Prateek Mechineni

SID: 437736660 | psmechin@syr.edu

MSADS Portfolio Milestone

Syracuse University

Introduction:

The Applied Data Science program at Syracuse University is designed to give students the skills necessary to become professionals who can analyze data to solve problems.

These skills include:

* Collect, store, and access data by identifying and leveraging applicable technologies
* Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle
* Apply visualization and predictive models to help generate actionable insight
* Use programming languages such as R and Python to support the generation of
* actionable insight
* Communicate insights gained via visualization and analytics to a broad range of
* audiences (including project sponsors and technical team leads
* Apply ethics in the development, use, and evaluation of data and predictive models

(e.g., fairness, bias, transparency, privacy)

Over the duration of the program, I was able to learn and apply each of the learning objectives and skills to achieve a deeper understanding of data science.

The projects include:

* IST 772: Vaccine Analysis Final Project
  + Creating actionable insight across a range of contexts
  + Using R
  + Applying ethics
  + Communicating Insights
* IST 707: Wine Data Analysis
  + Collect, store, and access data
  + Applying visualizations
  + Using R
  + Predictive Modeling
  + Communicating Insights
* IST 652: Flight Delay Data
  + Collect, store, and access data
  + Applying Visualizations
  + Using Python
  + Predictive modeling

IST 772:

Goals: For this project, I was analyzing California School Vaccine Data to then help write a report to a staff member in a state legislator’s office on how to allocate assistance to school districts to improve vaccination rates and their reporting compliance. I used R, predictive models, and analytical models.

I first explored the general vaccine data and create some visuals over the time series data:

A picture containing diagram, text, plan, line

Description automatically generated

The Dip in vaccine rates around the same time in 1980-1990 was very perplexing but that is where historical and political context come into play. That time period was the same time period that Anti-Vaccination information was spreading like crazy and explained the dips as parents saw the information and assumed it would hurt their children.

I then looked at the discrepancies between public and private school reporting in vaccination data:

A screenshot of a computer

Description automatically generated with low confidence

The data indicates that vaccination information was reported by 84.7% of private schools and 97.4% of public schools. One possible reason is that the U.S. Public schools receive funding from the states and are dependent on accurate and timely reporting for financial reasons while private schools usually do not have that burden.

A screenshot of a computer

Description automatically generated with medium confidence

When I looked further into the correlation between all the vaccines, we can see that when you are missing any one of the vaccines, you are highly likely (90% or higher on each vaccine to vaccine correlation) to be missing the other vaccinations. It shows that when a person is not receiving one vaccine, they are very likely to not have any of the other vaccines. Once again, it shows that not all the information can be looked at as pure data or numbers-wise and that the context and timing of the data that you are looking at is as important.

I then looked at the correlation between reporting and the various factors to see what factors affect reporting the most.

A screenshot of a computer

Description automatically generated with medium confidence

The results indicate a weak correlation between the variables and the dependent variable. However, there is a high correlation, as expected, between child poverty, free meal, and family poverty as these are logically linked.

A screenshot of a computer

Description automatically generated

By utilizing a frequentist approach, we observe that the poverty values and percentage of free meals exhibit high p-values and fail to reject the null hypothesis, suggesting that these independent variables do not significantly impact whether the district reports complete data. On the other hand, both the enrolled status and total schools demonstrate p-values below the 0.05 threshold, indicating that they could potentially reject the null hypothesis and show an influence on the determination of the dependent variable’s completeness in reporting.

Moving forward, we can perform a Bayes Factor analysis to ascertain whether we arrive at similar conclusions.

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, font, handwriting, black and white

Description automatically generated

The results from the Bayes Factor analysis are similar to those obtained from the standard logistic model. The number enrolled and total schools in the district are the most probable predictors of complete reporting. Their HDIs do not cross zero but are in close proximity to zero. Conversely, factors such as the percentage of family poverty, percentage of child poverty, and percentage of free meals served do cross zero and lack statistical support, thereby failing to reject the null hypothesis, similar to the frequentist model.

I then analyzed which variables predict the percentage of all enrolled students with completely up-to-date vaccines with the correlation and both freq uentist and bayes models.

A screenshot of a computer

Description automatically generated

Similar to before, we get a low correlation from the independent variables to the dependent (up to date) variable.

Frequentist Model:

A screenshot of a computer

Description automatically generated

The overall performance of the model appears satisfactory, as indicated by the high F value greater than 1, which suggests that the model performs well. This conclusion is further supported by the small P-value of the overall model, which is approximately 5.028e-13, indicating that the model is statistically significant, and we can reject the null hypothesis that the independent variables have no effect on predicting the percentage of students who are up to date on their vaccines.

Furthermore, after analyzing the different variables, it appears that all of them, except for PctChildPoverty, show some degree of significance. The Intercept, Percent Free Meal, PctFamilyPoverty, and Total schools and Enrolled have P-values that are below the 0.05 alpha level, implying that they all have a significant impact on the dependent variable. On the other hand, PctChildPoverty is the only factor that is not statistically significant since the pvalue is above an alpha of 0.05 and we cannot reject the null hypothesis for this variable. Therefore, we can conclude that all the factors except PctChildPoverty have an effect on the dependent variable.

When ran without PctChildPoverty, the F score does increase which is also an indicator that it does not affect up to date vaccinations.

A screenshot of a computer

Description automatically generated with medium confidence

The model does produce consistent results with our frequentist model. The PCTChildPoverty does pass through 0 so its not a good indicator of being up to date on vaccines. All the other factors do not and they can all reject the null hypothesis.

A screenshot of a computer

Description automatically generated with medium confidence

It is worth noting that the overall model has a Bayes factor of 5276913702 ±0.01%, which indicates significant support for the model under the Bayes approach. As such, we can reject the null hypothesis and conclude that the variables except have a statistically significant impact on the percentage of students who are up to date on their vaccinations.

The final part of the analysis included look a the percentage of students who have belief exemptions and what factors affect those students.

A screenshot of a computer

Description automatically generated

We see that the major independent variables all have a negative correlation to the dependent variable. Subsequently, we will conduct a linear model to evaluate the predictive ability of the dependent variable.

A screenshot of a computer

Description automatically generated with medium confidence

Upon reviewing the linear model, it becomes apparent that all variables, with the exception of Total Schools under the system, are significantly relevant in predicting the likelihood of students presenting a vaccination exemption. In the case of Total Schools in the district, we cannot reject the null hypothesis that it leads to a student presenting an exemption. Conversely, the other variables have P-values that are below the 0.05 alpha level, allowing us to reject the null hypothesis.

A screenshot of a computer

Description automatically generated with medium confidence

The Bayes model indicates that the Total Schools variable contains zero. Similar to the frequentist test, we would fail to reject the null hypothesis that the total number of schools has no significant impact on whether or not a student presents a vaccine exemption. On the other hand, the other variables have HDI ranges that do not cross zero and display significance. Therefore, we would reject the null hypothesis and contend that they can predict the likelihood of a student presenting an exemption.

Lastly, we need to verify the overall model’s Bayes factor.

A screenshot of a computer

Description automatically generated with medium confidence

As we can see from the outcome of the Bayes Factor, the model show ratio of 6.025845e+13 and thus the model is significant on its own to reject the null hypothesis and say that the dependent variables are showing an effect on the dependent variable.

Overall Conclusion and Recommendation To State Legislators on how to allocate funds to improve vaccination rates and reporting compliance:

The initial model reveals that schools are more likely to report vaccination statuses when student populations are grouped together, but this likelihood drops as the number of schools in the district increases. It appears that there is a threshold where schools must report the vaccine status of their students to the state, and as the number of schools in the district grows, they enter this threshold. Therefore, the data suggests a need for better reporting resources for districts with more schools, such as improved systems for centralizing information or utilizing a state reporting system like Pennsylvania’s.

Consolidating schools could also be a potential solution, as one new school causes a -.18 log drop, but an additional 100 students essentially cancels out the effect, based on the model where each additional student increases reporting likelihood by .0018 log. This could be used as a consolidation threshold for small districts.

To improve vaccination rates, reducing childhood poverty is the most impactful area to focus on, as the model shows an increase in the up-to-date vaccination rate for each percent drop in childhood poverty rate within the school district. Additionally, as childhood poverty rates decrease, the number of students who present with belief objections also decreases.

However, it is important to take a closer look at family poverty rates, as the model suggests that as these rates decrease, we see the opposite effect on vaccinations within the school population. This may be due to the recent trend among more affluent individuals to suspect a link between autism and vaccination, which has skewed the data. It could also be that the school districts included in the model are from more affluent areas, leaving out more urban districts and skewing the data in this way. Nonetheless, as poverty rates decrease and people have more disposable income for healthcare, we may see improved vaccination rates, although this is not currently supported by the models possibly due to increased misinformation about vaccines.

In summary, focusing on childhood poverty rates and total enrollment numbers are the most likely factors to lead to higher vaccination and compliance in reporting based on the provided data.

IST 707:

Goals: For this project, I was tasked with finding my own data set and using what I had to create some predictive models that can provide actionable insight for a business. I chose a red wine dataset and wanted to understand the difference between various qualities that affect the wine. I looked at what qualities of the wine are the most important to be considered high-end vs low-end. I also wanted to predict a wine’s rating based on just the technical factors since wine is still a subjective taste to most people. Finally, I wanted to use the insights gained to help a company market the wines to increase profits as a wine company overall.

Data Selection/Collection:

For many students, Kaggle is the first database of datasets that they gravitate to due to the ease of searching for and the details of the dataset. While it is a great resource, the quality/accuracy of the data is variable since it is a public database. To help avoid this, I used the University of California, Irvine Machine Learning Repository to select a wine dataset that has been vetted and used by many in an academic setting. The importance of selecting a dataset that can be used academically allows me to create real and actionable insight with industry-standard information.

Overall EDA:

A screen shot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Attribute Descriptions:

* Fixed acidity: Most acids involved with wine or fixed or nonvolatile (do not evaporate readily)​
* Volatile acidity: The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste​
* Citric acid: Found in small quantities, citric acid can add 'freshness' and flavor to wines​
* Residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter
* Chlorides: the amount of salt in the wine​
* Free Sulfur Dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine​
* Total Sulfur Dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine​
* Density: the density of wine is close to that of water depending on the percent alcohol and sugar content​
* pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale​
* Sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant​
* Alcohol: The percent alcohol of the wine

Quality Distribution of Wines:

A picture containing text, screenshot, diagram, plot

Description automatically generated

The quality distributions of the wines ranged from 1-10 with 1 being the lowest and 10 being the highest as deemed by professional wine tasters. However, the red wines in our datasets (1600 wines), all fell within the 3-8 range. My initial thought was that it is possible that extremely low-quality wines are not made in big quantities since the company and consumers know they will not be bought. Extremely high-quality wines are possibly made more by smaller companies and are not as mass-produced as some of the wines in the middle range of the lists since it takes more time and effort to create high-quality wines.

Since there were no observations within the 1-2 and 9-10 quality ranges, I removed this from our dataset and grouped the rest into three categories, Bad (3-4), Medium (5-6), and Good (7-8). This lowered the decision-making into three variables as opposed to six, which improved accuracy in our models drastically. Companies also wanted to market wine in 3 pricing groups of cheap, medium, and high-end to streamline marketing campaigns, we determined it was unnecessary to have so many categories. Additionally, I scaled the data to make it more usable and get the data points closer together, which led to improvements in our modeling accuracy.

I then looked at the correlation between all the variables:

A picture containing text, screenshot, font

Description automatically generated

A picture containing screenshot, line, colorfulness, graphics

Description automatically generated

Between the two correlation plot, the biggest correlation between Quality and another variable was Alcohol suggesting that higher alcohol contents leads to higher quality. It also suggests that high volatile acidity and quality are inversely related. This is also exemplified in the boxplot chart, comparing the quality score to each input variable.

A picture containing text, diagram, screenshot, plan

Description automatically generated

Predictive Models:

I started the analysis by running a decision tree as seen below. Our first tree resulted in a lackluster 57.4% accuracy, which led us to conclude that the data needed to be cleaned further to give us a more accurate representation. Originally, the decision tree was pointing to the individual wines rather than wine quality, so we decided to change that to reflect what we were looking for, that being quality.

A picture containing text, diagram, sketch, font

Description automatically generated

To improve our accuracy, I decided to scale the data to pull the data points together, as well as remove quality scores 1,2,9 and 10 since none of the wines scored in those categories. For ease of analysis, I then grouped the data into three categories, “Bad” (3-4), “Medium” (5-6) and “Good” (7-8). I then ran our decision tree again and received much more favorable results. 

A picture containing text, diagram, screenshot

Description automatically generated  
A screenshot of a computer

Description automatically generated with low confidence

By scaling and splitting the data into more generalized ranges, we were able to improve accuracy, from 57.4% to 85.21%. One major takeaway was the impact of fixed acidity and wine quality. The decision tree revealed that when fixed acidity was lower than .41, it always resulted in poor wine quality.

Random Forest:

I ran a Random Forest Model and again concluded that alcohol and volatile acidity had the most impact on wine quality. We used tuneRF to find the optimal value for mtry and that was determined to be 2. I tried different mtree sizes and minsplit sizes but that did not change the accuracy significantly. Once again, I assume that, since the data was scaled and already concentrated in the medium range, the data is already processed and changing the variables does not affect the prediction. The RF model also confirmed our EDA that alcohol and volatile acidity were the key factors in predicting the quality of the wine. Overall, I tried multiple attempts but, in the end, I was able to get accuracy as high as 89.06%.

A picture containing text, font, line, screenshot

Description automatically generated  
A screenshot of a computer

Description automatically generated with medium confidence  
A picture containing text, screenshot, number, font

Description automatically generated  
  

SVM:

Support Vector Machines is an algorithm that seeks the optimum separating hyperplane between the variables by maximizing the margin between the classes' closest points. When running this model, I kept all the factors the same but changed the kernel each time to get the most accurate result. What I found was that linear and radial produced the most accurate predictions, while Poly was the least accurate.

A picture containing text, screenshot, font, number

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Naïve Bayes:

I also ran a Naïve Bayes model and was able to get an accuracy of 77.19%. Despite changing the model, I could not improve our accuracy despite running it multiple times. I believe the reason for this is because we had already scaled our data and that already reduces over/underfitting. Due to this, the smoothing variable could not improve the model.

A picture containing text, font, screenshot, receipt

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with medium confidence

KNN:

I also conducted a KNN prediction on the dataset. With KNN, I can specify how many neighbors/clusters I want the algorithm to use while trying to predict. I used 3,5, and 7 as good starting points and found that they provide the best accuracy. I thought that maybe as the number of neighbors goes up, the accuracy would go up as well, but that was proven incorrect. 3 had a higher accuracy than 5 but 7 had a higher accuracy than both. It was not what I expected but, in the end, I was able to get an accuracy of 84.69% which was on par with our other predictive methods.

A picture containing text, screenshot, font

Description automatically generated

A picture containing line, plot, diagram, parallel

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Overall Conclusion:   
In conclusion, I was able to determine that alcohol and quality have a strong positive relationship, and acidity and wine quality are inversely related. The results of the predictive model analysis showed that Random Forest had the highest accuracy of all the models we analyzed.

From a business perspective, when marketing higher-quality wines, the push should be on the amount of alcohol in the wines. Companies can market lower-quality wines as wines that not only cheap out on the taste but also cheap out on the amount of alcohol in each bottle. Meanwhile, lower-quality wines with less alcohol can market themselves are light or healthier wines with fewer calories and alcohol compared to more expensive wines. They can also be marketed as cooking wines since cooking wines have to burn off the alcohol and expensive wines are generally not used for cooking. Medium-quality wines however have the most number of customers to compete for with very few differentiating factors. The best recommendation would be to focus less on taste and alcohol quality and more on branding and packaging. Creating differentiating factors other than the taste of the wine is what can make a medium-quality wine succeed over others.

IST 652:

Goals: For this project, I was analyzing flight data from the past year to see if I can analyze key factors that determine flight delays or if I can create a model that can predict if a flight will be delayed or not.

Business Questions to Answer:

* What airline has the most delays?
* Which airport has the most delays?
* Which day has the most delays?
* Which airports are best to fly in and out of?
* Does flight distance affect delays?
* Does flight time affect delays?
* Can I create an ML model to help predict if flights are going to be delayed or not?

Data Collection:

I used a Kaggle data set that contains all the flight data of US domestic airlines which was a CSV files that contains 539383 instances and 8 different features:

Text

Description automatically generated with low confidence

Data Cleaning:

I first performed some general analysis on the dataset and decided to split my dataset into two data frames with one that contained all the numerical columns and one that contained all the object columns. This allowed me to create graphs with Seaborn and Matplot and answer some of the key questions more easily.

To create the ML model, I had to split the data into a training set and a test set since ML models need a training set to help accurately predict models. I then used Python’s Scikit-Learn library to import ML models and tools to create and test the various types of models. I used Oridinal Encoding to process the data. I did a test run, a Decision Tree model, a Gaussian model, and a KNeighbor model. I chose these since they are well incorporated into the SciKit-learn library and each one is a slightly different model.

Metadata:

* ID: Just list of items
* Airline: Type of Arline
* Flight: Flight Number
* AirportFrom: 3 letter airport code
* AirportTo: 3 letter airport code
* DayofWeek: Number to represent day of the week
* Time: Time of flight (Add number of min in column to Midnight of that day to get time. Ex. 15 means 12:00am + 15min = 12:15am flight time.
* Length: Time of flight in Min
* Delay: 1 for delay, 0 for no Delay.

Answering Business Questions: To answer my key questions, I transformed the data into graphs to give a visual analysis.

1. What airline has the most delays?

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

1. Which airport has the most delays?

Chart, histogram

Description automatically generated

1. Which day has the most delays?

Chart, histogram

Description automatically generated

1. Which airports are best to fly in and out of?

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

1. Does flight distance affect delays?

Chart, histogram

Description automatically generated

1. Does flight time affect delays?

Chart, histogram

Description automatically generated

Looking at the business questions I wanted to answer, I was able to see that WN (Southwest) and DL (Delta) were the airlines with the most delays. Atlanta airport was shown to be the airport with the most delayed flights. For days of the week, Wednesday and Thursday had the greatest number of delays but the differences from the other days of the week were not drastic. Once again Atlanta airport is shown to be the worst flight to fly in and out of.

Interestingly, I found that flight length doesn’t seem to be influencing delays but flight time is a major indicator of flight delays. As time goes on later in the day, the more delays there are. I assume that during the start of the day that there are fewer issues in the airport to cause delays. Once a delay occurs, I think it creates a snowball effect that pushes other flights into delays as well.

Predictive Models:

Test Run with Dummy Classifier:

Text

Description automatically generated

Decision Tree:

Text

Description automatically generated

GaussianNB:

Text

Description automatically generated

KNeighbors:

Text

Description automatically generated

For the ML analysis, the dummy model came out with about 50.5% accuracy which is what we are expecting for our baseline models. In the end, the KNeighbors Classifier was able to give the best accuracy but it was only around 61%. Overall, I think I needed more data columns such as weather or the reason why the flight was canceled, and then incorporated that into my ML models to get better accuracy and precision scores.

Conclusion:

During this program’s classes, I learned how to gain insights from data and make recommendations to businesses and other organizations using visuals, statistical analysis, machine learning, and other analytical tools. I would like to thank all my professors, classmates, and advisors for helping me through this program and giving me an opportunity to succeed. I am excited to dive into the world as a career data scientist and hope to one day work for an NBA as part of their analytics program!