should continue to focus on these two areas because of their significance in Auburn's score differential.
- In addition, Bruce should try improving on the rest of the variables to increase the team's overall performance during games.

References

- Brown, Bryce. "Predictive Analytics for College Basketball: Using Logistic Regression for Determining the Outcome of a Game." *University of New Hampshire Scholars' Repository* , University of New Hampshire , Apr. 2019, https://scholars.unh.edu/cgi/viewcontent.cgi?article=1471&context=honors.
- Magel, Rhonda, and Samuel Unruh. "Determining Factors Influencing the Outcome of College Basketball Games." *Open Journal of Statistics*, vol. 03, no. 04, 13 July 2013, pp. 225–230., https://doi.org/10.4236/ojs.2013.34026.

Acknowledgments

This work was conducted with data provided by Dr. David Paradice. Any opinions, findings, and conclusions or recommendations expressed in this poster are that of the authors.

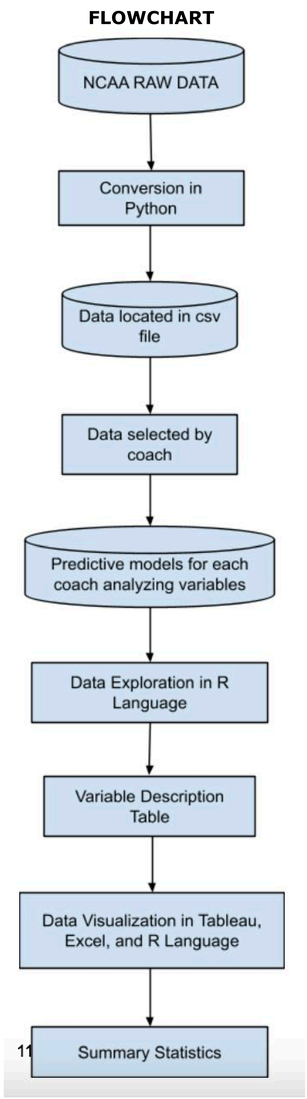
The Effect of Individual Coaches at Auburn University Basketball

Team 2 - Hadee El-Kattan, Tony Romano, Zachary Watson, Keaton Gum, Kevin Haskins, & Nico Carpio
Auburn University – BUAL 5860 002

PROBLEM STATEMENT

The scope of our problem can be widened to the various trends and associations that a coach may have in leading their team to success. We are going to compare, in two sections, the various coaches' historical basketball game data at Auburn University. Our problem can be narrowed down to: **Does a basketball team's performance reshape based on new coaches?**

PROJECT FLOW CHART



METHODS

For our final report, we decided to conclude our project by using the following models: decision tree, and SVC (support-vector classifier) machine, and logistic regression. To conceptually understand, we chose the following models because of the below reasons:

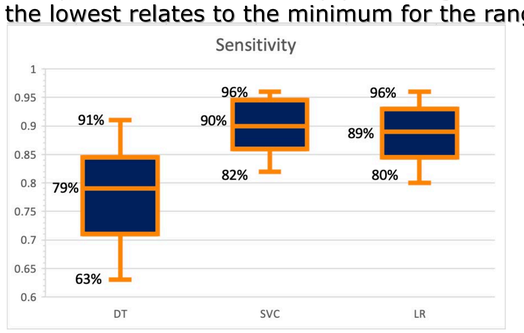
Decision tree Model:
- Easily understood and interpreted
- Handles many numerical and categorical variables for multi-output problems
- Little data preparation for predicting logarithmic data

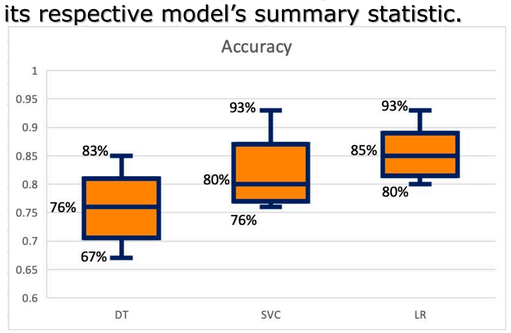
SVC Machine:
- Efficient memory space (dealing with large amounts of data)
- Clear margins between Coach Barbee and Pearl
- Effective with high dimensional spaces

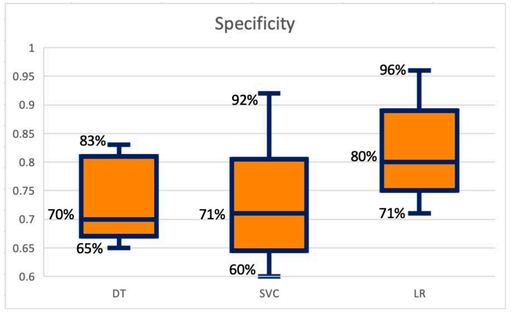
Logistic Regression:
- Easily implemented and evaluated
- Very trainable with a high number of observations (265 total)

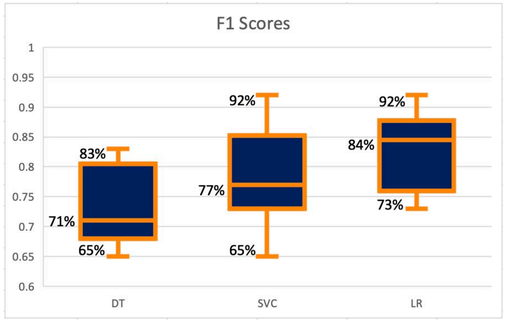
METHOD EVALUATION

Our code was set up to use 80% of our data for training and the other 20% for testing. We also made a robust process by rerunning the seed ten different times. Following, we calculated the average result. Each boxplot below represents the summary statistics of each model in terms of Sensitivity, Accuracy, Specificity, and F1 Score. The top percentage is the maximum, the middle percentage is the median, and the lowest relates to the minimum for the range of its respective model's summary statistic.







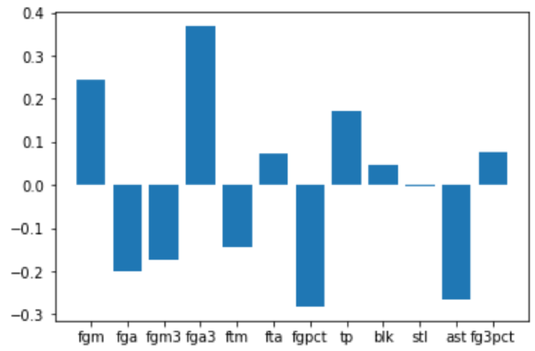


CLASSIFICATION MODELS' RESULTS

The table below illustrates the data that we chose to research for this problem. We decided to do trend analysis on season and non-season basketball games from Barbee versus Pearl.

y_variable	Year Started	Year Ended	Coach
0	2011	2014	Tony Barbee
1	2015	2019	Bruce Pearl

The graph below depicts the feature importance of our analysis during our research. In summary, this technique allows us to understand the score of our input features. We were able to determine the process of our trend analysis based on these results.



The variables in which had high importance during our research analysis included: field goal percentage, three-point percentage, free-throw percentage, blocks, assists, and steals.

FINAL CONCLUSION

After converting our data, exploring, and creating models, we can easily determine that Auburn Basketball is reaching top tier quality upon the NCAA leagues. Furthermore, the models that we chose demonstrate high statistics in determining our analysis. In that, we can conclude that Bruce Pearl had marginally higher statistics across all variables used in our problem.

ACKNOWLEDGEMENT

Thank you to Dr. David Paradice for the sports data!

Predicting Auburn Basketball Outcomes

Team 3 - The Minimalists

Kayla Ericksen, Thomas Cooney, Patrick Goggans, Will Richardson, Emily Valentine, and Charles Estes

Faculty Advisor – Xing Wang

Introduction & Project Motivation

Basketball at the collegiate level has become increasingly more competitive over the years. In recent years Auburn basketball has increased in popularity with our newfound success. With this Auburn basketball team's recent success has made it important to our group to see what it is that is helping Auburn win basketball games.

As a result, it is important to understand which factors significantly contribute to the total points scored by a team during a basketball matchup. Therefore, we have used statistics from Auburn Men's Basketball games from the 2010-2019 seasons in order to determine which factors will strongly influence the score of the team's next matchup.

Literature Review

In order to have a better understanding of basketball analytics, we sought to find literature on an already existing analysis that was comparable to our project scope. We were able to find two research articles that were relevant to our work and could be insightful to use as we bean to interpret the data. The first article is titled "Predicting National Basketball Association Winners" written by Jasper Lin, Logan Short, and Vishnu Sundaresan. This article was written by students at Stanford University as part of their final project. In this article some helpful takeaways One important point they reached was the importance of winning and the need for more statistics than a traditional box score. They also found that point differential and Win/Loss record were the greatest predictors. While this article was predicting NBA games while we were using NCAA basketball games, we found that they were similar enough to find this a very helpful article.

The second article is titled "Building an NCAA Men's Basketball Predictive Model and Quantifying Its Success" written by Michael J. Lopez and Gregory Matthews of Skidmore College.
In the first article, "Predicting National Basketball Association Winners", these students set out to determine what game factors are most important when determining the outcome of a basketball game. This research premise is very relevant to our project since we are also using box score statistics and records to predict, rather than looking at individual player statistics. They used three benchmarks to compare their models' predictions so that they can establish a scope of how accurate their models should be. The three benchmarks that they included are the following: point differential (difference between a team's average points per game and average points allowed per game), win-loss record (the win rate out of the total amount of games they played), and expert prediction (could be inflated because experts did not make predictions on games, they deemed too close to call).

Variables

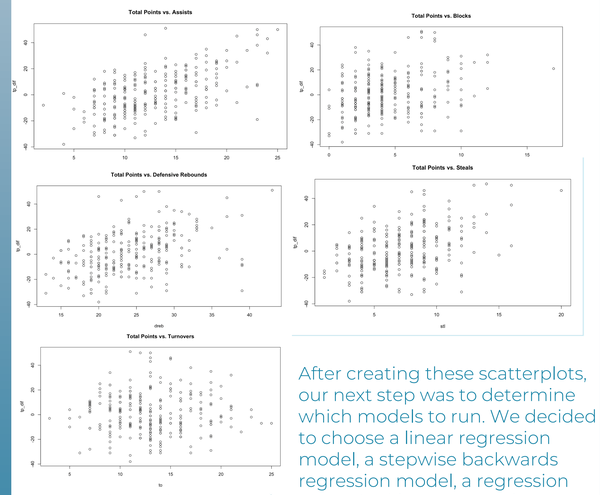
Our initial dataset consists of 22 total variables. However, we have narrowed our variable selection down to 7 independent variables and 1 dependent variable. We chose to leave out variables such as total field goals made and total free throws made, as they directly contribute to the overall score of a basketball game.

Variable Name	Variable Description
tp	**Total Points:** The total number of points scored by the team through field goals and free throws (Dependent Variable)
ast	**Assists:** Total number of assists by each team. An assist is successful when a teammate passes the ball to another player, resulting in a scored field goal.
blk	**Blocks:** In order to keep the opposing team from scoring, a defensive player might try to block an attempted field goal. If they are successful and the field goal is not scored, this is recorded in the dataset under the variable "blk".
stl	**Steals:** A steal occurs when a defensive player forces the offense to turn the ball over by snatching or swatting the ball away. This variable is listed as "stl" in the dataset.
oreb	**Offensive Rebounds:** When a player on offense recovers the ball after a field goal attempt, this is known as an offensive rebound. The variable "oreb" corresponds to the total number of offensive rebounds by a team.
dreb	**Defensive Rebounds:** When a player on defense recovers the ball after the opposing team attempts a field goal, this is known as a defensive rebound. The variable "dreb" represents the total number of defensive rebounds by a team.
pf	**Personal Fouls:** The number of personal fouls committed by a team. These fouls usually occur due to illegal physical contact with an opposing player.
to	**Turnovers:** A turnover is when a team loses possession of the ball to the other team before they are able to make a field goal attempt. By forcing multiple turnovers, the opposing team will not be able to score as often. Turnovers are recorded as "to" in our dataset.

Methodology

In order to begin the analysis, the first step we had to take was parsing the XML files and converting them to CSV files. We did this by using the program Python.

Once this was finished, we created scatterplots to show the relationship between our independent variables and the total point differentials.



After creating these scatterplots, our next step was to determine which models to run. We decided to choose a linear regression model, a stepwise backwards regression model, a regression tree model, and a random forest model. Linear regression models explain the relationship between one dependent variable and a set of explanatory variables. Regression Tree Models are easy to understand and fairly accurate when it comes to predictions. Lastly, random forest models are used to help increase the accuracy of a regression tree model.

Descriptive Analysis & Modeling

The first model that we conducted was a linear regression model. Regarding this model, we used the total points differential as the dependent variable. Our independent variables consisted of assists, blocks, steals, offensive rebounds, defensive rebounds, personal fouls and turnovers. The output from our model results are provided below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.9825    11.0472  -2.714 0.009389 **
blk           0.6022     0.8294   0.726 0.471571
ast           1.5309     0.4283   3.574 0.000853 ***
dreb          0.6312     0.4381   1.441 0.156603
stl           0.3078     0.5502   0.559 0.578600
oreb         -0.2658     0.3422  -0.777 0.441325
to           -0.5510     0.3689  -1.494 0.142234
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.51 on 45 degrees of freedom
Multiple R-squared:  0.4063,    Adjusted R-squared:  0.3272
F-statistic: 5.133 on 6 and 45 DF,  p-value: 0.0004319
```

This model was not very accurate at determining the significant variables. It also contained an R^2 of approximately 0.40.

The next model that was conducted was a stepwise backwards regression model. The purpose of this model was to drop the variables that are least significant in order to help create our best fit model.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -30.9836     8.7438  -3.543 0.000891 ***
ast           1.5189     0.4011   3.787 0.000425 ***
dreb          0.7554     0.3961   1.907 0.062515 .
to           -0.5835     0.3501  -1.667 0.102095
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.33 on 48 degrees of freedom
Multiple R-squared:  0.3876,    Adjusted R-squared:  0.3494
F-statistic: 10.13 on 3 and 48 DF,  p-value: 0.00002785
```

After running this model, we found that assists, defensive rebounds, and turnovers were the most significant variables. This was due to the fact that these variables helped to generate the best fit model with the smallest AIC value.

The next model that we generated was done through the program Python. This model was our regression tree model. In order to create our most accurate regression tree model, we used blocks, steals, assists, turnovers, and defensive rebounds. We also set our random state at 55. The maximum leaf nodes on the regression tree were set at 8 nodes.

Decision Tree RMSE: 12.7283

Decision Tree MAE: 10.4270

R-squared 0.40433544344270267

Listed above are the results from our regression tree. Although these are the most accurate results we could produce, the RMSE and MAE are still very high. Our R^2 is also noticeably lower than the R^2 from our initial linear regression model.

Descriptive Analysis & Modeling

Random Forest RMSE: 11.5735 **R-squared 0.5075174510144699**
Random Forest MAE: 9.3524

The results shown above are the results generated from our random forest model. Similarly to the regression tree model, these results are not what we expected Once again, the RMSE and MAE are incredibly high. However, the R^2 of 0.50 is significantly better than the R^2 from the regression tree model.

Overall, our best fit model was our initial linear regression model. However, we had problems with all of our models. We though that adding more data points to our initial dataset and converting total points to total point differential would fix these problems, but we were still left with unexplainably high RMSE and MAE values.

Conclusion

- Blocks, Steals, Defensive Rebounds and Assists were the significant variables in the models
- Through the decision tree model, we saw that the Assists variable was the most significant in the dataset
- Linear Regression model proved to be the best fitting predictive model to get our results with an R^2 of approximately 0.50
- The models shown correlated with the results in the literature we referenced, showing medium correlation between the variables tested and total points scored.

References

Jasper Lin, Logan Short, Vishnu Sundaresan. 2014. Stanford University. "Predicting National Basketball Association Winners," Retrieved from: http://cs229.stanford.edu/proj2014/Jasper%20Lin,%20Logan%20Short,%20Vishnu%20Sundaresan,%20Predicting%20National%20Basketball%20Association%20Game%20Winners.pdf

Michael J. Lopez, Gregory Matthews, "Building an NCAA men's basketball predictive model and quantifying its success," Journal of Quantitative Analysis in Sports, 2015, 11-1. https://www.degruyter.com/view/j/jqas.2015.11.issue-1/jqas-2014-0058/jqas-2014-0058.xml?format=I

Effects of Sleep on the Cardiovascular System

Listen to your Heart

Prof. Xing Wang || Communicating Quantitative Results in business

Alex Vojnovic, Tai Giang, Cindy Guo, Renhao Li, Noah Lynn, Sichen Tong

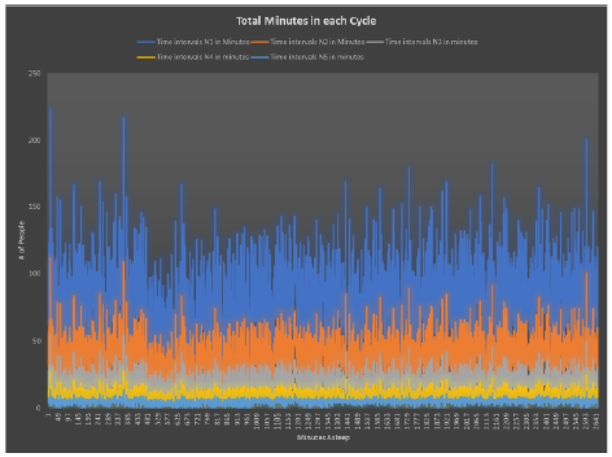
Abstract

The goal of our project was to analyze the link between heart disease, cardiovascular health and sleep apnea. Through extensive observations and data collection, we have made significant breakthroughs in the overall progress of the project. First, we analyzed the sleep cycles of over thousands of patients using data extracted from an academic article, and identified the difference in the duration of the non-stop phase between the general population and the patients. We also analyzed age, gender and race to detect the probability of such diseases. Although it is not possible to distinguish between sleep and age as the underlying cause of such disorders with the help of modern technology, our data have done as much as we can at this stage.

Introduction

Sleep is an essential part of our everyday lives. It does not discriminate between races, age, nor status. Cardiovascular health could be directly linked to the quality of sleep. The motivation behind our project is to figure out the correlation between cardiovascular health and the quality of sleep. With heart disease being the leading cause of death in America, particularly with cardiovascular disease causing the death of an individual every 36 seconds, our study hopes to provide analysis and clarity on how improving one's sleep could also improve the health and longevity of life. Our analysis of the association between cardiovascular health and sleep apnea aims to help everyone since sleep is a basic human need.

Methodology



This is a chart of the number of people and how many minutes that person stays in each cycle. We see that people are staying in the N1 and N2 stage the most and less in N3, N4, and N5. This could be a potential reason why many aren't getting enough sleep. Since N1 and N2 sleep is where one is easily woken up, it could point out that many are being disturbed when they try to sleep, whether it is because of their phones, emotions, others like family, etc,.

Data

- Includes data from the SHHS survey of 6,600 adult participants aged 40 and older with various cardiovascular-related conditions such as heart disease and hypertension

Variables includes:

- 2651 patients in the dataset
- 5 steps in the sleep cycle (n1-n5) counted by 30 seconds per column
- 37 symptom and patient information based variables,

such as age, gender, race

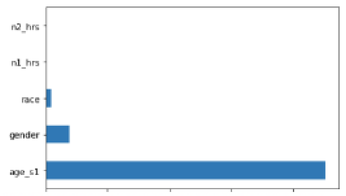
- Any_cvd, Any_chd

Results

When it comes to having "Any Cardiovascular disease" or "Any Coronary Heart Disease", our model shows that age is the most important feature when the x variables are age, gender, race, and numbers of hours asleep in stages n1 to rem.

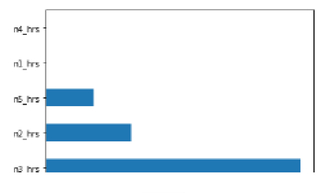
Although our diagnostic performance may not be optimal, it could be explained by other factors like lifestyle. Since Age is the biggest contributing factor, it is hard to tell whether the heart diseases were developed because of sleep or because of natural causes in aging.

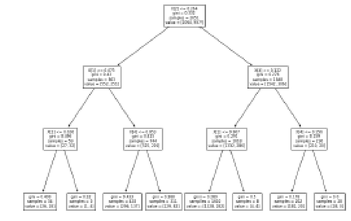
Any_cvd with n1:n5, age,race,gender



Accuracy = 0.7784431137724551
Sensitivy = 0.8
Specitivity = 0.7628865979381443

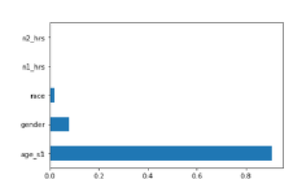
Any_cvd with only n1:n5

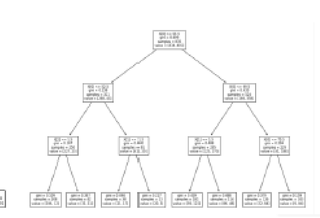




Accuracy = 0.7966101694915254
Sensitivy = 0.0
Specitivity = 1.0

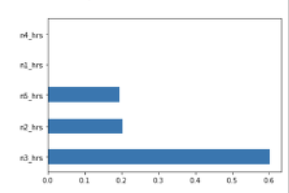
Any_chd with n1:n5, age, race, gender

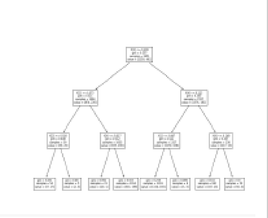




Accuracy = 0.7844311377245509
Sensitivy = 0.7647058823529411
Specitivity = 0.797979797979798

Any_chd with only n1:n5





Accuracy = 0.8531073446327684
Sensitivy = 0.0
Specitivy = 1.0

Conclusion

- Through our data and observations, age and emotion are both important components of sleep quality, which can also be directly associated with cardiovascular disease

- We recommend that people go to bed 20 minutes earlier and get 7.5 hours of sleep each day to experience a full sleep cycle

- In addition to sleep duration and quality, we also searched for other factors and came up with the best ways to prevent cardiovascular disease. Maintain a good state of mind, be physically active, lower blood cholesterol, and control nicotine intake.

Reference

Nagai, Michiaki, et al. "Sleep Duration as a Risk Factor for Cardiovascular Disease- A Review of the Recent Literature." Current Cardiology Reviews, Bentham Science Publishers Ltd., Feb. 2010, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2845795/.

Zhang, Jihui, et al. "Sleep Patterns and Mental Health Correlates in US Adolescents." The Journal of Pediatrics, Mosby, 7 Dec. 2016, https://www.sciencedirect.com/science/article/abs/pii/S0022347616312355.

Kuehn, Bridget M., and Bridget M. Kuehn Search for more papers by this author. "Sleep Duration Linked to Cardiovascular Disease." Circulation, 20 May 2019, https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.119.041278.

Predicting Cardiovascular Disease Based on Irregular Sleep Habits

Jack Buck, Rebecca Self, Alex Edwards, Michael Fulda, Will Ogden & Jonathan Hong
Auburn University Harbert College of Business – Dr. Xing Wang

Introduction

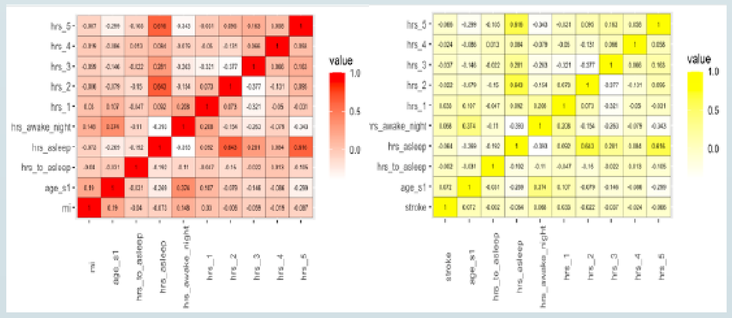
The Case Study we reference is The Sleep Heart Health Study (SHHS). It is an organization that has created this cohort study for looking into the issue of obstructive sleep apnea (OSA) and other sleep-disordered breathing (SDB). The data for the sleep risk factors were collected using a polysomnogram (Sleep Study).The dataset given has a total of 2651 patients. The heart related variables were collected from the six previous other studies, while the bulk of the data is the sleep cycle data collected from the SHHS using a polysomnogram.. In total, there are five different sleep cycles and in the dataset each cell represents a 30 second interval of that stage.

Problem Statement

Our goal is to create a model that can accurately predict cardiovascular health problems using sleep habits. This would help doctors identify potential risks of stroke or heart attack in patients from their sleeping patterns.

Variables

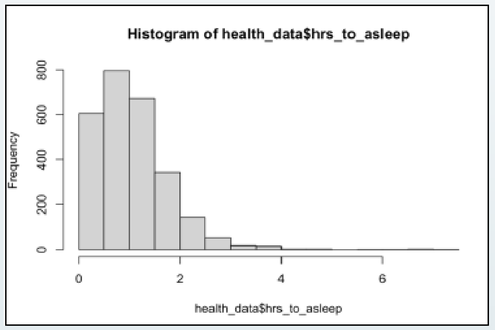
Variable Name	Variable Definition	Variable Type
Hrs_to_sleep	Time it takes patient to fall asleep	Continuous Data- Ratio
Hrs_asleep	The amount of time patient is asleep	Continuous Data- Ratio
Hrs_1	Time spent in stage 1 sleep cycle	Continuous Data- Ratio
Hrs_2	Time spent in stage 2 sleep cycle	Continuous Data- Ratio
Hrs_3	Time spent in stage 3 sleep cycle	Continuous Data- Ratio
Hrs_4	Time spent in stage 4 sleep cycle	Continuous Data- Ratio
Hrs_5	Time spent in stage 5 sleep cycle (REM sleep)	Continuous Data- Ratio
Gender	1= male and 2= female	Categorical Data- Nominal
Race	1= white, 2=Black, and 3= Other	Categorical Data- Nominal
Age_s1	Age of patient during study	Continuous Data- Ratio
MI (Outcome Variable)	Number of myocardial infarctions (MIs) Since Baseline	Continuous Data- Ratio
Stroke (Outcome Variable)	Number of Strokes Since Baseline	Continuous Data - Ratio

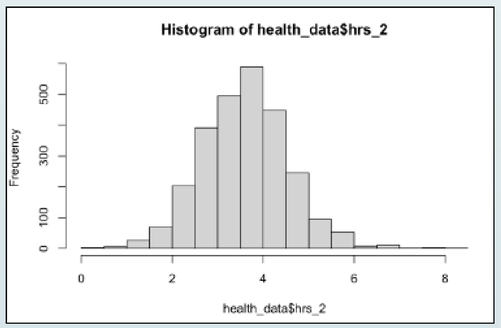


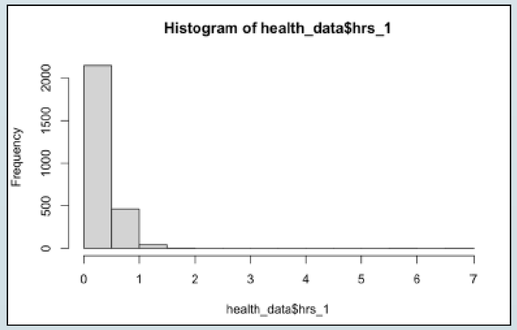
Data Processing

The first step of our project was to find a way to utilize our polysomnogram data. The data captured the sleep cycles of each patient throughout one night of sleep. In order to use this data, we had to engineer new variables that could be used by our models.

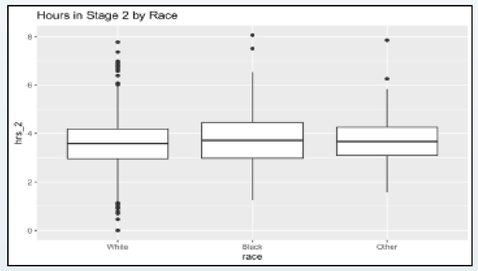
The variables we created measure the hours it takes to fall asleep, the time spent awake in the night, and total hours spent asleep.







Data Processing Continued



Methods and Cross Validation

After running our initial models, we had achieved a good accuracy score, but very low sensitivity. We believed that this was due to our data set having more healthy patients than patients with heart disease. We adjusted our data set to train the model with equal numbers of healthy and unhealthy patients. This improved the sensitivity and specificity of our models, while accuracy only decreased slightly. We divided the dataset into five sets of as equal size as possible. Each set was used once as a testing set, while the rest were used as training sets. This gave us an 80/20 training/testing split.. This allowed us to run 5 seeds of each dataset, giving the models robustness.

Results

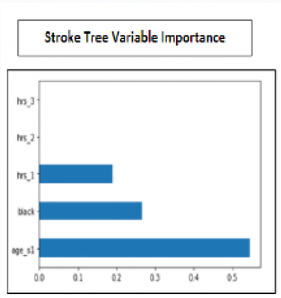
The values in the tables below are averages of all seeds run. The Decision Tree is our best model for stroke, and the K-Nearest Neighbors is our best model for myocardial infarctions.

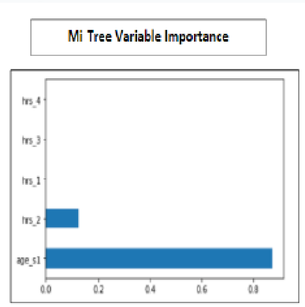
Mi	Accuracy	Sensitivity	Specificity	F1-Score
Decision Tree	62.26	44.33	65.23	0.48
Log. Regression	64.2	52.44	72.82	0.54
K-Nearest Neighbors	67.61	67.08	57.43	0.62

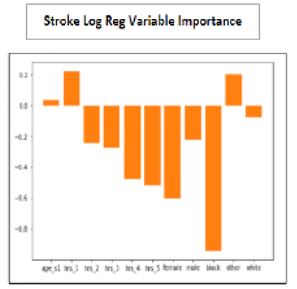
Stroke	Accuracy	Sensitivity	Specificity	F1-Score
Decision Tree	72.93	57	76.83	0.72
Log. Regression	70.36	49.86	84.19	0.56
K-Nearest Neighbors	65.91	67.24	57.26	0.61

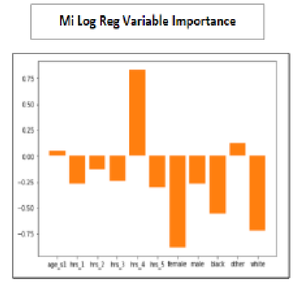
Results Continued

The most consistent variable of importance was the participant's age. We found that the sleep data itself was not a huge factor in predicting heart issues.









Conclusion

From our results, the best model for predicting strokes was the decision tree model with an accuracy of 72.93%, and for mi the best for prediction was KNN with an accuracy of 67.61%. We found that the most significant factors when predicting were the age of the patient and sleep stages 1 and 2. This pattern shows that the early stages of the sleep cycle are important in determining future heart related issues.

Acknowledgements

This work was conducted with the Health Care Sleeping Data by Dr. Rupesh Agrawal from Northern Kentucky University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of Northern Kentucky University.

Maximizing Auburn Men's Basketball's Chance of Winning A Championship in the Following Season

D1 NCAA business analysts: Hao Feng; Will Nutter; Connor Tidwell;
James Robinson; Lauren Huether; Wanling Zhu; Instructor: Dr. Xing Wang

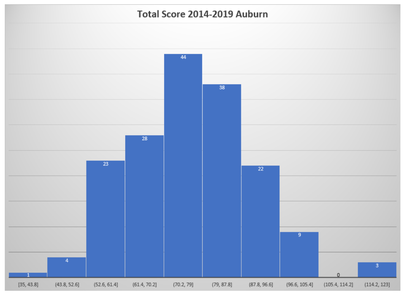
Motivation

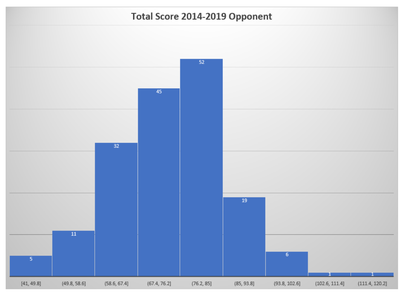
The Auburn basketball team made it in 2017-2018 and 2018-2019 when Bruce Pearl took over as head coach, but the squad's ultimate objective is to win the NCAA championship. We intend to better comprehend an opponent's talents and overall style of play by analyzing and researching data, which will help our basketball team make tactical adjustments.

Data Summary

- Our dataset focused on the statistics from the Auburn basketball team's games from 2014 to 2019.
- An analysis of group pairs, offensive and defensive styles, as well as a comparison with current season championship teams, comprise the project.
- The model will be built in Python, with data processing, analysis, and visualization in Excel and SQL.

Data Measuring

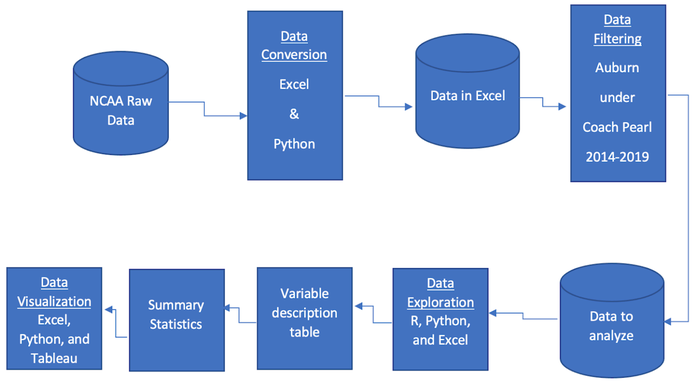




Variables

Name	Description	Explanation
W/L (y variable)	Win or Loss	Result of the game(win or loss)
BLK	Blocks	A block occurs when the defense player tips the ball, blocking their chance to score
ftm	Free throws made	The % made field goals while on the court
fta	Free throws attempted	The number made field goals while on the court
Total 3pt	3 point Field goals made	The number of a team's 3-point field goals made while on the court
Made 3PT	3 point Field goals percentage made	The % of a team's 3-point field goals made while on the court
ofeb	offensive rebounds	The number of rebounds gathered while they were on offense
Dreb	Defensive Rebounds	The number of rebounds a player or team has collected while they were on defense
STL	Steals	Number of times that takes the ball from a player on offense, causing a turnover
to	Turnovers	The number of turnovers while on the court
ast	Assists	The number of assists while on the court

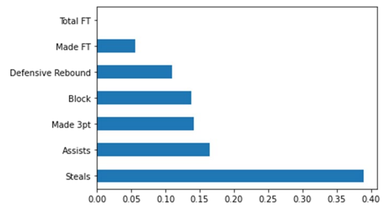
Method



Models and Results

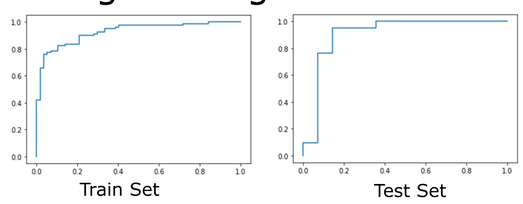
The following models were used:

Decision Tree



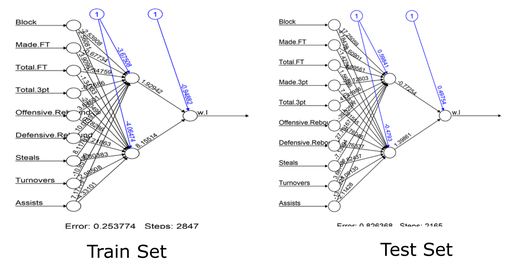
	Train Set	Test Set
Accuracy	0.8015	0.6286
Sensitivity	0.6538	0.4211
Specificity	0.8929	0.8750

Logistic Regression



	Train Set	Test Set
Accuracy	0.8456	0.8286
Sensitivity	0.8987	0.7619
Specificity	0.7719	0.9286
AUC	0.92604	0.90816

Neural Network



	Train Set	Test Set
Accuracy	0.7462	0.1736

Conclusions and Implications

- We have analyzed the data using three different tools and three different models. The defensive data will always account for the majority of the contribution to the victory. For example, defensive rebounds and steals, these two variables that always play the most critical role in winning or losing a game, regardless of the all model.
- In the offensive stats, free throws and made 3pt that take up the biggest contribution to the winning percentage. The number of three-pointers made becomes the key to winning the game.
- Basketball is a sport of defense over offense. If the Auburn basketball team wants to get better the rest of the years, it must focus more on defensive play.

Reference

[1]Magel, Rhonda, and Samuel Unruh. "Determining Factors Influencing the Outcome of College Basketball Games." Open Journal of Statistics, vol. 03, no. 04, 2013, pp. 225–30. Crossref, https://doi.org/10.4236/ojs.2013.34026.

[2]Mikołajec, Kazimierz, et al. "Game Indicators Determining Sports Performance in the NBA." Journal of Human Kinetics, vol. 37, no. 1, 2013, pp. 145–151., https://doi.org/10.2478/hukin-2013-0035.

[3]Csataljay G , James N , Hughes M , Dancs H . Performance indicators that distinguish winning and losing teams in basketball . Int. J Perform Anal Sport . 2009 ; 9 : 60 – 66 .

Acknowledgements

This work was conducted with data from NCAA Division 2 basketball game data provided by Dr. David Paradice from Auburn University, Harbert College of Business.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Dr. David Paradice.

Team 7

People's Choice Winner

Predicting the Winner at Halftime in College Basketball

BUAL Ballers: Jiawei Tong, Tiancheng Jiang, Tanner Rowburrey, Madison Morrow, W. Buddy Haas, Annabel Antoniak Mentor: Dr. Pankush Kalgotra

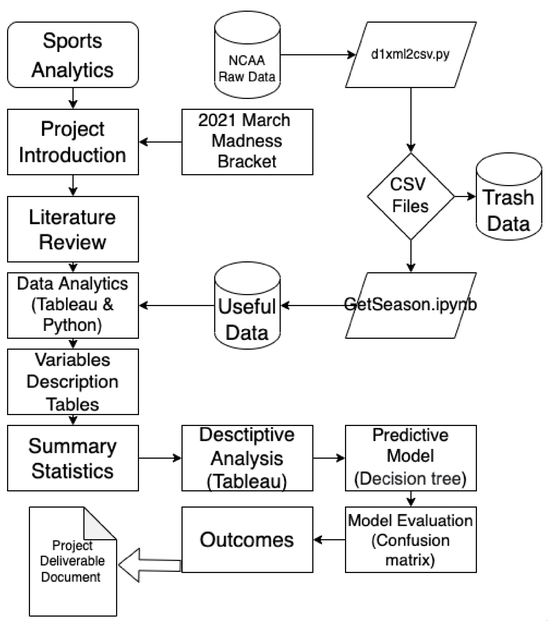
Introduction

- Auburn's basketball program has become more competitive
- Previous literature states 3 pointers are primary variable
- We created descriptive analytics to gather an understanding of the data
- Using artificial intelligence, we used the first half of games to predict which team will win

Data Set Description

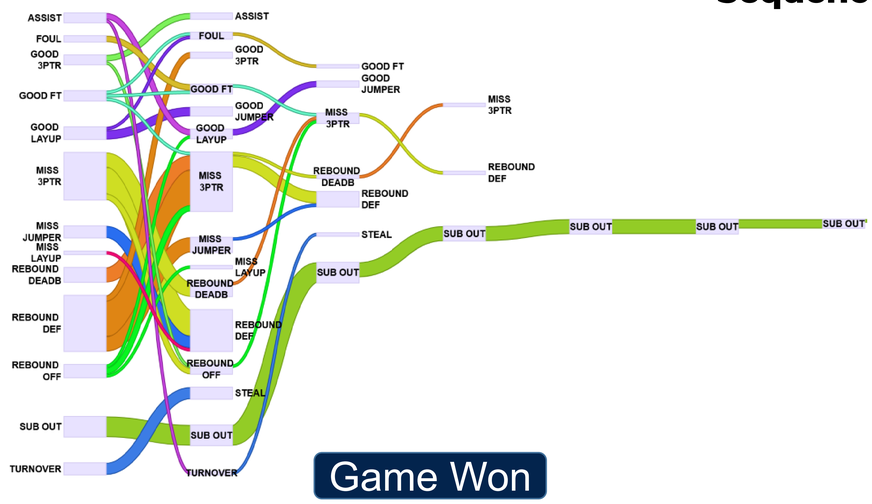
- 10 seasons, 2009-2019
- +6000 games in a season
- Play-by-play info in each game
- Variables
  - Field goal made (any scoring), 3 pointers, Free throws, Assists, Fouls, Rebounds, Turnovers
- Selected teams:
  - Duke, Kentucky, Gonzaga, Michigan, Villanova, Arizona, Kansas, Louisville, UConn
- Auburn analyzed separately

Methodology



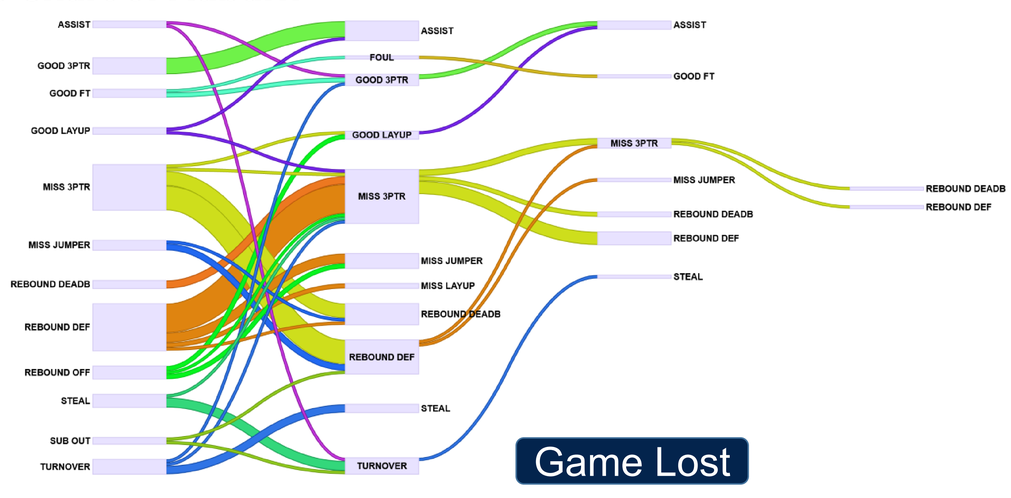
Acknowledgments
Special thanks to Dr. David Paradice for providing us with the data set and notebooks for our research, and thanks to Dr. Pankush Kalgotra for his support and guidance.

Sequence Chart of Auburn



Game Won

- It shows Auburn's gameplay when they win and lose.
- When Auburn has won, they have frequently substituted players in the game.

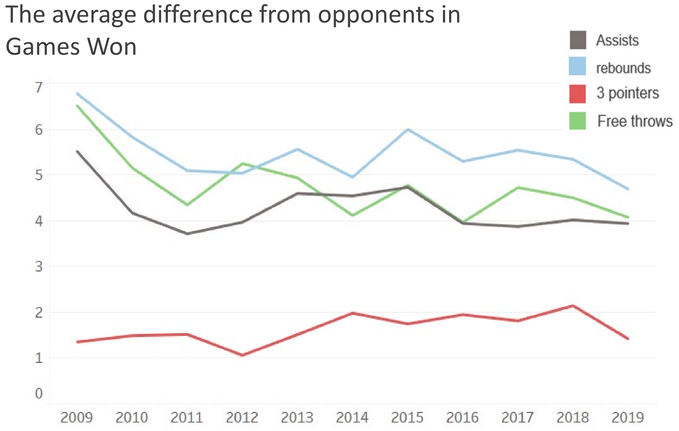


Game Lost

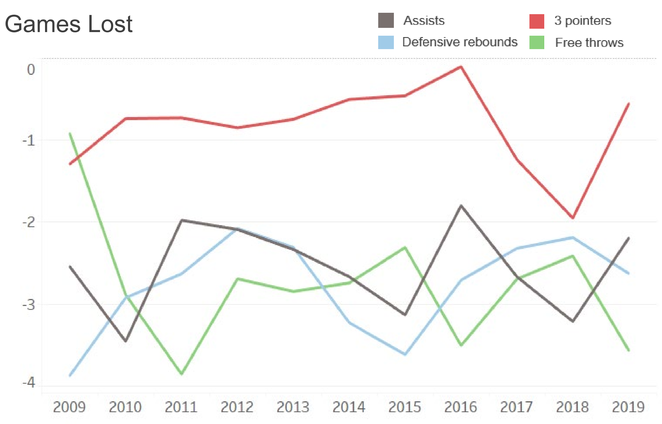
- In games Auburn has lost, rebounds are often followed by missed 3 pointers.
- Also, in games lost, good 3 pointers always lead to more assists.

Win vs Lose

The average difference from opponents in Games Won



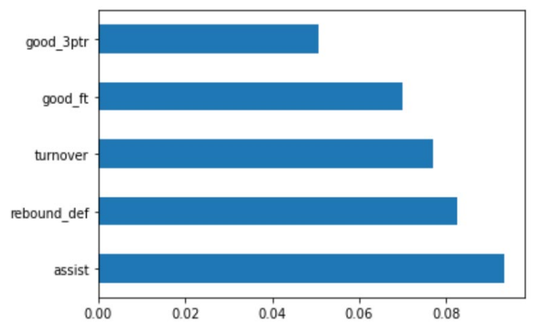
Games Lost



AI: Building the Model

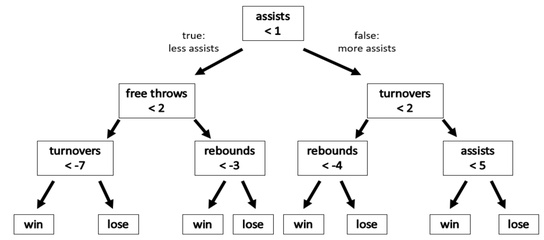
- We used a supervised predictive model.
- It took data from the first half of 1558 games from selected teams.
- Independent (x) variables are the difference between the team and their opponent.
- Dependent (y) variable: win-lose
- 60% (train) : 40% (test)

Overall accuracy: 80.7%
Sensitivity: 78.7%
Specificity: 82.8%



- Above shows variable importance in the model.
- Assists and rebounds have the most significant impact on whether a team will win or lose.

AI: Decision Tree



For example,
if a team has more assists than the opponent, less than 2 turnovers, and the opponent has more than 5 rebounds
the team is predicted to win.

Conclusion

- In contrast to the NBA, college basketball has less emphasis on 3 pointers.
- Successful college basketball teams rely more heavily on teamwork since more assists are likely to produce positive results.
- The cohesive gameplay of high performing teams is similar to Auburn's. Auburn has more frequent player substitutions in winning games, which shows a dynamic playstyle.

References
Onwuegbuzie, Anthony J. "Factors Associated with Success among NBA Teams." The Sport Journal, 27 Nov. 2013, https://thesportjournal.org/article/factors-associated-with-success-among-nba-teams/.

Team 8

Winner – First Place

Do You Wake Up at Night Frequently? Early Detection of Heart Conditions using Sleep Patterns

Chloe Mikus, Isabel De Armas, Pierce Dickson, Brady Watts, and Anne Hays Wright
Auburn Analysts
Faculty Advisor: Dr. Pankush Kalgotra

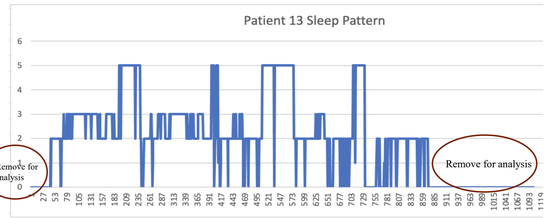
INTRODUCTION

- Sleep has 5 stages
- Sleep and heart health are related
- Relationship between gender, race, age, and heart health
- Every year, about 647,000 Americans die from heart disease, making it the leading cause of death in the United States

Is it possible to detect angina, cardiovascular disease, and coronary heart disease through the analysis of a patient's sleep pattern?

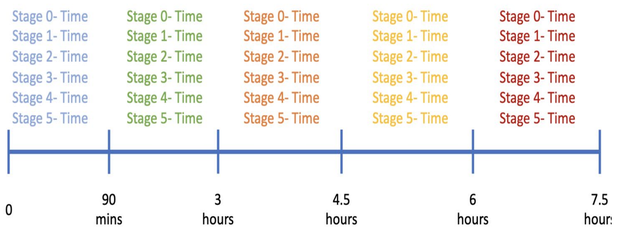
DATA

- The dataset comes from The Sleep Heart Health Study
- This study is in affiliation with the Department of Medicine and Respiratory Sciences Center at University of Arizona
- There are 2,651 patients in the dataset
- Brain wave activity was measured every 30 seconds for each patient



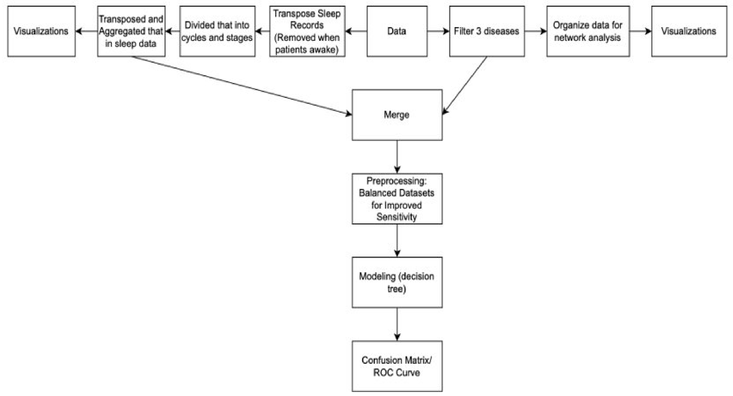
- Stage 0 is when the patient is still awake
- Stage 1 is the lightest stage of sleep
- Stage 2 makes up 50% of our sleep
- Stage 3 is when we begin deep sleep
- Stage 4 is exclusively deep sleep
- Stage 5 is REM sleep

Patient Timeline



A patient timeline sleeping for 7 ½ hours.

METHODS

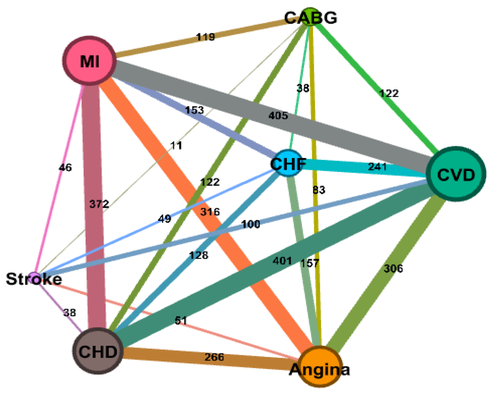


Description

	Cardiovascular Disease	Coronary Heart Disease	Angina
Number of Patients since Baseline	557 patients	401 patients	1208 patients
Average Age	69 years old	68 years old	64 years old
Gender	Male: 347 patients Female: 210 patients	Male: 263 patients Female: 138 patients	Male: 592 patients Female: 616 patients
Race	White: 484 patients Black: 53 patients Other: 20 patients	White: 352 patients Black: 34 patients Other: 15 patients	White: 1161 patients Black: 38 patients Other: 9 patients

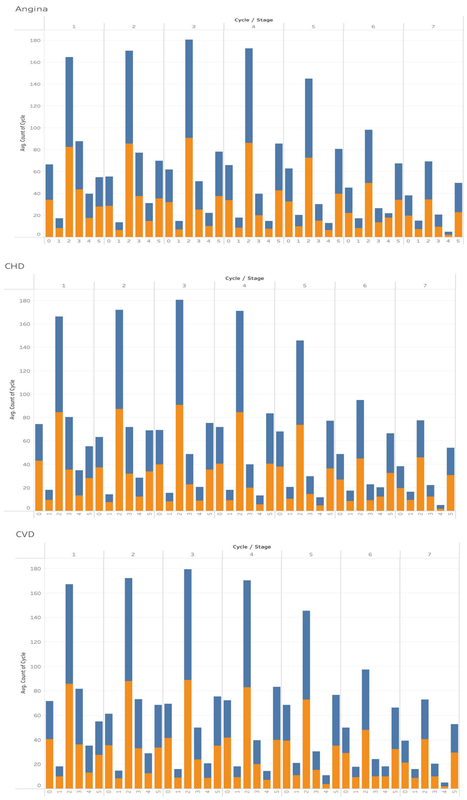
Network Diagram

- 306 patients with cardiovascular disease also have angina
- 401 patients with cardiovascular disease also have coronary



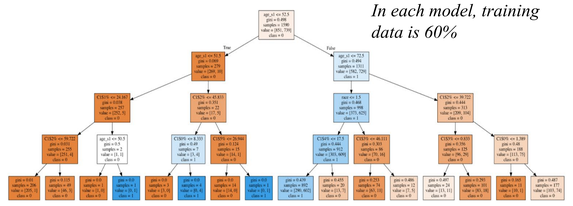
Descriptive Statistics

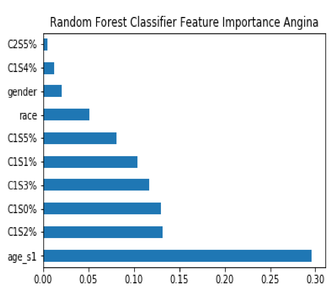
- Those with CHD and CVD tend to stay awake longer than those without
- Average count of cycle for those with CVD is 40.65 and without is 31.07 in cycle 1 stage 0
- Average count of cycle for those with CHD is 42.93 and without is 31.33 in cycle 1 stage 0

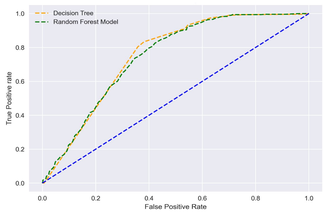


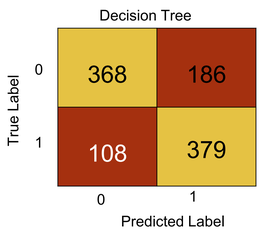
Angina Results

In each model, training data is 60%



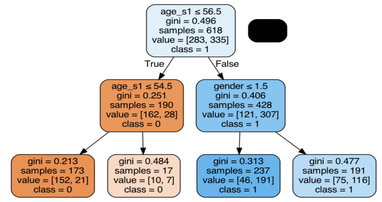


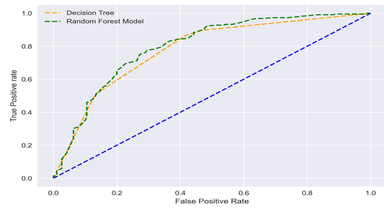




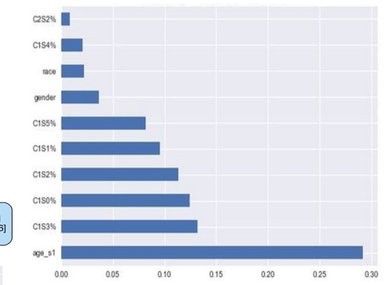
Accuracy: 71.8%
Sensitivity: 77.3%
Specificity: 67.1%

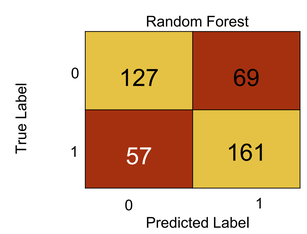
Cardiovascular Disease Results



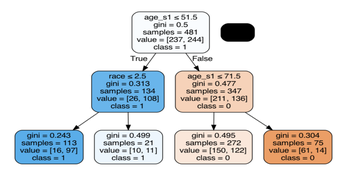


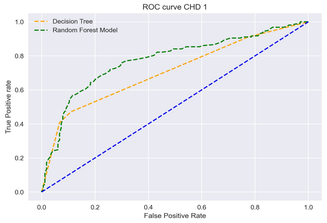
Accuracy: 69.6%
Sensitivity: 69.0%
Specificity: 70.1%



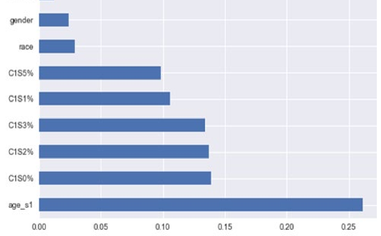


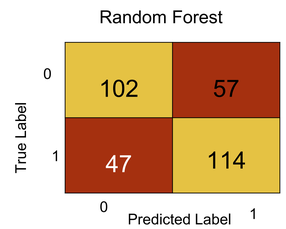
Coronary Heart Disease Results





Accuracy: 67.6%
Sensitivity: 68.4%
Specificity: 66.9%





Conclusion

- We ran models with unbalanced data sets which resulted into very low sensitivity
- It is possible to use patients sleep data to predict early detection of heart disease including angina, cardiovascular disease, and coronary heart disease

Acknowledgements

This work was conducted with data from a dataset provided by Dr. Rupesh Agrawal. All of the work and opinions on this poster are those of the authors.

Team 9

Forecasting Top Three Pickle Jar Products

Pickle Jar People: Colin Scollard, Nick Morgan, Rosie Gallucci, Jack Jones, Jack Rubisch
Faculty Advisor – Dr. Pankush Kalgotra

Introduction

Motivation:

- The Importance of sales forecasting
  - Poor forecasting leads to unnecessary amount of time, money, and problems
  - 12% companies forecast correctly
  - Most successful companies have a thorough understanding of inventory management and for any company to thrive they must obtain this same understanding

Research Questions:

1. What trends will our top 3 sellers for 1 of their locations experience in the next 12 months?

2. How does seasonality affect the purchases of these products?

3. Which stores are purchasing the highest quantity of pickle jars?

4.How can we utilize our knowledge of our top 3 pickle jar sales at all three levels?

Dataset Description

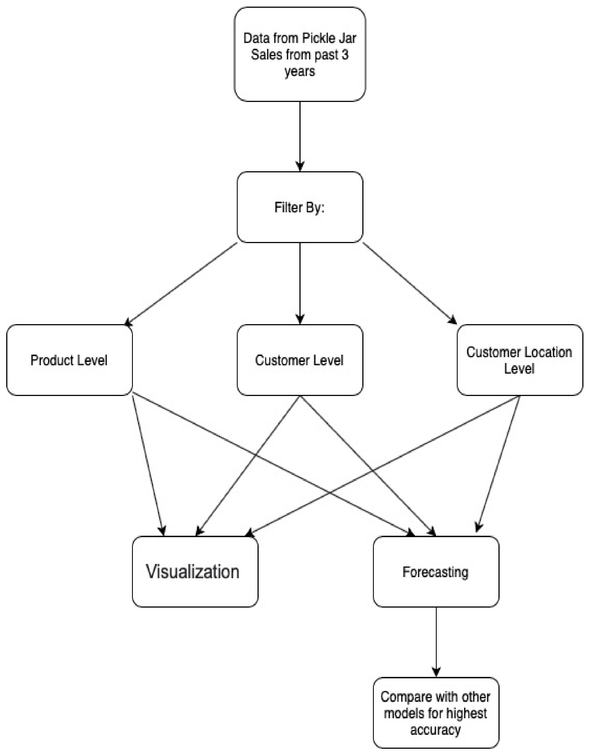
Table 1. Description of Dataset

Title	Description	Unique Quantities
MonthID	Six Digit number depicting the year and month of the product sale	August 2014-August 2017
Product	Type of pickle jar sold to a customer group	528
CustomerGroup	Market Distributor such as a chain store, capable of sending the pickle jars to their individual stores	302
CustomerShipTo	Individual store that the customer group sends the pickle jar products to	9,985
Quantity	Amount of jars sold through the transactions on the dataset	

An excel sheet was given from a pickle jar distribution company with their historical information.

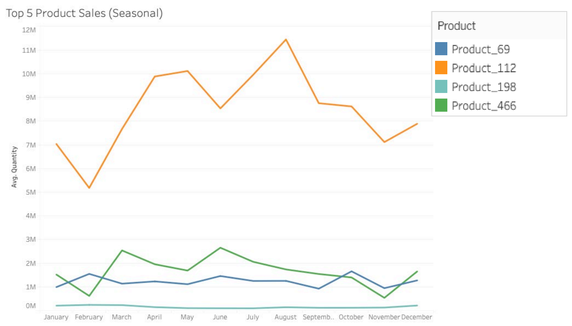
Methodology

Figure 1. Methodology

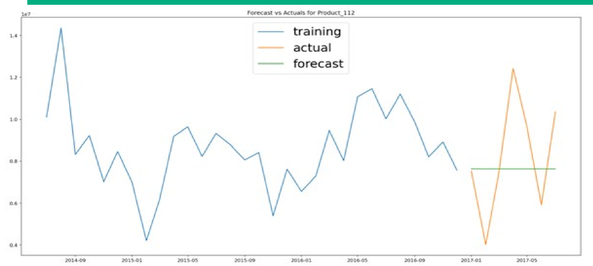


Models Used: ARIMA, Historical Forecasting, and Exponential Smoothing

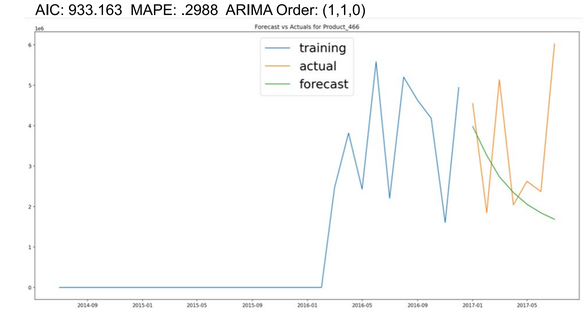
Visuals & Models



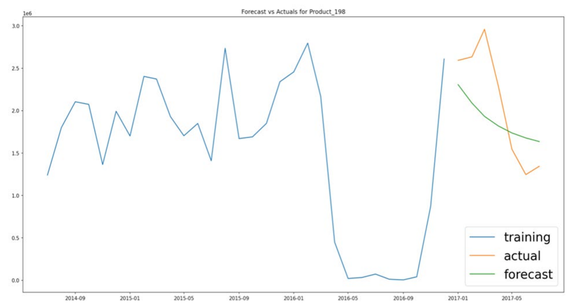
Results



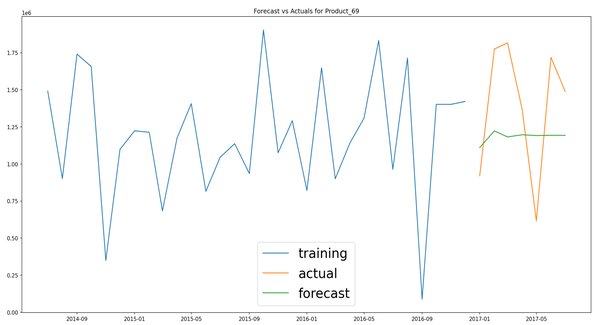
AIC: 933.163 MAPE: .2988 ARIMA Order: (1,1,0)



AIC: 938.325 MAPE: 0.221 ARIMA Order: (1,0,0)



AIC: 891.858 MAPE: 0.3812 ARIMA Order: (1,0,0)



AIC: 862.187 MAPE: 0.3590 ARIMA Order: (1,0,0)

Conclusions

1: The trends for our top 3 sellers at one location forecasted for the next 12 months are as follows:

Product 112: We forecast product 112 to flatline in sales based on our January 2017 sales number.

Product 466: For product 466 we predict it to have a rapid decline in sales as it approaches the months of June and July in 2017.

Product 69: Seems about every 3 months sales go down . Our forecast predicts a more accurate trend of a product sales at

2: Product 112:

* experienced a drastic decline in sales in the fall and winter months
* typically beginning in September
* ending in the spring/summer where sales begin to increase.

Product 198: experiences a positive seasonal trend in the Fall months such as September October and November. The sales pattern outside of Fall season don't have any consistent trend.

3: The top three location sales:

1st: Product 112 at customer group 78 and store location 179 with a quantity of 306,435,301

2nd: Product 466 at customer group 78 and store location 179 with a quantity of 61,674,322

3rd: Product 69 at customer group 48 and store location 96 with a quantity of 45,504,050.

4: Product 466:

* Wasn't on the market until early 2016.
* Successful with some seasonal fluctuations in June and July.
* With this knowledge we plan to decrease inventory of this product in the months of June and July and possibly cut the price to avoid inventory surplus.

Product 112:
* We see that when its first introduced to the market it was very successful
* Since then has fluctuated in the fall and winter months.
* Our forecast predicts a flatline in sales from 2016 moving into 2017 and 2018.

Product 69:
* Based on our forecast we want to have less inventory as stated in the actual trend
* We need to keep a balanced inventory of this product due to the 3-month fluctuations

Acknowledgements and References

Special thanks to Professor Pankush Kalgotra for his guidance and advice throughout the entire semester for completion of this project. We thank Mr. Mohit Dobhal from Decision Spot, LLC for providing us the dataset.

Team 10

Winner – Second Place

Forecasting Pickle Jar Sales with ARIMA Time Series Model and Exponential Smoothing

Dill or No Dill | Jack Ray, Danny Trainer, Kayla Taylor, Anthony Bostany, and Noah Vaughn

Faculty Advisor – Dr. Pankush Kalgotra

Introduction

Problem: The sales forecast for a pickle jar company will be conducted for the last 10 months (10/2016 - 7/2017) of our data based on the previous 27 months' figures. Performance will be measured against actual results for the most accurate conclusion.

Motivation: 67 percent of companies lack a formalized approach to sales forecasting; however, companies with accurate forecasts are 10 percent more likely to grow revenue year-over-year, on average (Roberson.)

Importance: The goal is to minimize any opportunity costs that may occur by uncovering a defensible competitive edge. This research is important because without sales forecasting, growth will be consistently stagnant.

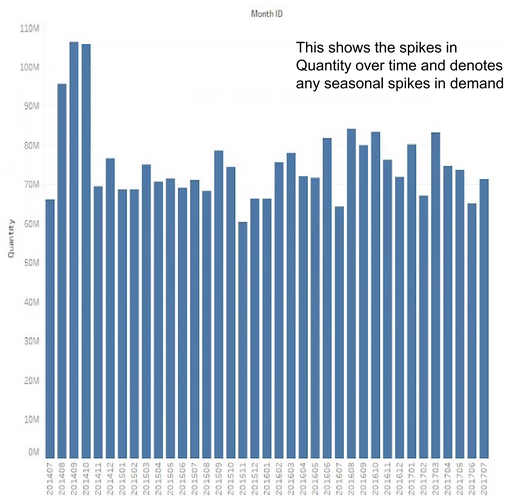
Research Question: Can the previous measures conducted related to sales forecasting be outperformed through ARIMA Time Series and Exponential Smoothing?

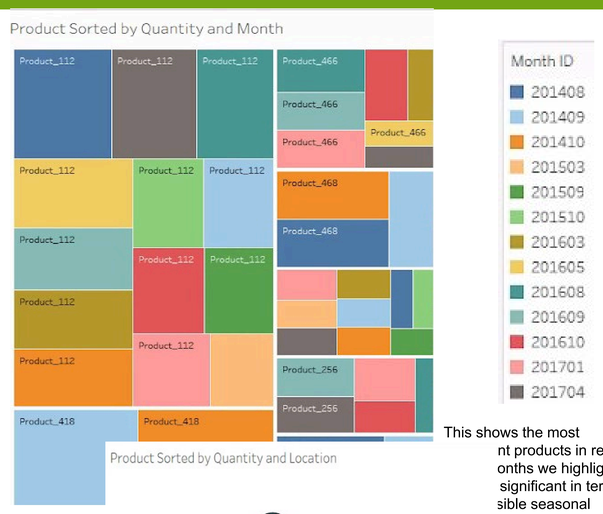
Dataset Description

We had 6 columns of data in a date range from 2014 to 2017

- **Product** - 528 unique products
- **CustomerGroup** - anonymous "name" for each customer
- **CustomerShipTo –** anonymous "name" for the location of each customer
- **CustomerShipToName -** CSTN is insignificant due to our variables all being anonymous. Therefore, CST and CSTN are identical for the purposes of analysis.
- **MonthID (YYYYMM) -** the month and year tied to each transaction
- **Quantity –** number of pickle jars sold

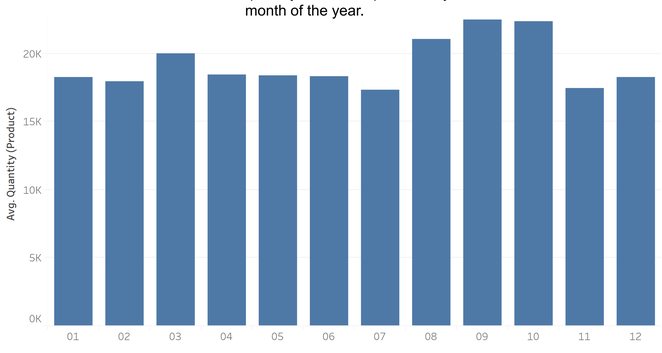
Visuals

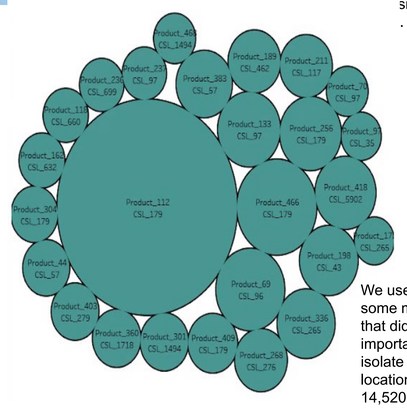




This shows the most nt products in relation onths we highlighted significant in terms of sible seasonal .

This shows the average quantity sold for all products by month of the year.

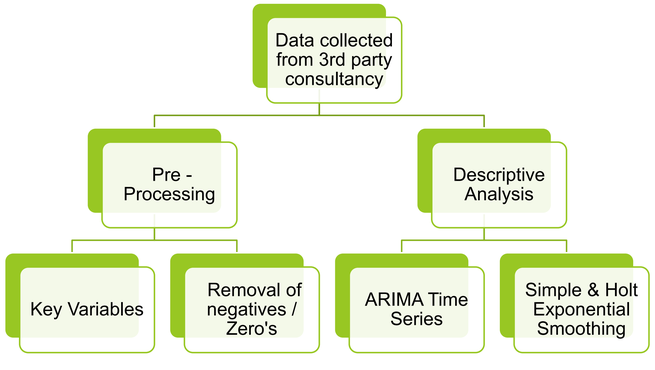




We used this model to show some more significant products that didn't show up via seasonal importance. We were able to isolate the highest selling locations as well. Min value of 14,520,615 units and max of 316,549,404.407 shown

Methodology

Purpose: The cleaning and modeling techniques serve to meet customer needs quicker, cheaper, and at greater value.



Modeling Results

- We ran models for three different levels of the data, with training data of 27 months and testing data of the following 10 months
- We models were for each product, then every product-customer group combination, and every product-customer group-customer ship to combination.
- For each level, we used four different models to do the forecasting.
- These included ARIMA Time Series, simple exponential smoothing, Holt exponential smoothing, and the company's current forecasting method.
- The current forecasting method involves taking the sales quantity from the selected month in the previous year and using that number to forecast for the current month.

Product Models Results

Pf

	MAPE	MSE
ARIMA Time Series	1.24	107,163,530,196.73
Simple Exponential Smoothing	1.05	73,634,361,604.23
Holt Exponential Smoothing	1.23	235,933,387,144.79
Current Forecasting Method	1.26	122,726,939,147.49

Product-Customer Group Models Results

	MAPE	MSE
ARIMA Time Series	1.09	21,884,852,835.01
Simple Exponential Smoothing	0.99	14,844,116,367.19
Holt Exponential Smoothing	0.89	44,484,935,802.74
Current Forecasting Method	0.94	25,662,130,467.20

Product-Customer Group-CustomerShipTo Models Results

	MAPE	MSE
ARIMA Time Series	1.14	14,355,500,381.50
Simple Exponential Smoothing	1.07	10,804,789,319.55
Holt Exponential Smoothing	0.89	24,740,675,296.91
Current Forecasting Method	0.95	16,819,173,173.55

Conclusions and Implications

Best Performing Models for Each Grouping by MAPE:

Product:
Simple Exponential Smoothing

Product-Customer Group:
Holt Exponential Smoothing

Product-Customer Group-Customer Ship To:
Holt Exponential Smoothing

For sales forecasting, it is vitally important to have cleaned and accurate data so that models can run smoothly. Sales forecasting is a necessary undertaking if a business wants to be successful, but time and resources must be used to ensure it is done correctly.

Acknowledgement:
We would like to thank Mr. Mohit Dobhal from Decision Spot, LLC for sharing this dataset with us.

Team 11

Can sleep patterns predict a heart disease?

Sleeping Tigers: Morgan Lyons | Kathryn Storey | Benji Parker | Tanner Freise

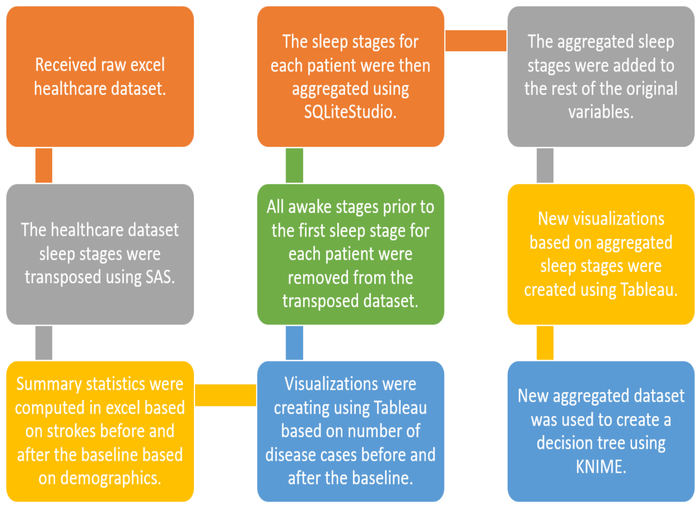
Professor Kalgotra

Introduction

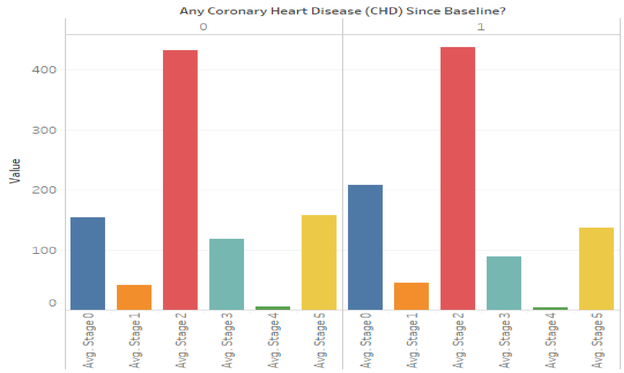
- Sleep-Disorder breathing with heart disease
- Major developments: lung, cardiovascular, heart disease
- Test whether sleep-related breathing is associated with an increased risk of coronary heart disease, stroke, all cause mortality, and hypertension.

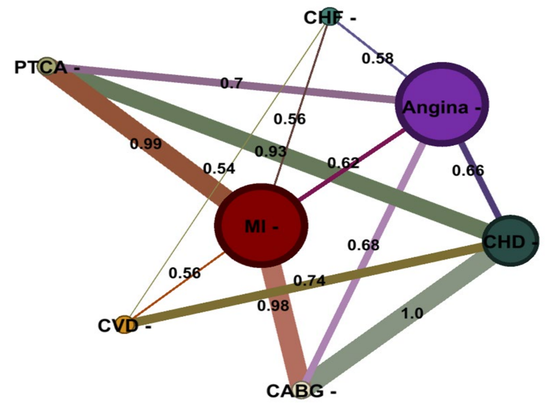
Method

- 2,651 Patients
- Sleep data per patient 1,364 30-second interval
- 6 sleep patterns



Descriptive Analytics



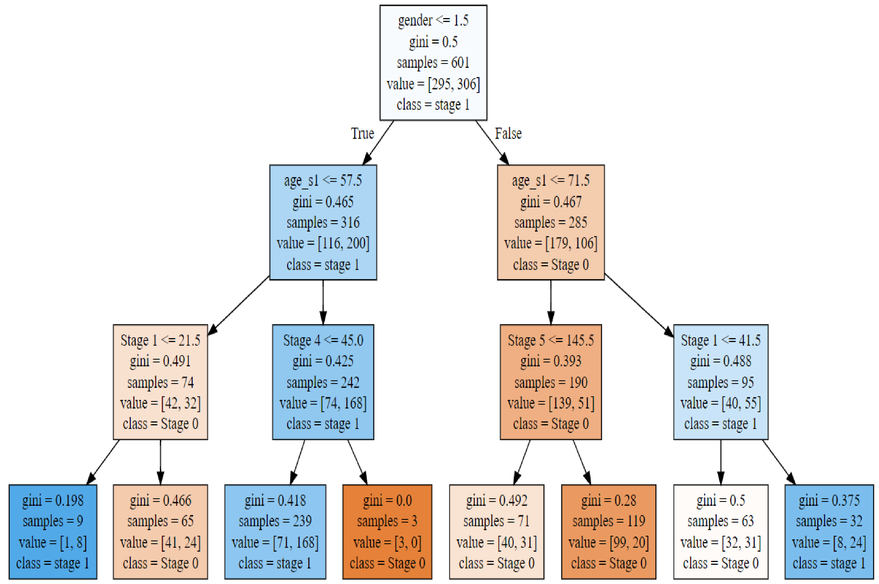


Association Rule Mining

Items	Consequent	Support	Confidence	Lift
Angina, CABG, MI	CHD	0.054	1	3.726
CABG, Angina	CHD	0.056	1	3.726
CABG, MI	CHD	0.08	1	3.726
PTCA,CHD	MI	0.11	0.994	2.906
Angina, PTCA	MI	0.084	0.992	2.9
Angina, PTCA, CHD	MI	0.076	0.991	2.898

Modeling Results

For our model, we ran a Decision Tree Classifier model, with Stages 0-5, age, race, and gender as the independent variables, and Cardiovascular Disease as the dependent variable.



- Along with our main model of Decision Tree, we also decided to run a Logistics Regression Model and a Random Forest model.

Decision Tree Model:

```
[[72 34]
 [29 66]]
```

Sensitivity = .69
Specificity = .68
Accuracy = .69

Logistic Regression:

```
[[68 38]
 [32 63]]
```

Sensitivity = .66
Specificity = .64
Accuracy = .65

Random Forest Model:

```
[[73 33]
 [31 64]]
```

Sensitivity = .67
Specificity = .69
Accuracy = .68

Conclusion

- Irregularities in sleep patterns can be used to predict the development of cardiovascular diseases.
- There are correlations between individuals with multiple cardiovascular diseases and their sleep pattern.

Thank you Dr. Rupesh Agrawal for the sleep dataset.

Team 12

Dill or No Dill?
Team: BUAL Mafia

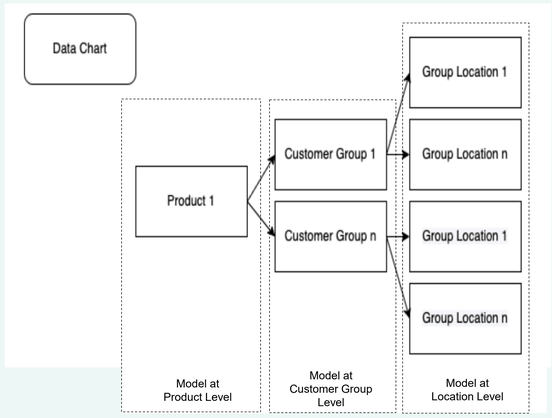
Alex Hernandez, Emma Parkhurst, Hannah Meehan, Jarrett McMeans, Shuhua He
Auburn University BUAL 5860 Capstone Project – Dr. Pankush Kalgrota

Introduction/Problem

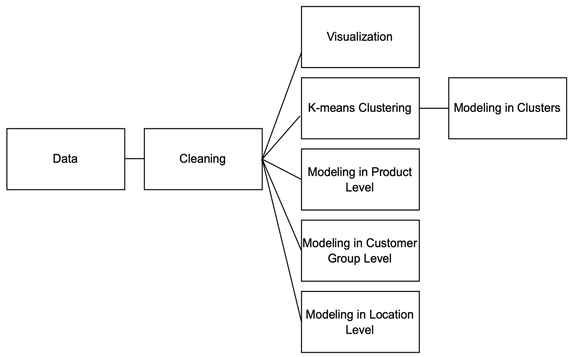
- We are partnering with a consultant company, Decision Spot, to build a sales forecasting model for a Pickle Jar Company.
- Our goal was to create a model for the next 12 months to predict their sales.
- Our model will deepen the understanding of product demand and gain more insight into the seasonality and quantity sold of each product.
- A challenge we faced was that with more than 520 products, not all of them behaved similarly. Therefore, we used our own iterative approach to develop models for them.

Variables

- The data from our model is from July 2014 to July 2017, where the company sold 528 different products to 836 customers.
- The chart below demonstrates how we organized our variables.

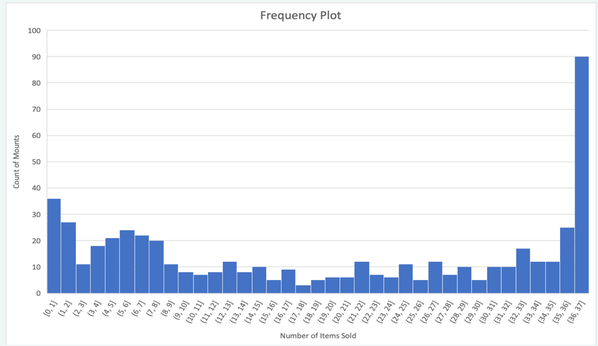


Method Flowchart

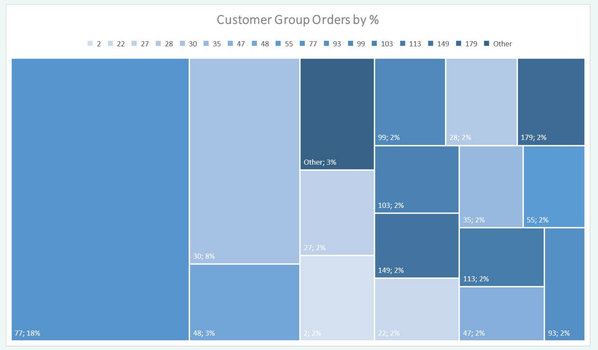


- Found that two products (284 and 368) were only sold within the first year, we removed them to run our models.
- In total 155 original products were sold for less than a year indicating that they were either discontinued or a relatively new product.

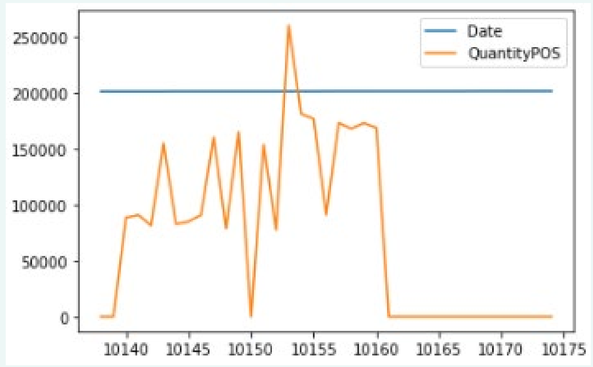
Analysis



- This histogram shows the frequency of how many months a product was ordered.
- The highest frequency represents products that were sold all 37 months, 90 out of 528 products fall into this category. They represented 17.05% of the company's products.



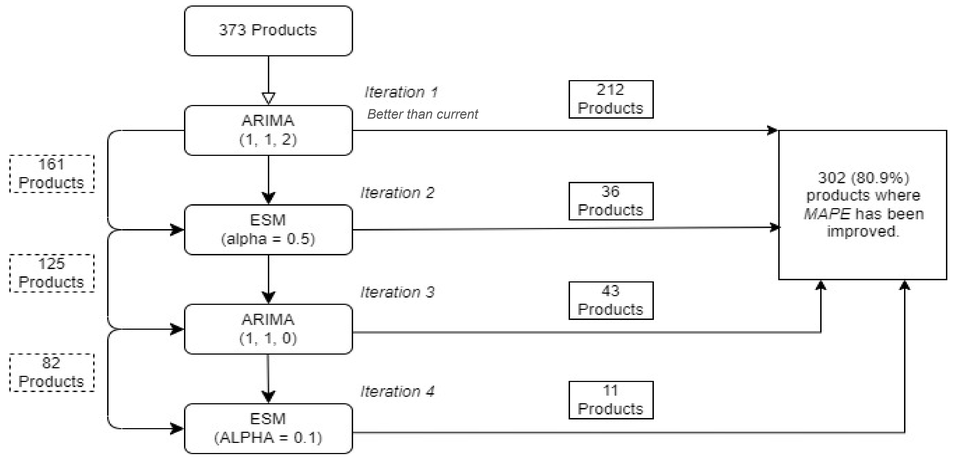
- This visual displays the customer groups and their percent of the total quantity of products ordered.
- 16 customer groups ordered 97% of the total quantity of products sold over 37 months.



- This line chart allowed us to determine our parameters for the ARIMA model and that pre-modeling differencing was not needed.

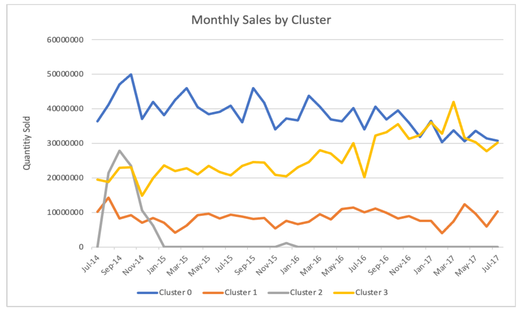
Results

- The training set contained the first 25 months of data for each entity and the testing set contained the last 12 months of data.



Product Level	Model	# of Products	% of Products	Average MAPE	Current MAPE
Iteration 1	ARIMA (1, 1, 2)	212/373	56.8%	0.68	1.98
Iteration 2	ESM (α = 0.5)	36/373	9.6%	0.52	0.91
Iteration 3	ARIMA (1, 1, 0)	43/373	11.5%	0.42	0.79
Iteration 4	ESM (α = 0.1)	11/373	2.9%	0.42	0.80
Totals	-	**302/373**	**80.90%**	**0.51**	**1.12**

Results



Level	ARIMA	ESM	Current
Cust Group	0.72	0.86	0.94
Cust Loc	0.71	0.82	0.77
Clustering	0.19	0.25	0.44

Conclusion

- We used Holt's method, an Exponential Smoothing Method, to predict sales for products with trends since we found that the ARIMA ran better without trend being considered.
- We compared our predictions against the current model the company is operating with using MAPE and MSE.
  - We cycled through 4 iterations for the Exponential Smoothing Method Model.
  - Our model resulted in a better MAPE value in 302 of the 373 products or 80.9%.

Acknowledgements

We thank Mr. Mohit from Decision Spot LLC for sharing this dataset with our team.

Team 13

Winner – Third Place

Do Photos Determine What You Eat?

Predicting the Popularity of a Restaurant Based on Image Analysis

The Miner League: Frank Hudgins, Riley Spengeman, Mary Koch, Kyle Travelstead, Grace Crosson

Faculty Advisor: Dr. Pankush Kalgotra

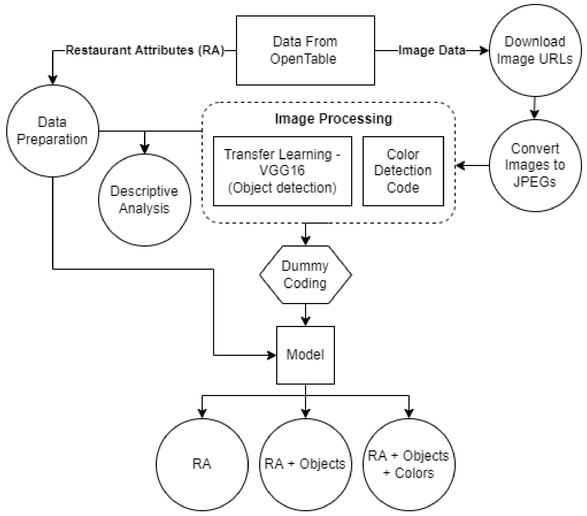
Introduction

- **Overview:** Online restaurant reservation service companies allow customers to skip the line and make an online reservation at their favorite places.
- **Problem:** The problem we are researching focuses on what aspects of a restaurants profile affect how many reviews it gets. We are using the number of reviews as a determination of popularity.
- **Purpose:** This project is important because restaurant reservation service companies are very popular and help people find restaurants to dine at. We decided to analyze color and object detection to test its correlation with popularity based on reviews. This can help restaurants boost traffic on the OpenTable platform.

Dataset Description

Name of Variable	Description	Count
Popularity	Based on review count.	
	Reviews < 300: Less	1,487
	Reviews 300 – 1000: Moderate	1,498
	Reviews > 1000: High	1,714
Region	Regions of the U.S.	
	Midwest	717
	Northeast	862
	Northwest	140
	Southeast	832
	Southwest	1,140
	West	827
Cuisine	Genre of food	
	American	2,076
	Asian	333
	European	1,053
	Latin	374
	Mediterranean	292
Dining Style	Restaurant environment	
	Casual Dining	2,584
	Casual Elegant	1,534
	Fine Dining	499
	Elegant Dining	56
	Home Style	8
Review count	Number of reviews per restaurant	1182 avg
Price range	Range of price per dish	
	$30 and under	2,807
	$31 to $50	1,537
	$50 and over	355
Color	Level of colors that appear in image; Extracted Variables	
Object	Objects in images; Extracted Variables	

Method



Object and Color Detection – Image Processing

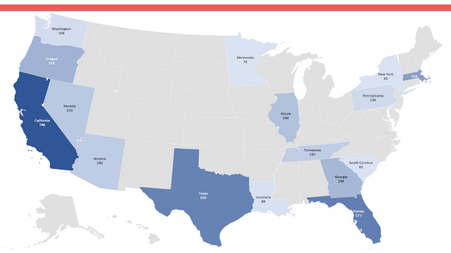
Color Detection – RGB Values (128X128) – Euclidean distance from 10 basic colors
Object Detection – Transfer Learning - VGG16 Pre-Trained Model

Color Detection Pixel Count:
Black Pixels: 388
White Pixels: 9708
Red Pixels: 17
Pink Pixels: 0
Blue Pixels: 0
Yellow Pixels: 0
Brown Pixels: 3721
Orange Pixels: 1386
Green Pixels: 144
Purple Pixels: 1020

Object Detection:
Object 1: Pizza (18.31%)
Object 2: Menu (12.74%)
Object 3: Plate (11.79%

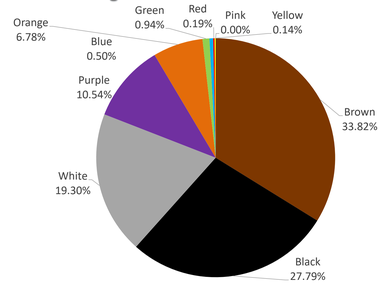
Top 10 Image Objects:
Restaurant: 31.24%
Plate: 7.54%
Patio: 4.13%
Dining Table: 2.71%
Grocery Store: 2.34%
Bakery: 1.95%
Meat Loaf: 1.44%
Pizza: 1.33%
Hot Pot: 0.96%
Confectionery: 0.80%

Analysis

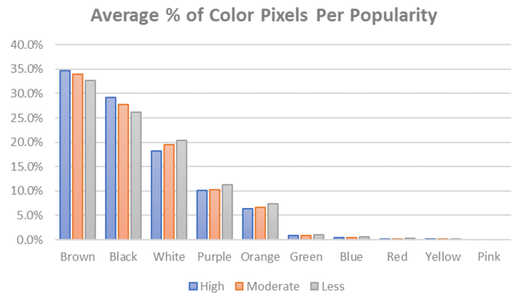


- This map above represents number of restaurants per state

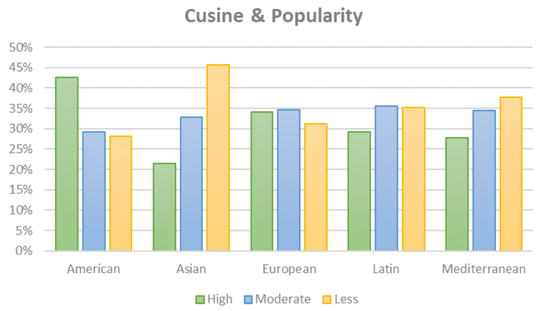
Color Detection Percentages



- The pie chart above represents detected colors in image analysis



- The bar chart represents the breakdown of color percentage per each popularity level

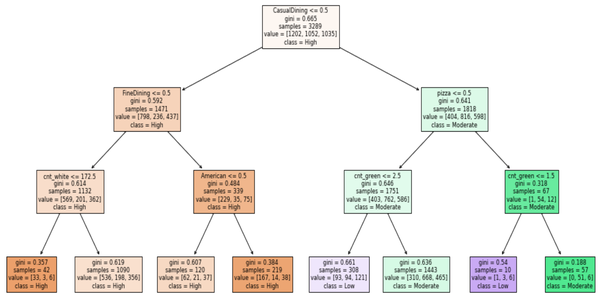


- The bar chart above represents the popularity between different restaurant cuisines

Model Creation

- Y-variable: Popularity (High, Moderate, Low)
- Training Data: 70% of data (randomly selected)
- Model – Decision Trees, Fixed Depth of 17
- Model 1: Just Cuisines, Dining Styles & Price Ranges (Accuracy: 50.21%)
- Model 2: Model 1 + Objects (Accuracy: 56.81%)
- Model 3: Model 2 + Colors (Accuracy: 85.25%)
- Best Model: Model 3

Model Evaluation



		Predicted		
		High	Moderate	Low
Actual	High	464 (90.63%)	31 (6.05%)	17 (3.32%)
	Moderate	59 (12.74%)	367 (79.27%)	37 (7.99%)
	Low	30 (6.90%)	34 (7.82%)	371 (85.29%)

Results

- During our model evaluation and analysis, we tested three different types of models: decision tree, random forest, and logistic regression.
- We selected decision tree model because of its simplicity and high performance.

Limitations

- Year that the restaurants opened
- How long the restaurants have been on OpenTable
- Demographics of the people writing the reviews
- Obscure objects being detected when doing analysis
- "Cuisine" variable had 5 different variables but really should've been many more

Conclusion

- Our results concluded that colors are by far the most important variable when looking at images from the OpenTable website.
- When we ran the model without color there was a 30% decrease in accuracy.
- While we found color to have the most impact on accuracy, it was interesting to note that casual and fine dining were both variables that impacted our results.
- Reservation booking platforms can utilize our model by understanding what colors and objects are found in images and the impact that they have on the number of reviews each restaurant received.

Acknowledgements

- We thank Sijun Liu for providing us with this dataset.

Team 14

got sleep?

Detecting Risk of Myocardial Infarction Through Sleep Patterns

Tree Climbers: Erin McGarry, Anna-Katherine Killian, Alayna Priebe, Ziqi Huang, and Eddie Loew
Faculty advisor – Dr. Pankush Kalgotra

AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS

Introduction

- Data collected from the The Sleep Heart Help Study (SHHS)
  - organized by the National Heart Lung & Blood Institute.
- Monitored during sleep cycle to detect 5 sleep stages

According to the CDC**, 1 in 4 deaths in America are from heart disease.**

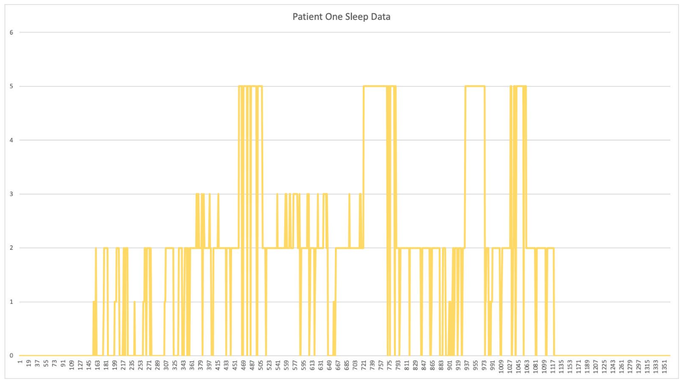
Problem

Can sleep patterns identify issues related to heart? Our objective is to create a machine learning model to diagnose Myocardial Infarction.

With this analysis, and next steps, we hope that awareness can be spread and preventative measures of MI can be taken early on

"Approximately, every 40 seconds, someone in the US will have a myocardial infarction." –American Heart Association

Sleep Data

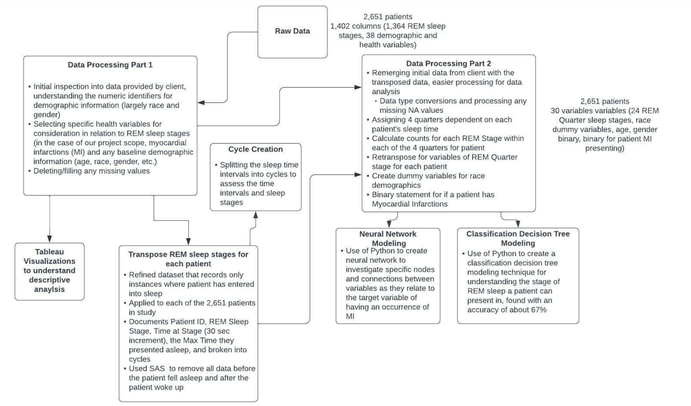


Data description

- **2,651 total patients**

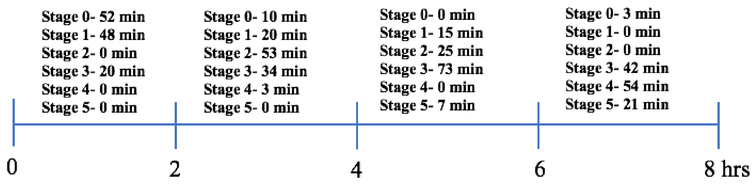
11.3% of patients had MI (one or more times)

Average age of patients with MI= 68.43



Methodology

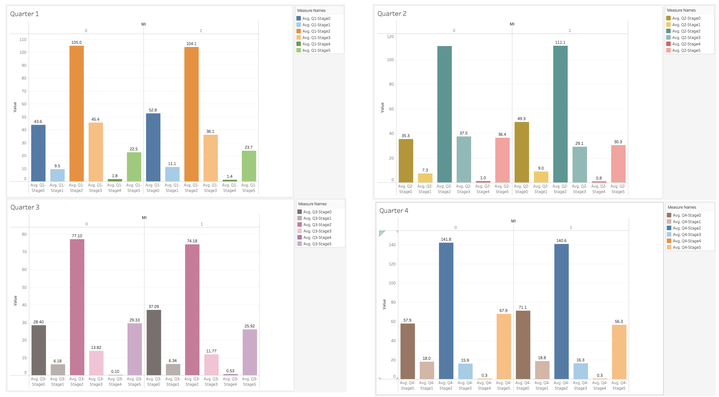
- Computed new sleep variables (aggregated)



Ex. Patient Sleeping 8 hours- broken down into 4 quarters and how long they were in each stage during each quarter

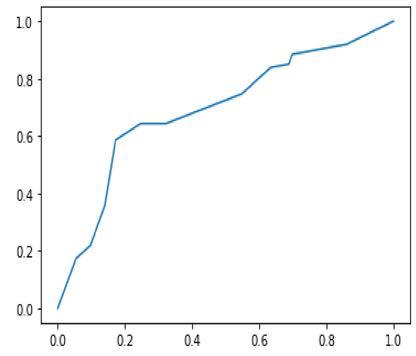
Descriptive Analysis

	Myocardial Infarctions in Patients (MI)	Patient with MI (1 or more)	Variable Type
		300/2649= 11.33%	**Binary (1 if MI occurred, 0 if no MI)**
AGE	Adults= 1486 Seniors(65+)=1167	Adults=93 Seniors= 207 *Avg Age= 68.43*	Integer
GENDER	Female= 1425 Male= 1226	Female= 111 Male= 189	Numeric
RACE	White= 2311 Black= 181 Other= 159	White= 258 Black= 29 Other= 13	Numeric



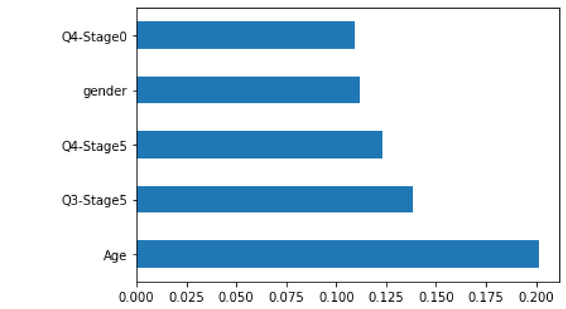
Decision Trees

	True Class: Positive	True Class: Negative
Predicted as: Positive	65	23
Predicted as: Negative	36	56



Accuracy= 67.04%
Sensitivity= 62.63%
Specificity= 71.05%

AUC= .6950



Above are the most important statistics in identifying risk of MI. The x-axis shows the level of importance, showing age to be the top variable

Limitations and Conclusions

- Other variables such as blood report, clinical characteristics can improve the performance
- Can help lead to early detection of risk
- Advise the normal user to see the doctor

Our model can be used as a part of an app tracking the sleep that notifies user if their patterns show irregularities or risk

https://professional.heart.org/en/science-news/heart-disease-and-stroke-statistics-2022-update
https://www.cdc.gov/heartdisease/facts.htm

Acknowledgement – We thank Dr. Rupesh Agrawal for providing us access to this dataset.