Project Executive Summary

Project Executive Summary

Introduction:

Life insurance is an old and important protective financial tool for many people in the US, and is a large sector of the insurance industry (trillions of dollars of benefits managed by companies in the states). However, the payouts from a life insurance policy tend to be much, relatively, higher than the premiums compared to other forms of insurance. Add to this the costs associated with underwriting in a manner that is quick, efficient, and (hopefully) accurate, and it is not hard to see that life insurance companies would benefit from the power of machine learning. Machine learning itself is a powerful tool whereby one can analyze large quantities of data to make predictions on an outcome and, in the case of life insurance underwriting, could be used to either automate the process of underwriting or act as a supplement to it making risk recommendations.

The purpose of this project is to determine how machine learning tools can be used in the process of life insurance underwriting, either as a supplemental tool to hasten the process or as a means to fully automate it. Implementing machine learning could potentially cut costs and reduce inefficiencies in the process of insurance underwriting which is, traditionally, a very time consuming task that necessarily involves a degree of subjectivity. Some questions we are considering while working on this project are:

1. What type of machine learning model should be used (e.g., classification or regression?)
2. Which specific machine learning model would work best for our dataset (after determining the model type to use)?
3. How accurate can we make the machine learning model and how well can we tune it (e.g., which parameters to tune and by how much)?
4. What other parts of underwriting can we automate?
5. Can we also use machine learning to predict the outcome of a claim assuming it is legitimate?
6. How much cheaper or efficient can using machine learning be compared to traditional underwriting?
7. How much money do life insurance companies spend on administrative costs per year?
8. What characteristics of a person put them at a higher risk level?

Questions four, five, and six are more exploratory questions that we will be unable to answer concretely for this project. Instead, we will look to answer these questions theoretically, based on how successful our other questions were answered and based on some of our own independent research.

Research:

Datasets:

Project Executive Summary

There were two datasets that were used in this project, a census dataset and a dataset from Prudential Financial which contained numerically encoded data of different life insurance applicants and their risk/response rating (ordinal from one to eight).

Census:

We were given various links to different census datasets to make use of in our capstone. We chose to use the link titled "Economic Data from the Census". From there we navigated to the "Finance and Insurance (NAICS Sector 52)" page, and then finally to the link titled "Finance and Insurance: Administrative Expenses and Benefits Paid for Life, Health, and Medical Insurance Carriers for the U.S.: 2017" located under "Miscellaneous Statistics".

This data set contains aggregates of financial data used in getting a big picture overview of costs associated with life insurance companies.

Prudential Life Insurance:

This dataset was found on kaggle.com as part of a competition to see how accurately machine learning was in predicting the risk category placement (ordinally encoded variable from 1-8) given several other parameters of life insurance applicants. There are 59,381 such applicants (or records) and 126 columns (features). All of these features are summarized in the figure below:

Variable	Description
Id	A unique identifier associated with an application.
Product_Info_1-7	A set of normalized variables relating to the product applied for
Ins_Age	Normalized age of applicant
Ht	Normalized height of applicant
Wt	Normalized weight of applicant
BMI	Normalized BMI of applicant
Employment_Info_1-6	A set of normalized variables relating to the employment history of the applicant.
InsuredInfo_1-6	A set of normalized variables providing information about the applicant.
Insurance_History_1-9	A set of normalized variables relating to the insurance history of the applicant.
Family_Hist_1-5	A set of normalized variables relating to the family history of the applicant.
Medical_History_1-41	A set of normalized variables relating to the medical history of the applicant.
Medical_Keyword_1-48	A set of dummy variables relating to the presence of/absence of a medical keyword being associated with the application.
Response	This is the target variable, an ordinal variable relating to the final decision associated with an application

Figure 1 Table displaying the feature (variable) name and description as provided by Prudential. The table was taken from: https://towardsdatascience.com/life-insurance-risk-prediction-using-machine-learning-algorithms-part-i-data-pre-processing-and-6ca17509c1e

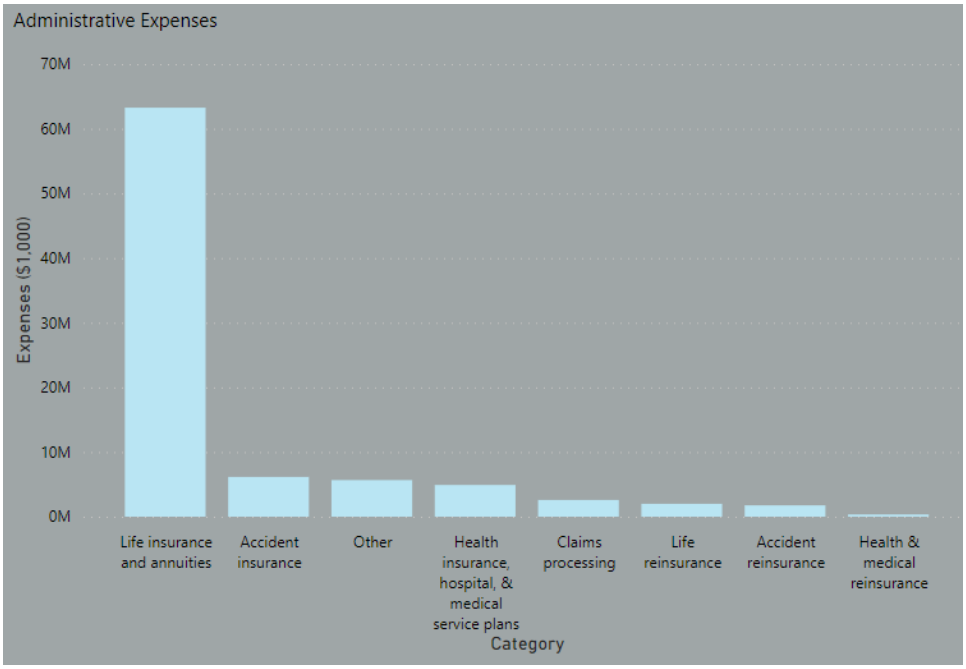
Machine Learning Models:

Since the data will be used to determine an applicant's risk level from 1-8, it was decided to use classification algorithm models. We thought it best to stay within the models already provided by the Sklearn Python library. The models tested are naïve bayes, support vector machines, neural networks, k-nearest, and random forest. These models were chosen because they all have very different unique algorithm calculations and are some of the most popular and common machine learning models used in industry today. Another reason is that some of them are known to usually produce accurate results, so we want to test if that remains true.

Project Executive Summary

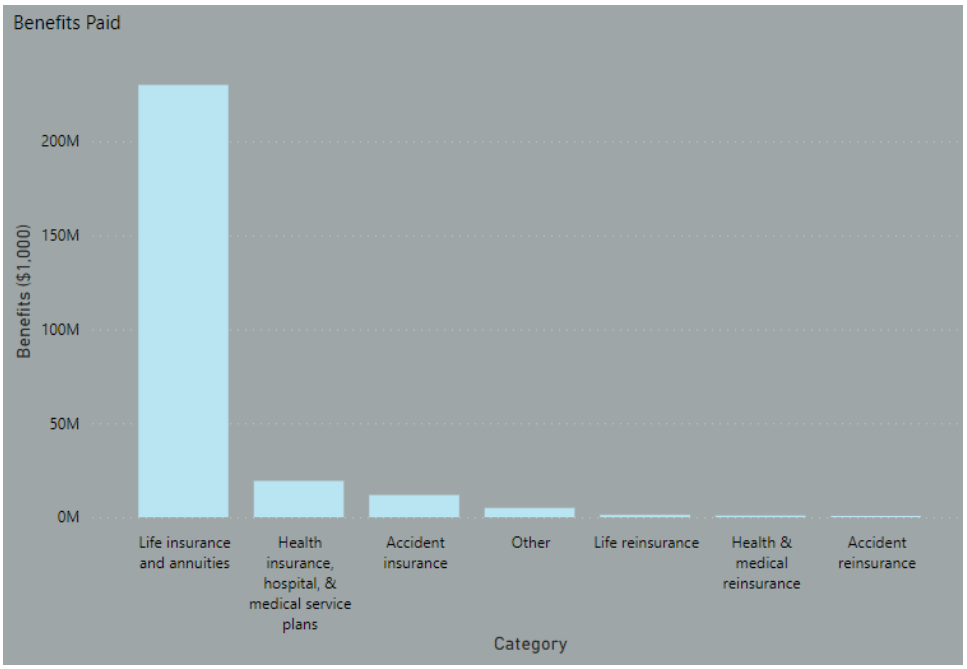
Exploratory Data Analysis (EDA):

Through our research we discovered that administrative expenses within the insurance industry totaled about $86B in 2017. As we can see in the graph below, about $63B of the total administrative expenses came from the life insurance department. This means that more than 70% of administrative expenses within the insurance industry came from the life insurance department.

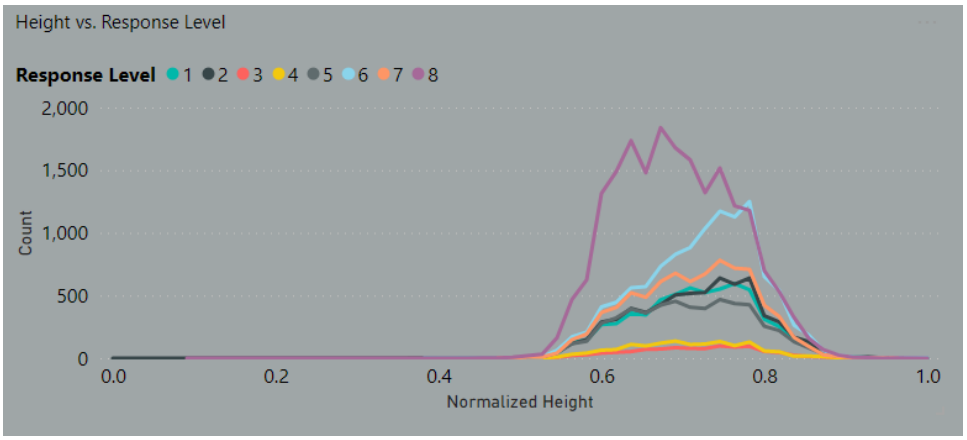


We then decided to look at the distribution of benefits paid for the insurance industry and discovered that the total benefits paid for the industry in 2017 was about $268B. As we can see in the graph below, the benefits paid for the insurance industry alone were about $230B. This means that more than 85% of the benefits paid within the insurance industry came from the life insurance department. This would explain why the administrative expenses for life insurance represent more than 70% of the overall administrative expenses in the insurance industry.

Project Executive Summary

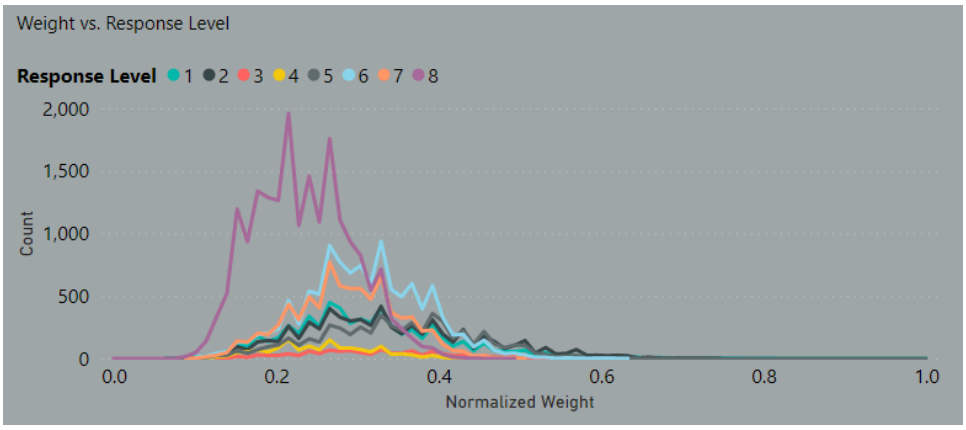


Continuing to the Prudential Life insurance data, we compared the amount of people in each risk level to the distribution of height, weight, and age. For height we found that most of the shorter people fall under a risk level of eight, and most of the taller people fall under a risk level of six or eight. This seems to suggest that being at either end of the height spectrum makes you a higher risk candidate for life insurance.

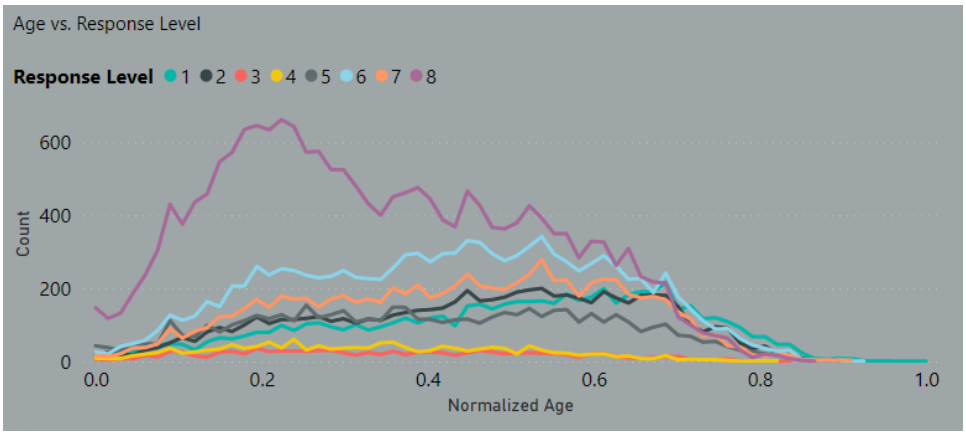


For weight we found that most of the lighter people fall under a risk level of eight, and the heavier people are pretty evenly distributed across each risk level. This data somewhat counters the typical teaching that being overweight leads to a plethora of health problems and would presumably make a person look like more of a risk to a life insurance company. Future research into how closely a life insurance company uses weight to determine a person's risk level could be interesting.

Project Executive Summary



Finally, for age we found that most of the younger people fall under a risk level of eight, and the older people are pretty evenly distributed across each risk level. Like with weight we see that this data counters the typical thought process that an older person would look like more of a risk to a life insurance company, given they are much closer to the average life expectancy. We hypothesize that younger people fall into higher risk categories because young people are most likely to not take out a life insurance policy unless they have some wort of underlying condition that could cause them to die at a relatively young age. This underlying condition could cause them to be viewed as a higher risk to a life insurance company.



Machine Learning Results:

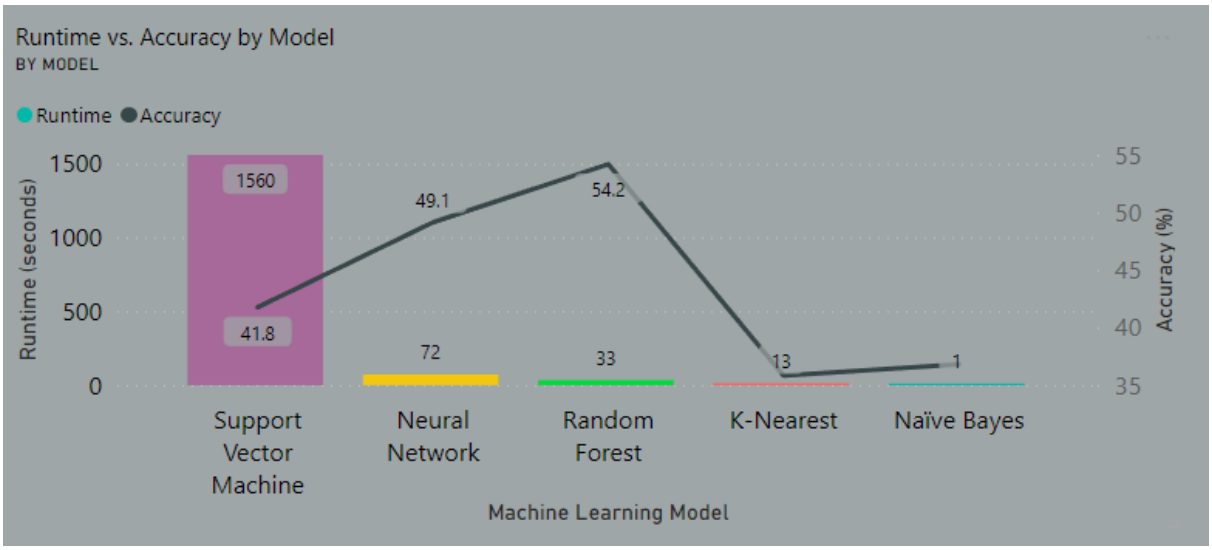
We decided to explore our data using classification machine learning models because it would be a better fit for determining and classifying an applicant's risk level from 1 through 8. The five different models implemented were: naïve bayes, support vector machines, neural networks, k-nearest, and random forest. Each model has been tested multiple times and optimized to produce the highest accuracy results. The exact tuning parameters can be found in the code files in our shared GitHub.

To run our models, we separated the first 5,000 rows of data from the initial 60,000. From the remaining 55,000, we then split it into 75% training-data and 25% testing-data and then

Project Executive Summary

ran our five different machine learning models. Once complete, we used the model to predict the "Response" (risk level) column on the separated 5,000 rows in order to simulate "new data." We then compared the result of the newly predicted column with the original "Response" column and calculated the accuracy scores which can be seen in the figure below.

We also timed the duration to run each model. The time (in seconds) is represented in the bar columns below and each model's accuracy score is represented by the line graph. Most of the models ran in approximately one minute or less. Support Vector Machine was the most time consuming at approximately 26 minutes per run. Due to its long runtime and low accuracy, this model was not chosen as the best option. The Random Forest model ended up being the winner with the highest accuracy score of 54.2% as well as a low model runtime of 33 seconds.



Conclusion:

Through our research we discovered that within the insurance industry, life insurance is by far the largest insurance type in terms of both claims and the amount of resources companies put into their administrative duties. We also discovered that for some reason, shorter, lighter, and younger people tend to be placed at a higher risk level even though we tend to think of these attributes as being associated with good health.

As for the machine learning portion of our project, we found that the best and worst accuracy for the models is driven by the two same factors. Size of the dataset as well as the large number of features is an advantage for the random forest model and a bane for the K-Nearest model. We decided to focus on classification models and discovered that the random forest model has the highest accuracy at 54.2% while only taking 33 seconds to run. Unfortunately, 54.2% isn't a high enough accuracy score to be able to rely on machine learning when it comes to something with as much money involved as life insurance. With that being said, we believe that machine learning could still be used within the underwriting process as some sort of supplemental prescreening to get the ball rolling in the right direction. We also believe that, given more time a more suitable machine learning algorithm could be found and utilized, or with more tuning random forest could become more accurate.

Project Executive Summary

Future studies could look into how much each attribute considered by insurance companies contributes to a person's overall risk level. Future studies could also look to use regression machine learning models instead of classification to see if higher accuracy scores could be attained. Finally, future studies could attempt to dive much deeper into the actual underwriting process and look to implement machine learning on specific parts of the underwriting process instead of using machine learning to take care of the entire process.