POMS. 6220 DECISION ANALYTICS

Fall 2018 Section 062

FINAL PROJECT

WHAT IS YOUR LIFE WORTH: A PREDICTIVE ANALYTICS APPROACH TO ASSESSING LIFE INSURANCE RISKS.

By

MODUPE AJALA

DANIEL DEWITTE

UNIVERSITY OF MASSACHUSETTS LOWELL

DUE- NOVEMBER 10TH 2018

*Abstract*

Research has shown that only about 40% of U.S households own individual life insurance. This has to do with the long and intruding process of getting individual life insurance policy. The process Involves collecting extensive customer information to identify risk classification and eligibility and scheduling numerous medical exams, which takes an average of 30 days. The process turns people off especially in this one-click shopping world with on-demand everything. This paper seeks to analyze a dataset from Prudential, one of the largest issuers of life insurance in the USA, who want to make the process of getting life insurance quote easier and less labor intensive for new and existing customers, while maintaining privacy boundaries.

This paper seeks to use a combination of various data mining and statistical techniques such as CART, C5, KNN, to predict life insurance quotes of individuals while using only a few predictors/ variables (family history, medical history, age, height, weight, and so on.). This technique will help by significantly cutting down the long process of getting life insurance policy. This paper will show a predictive model that accurately classifies customer risk using a more automated approach (CRISP-DM), which can enormously impact public impression of the industry.

*Keywords*: Life Insurance, Medical history, Predictive analytics, CRISP- DM, Risk factor, Data mining.

**1. Introduction**

Life insurance like any other insurance is a contract between an insurer and a policyholder (usually the insured). The policyholder pays a monthly or annual premium to the insurer and in exchange, the insurer promises to pay the beneficiary a sum of money or benefits upon his/her death, or other occurrence like a terminal or critical illness. The insurer sometimes also agrees to cover other expenses such as the funeral expense of the insured.

**How life insurance works**

Life insurance transactions consist of three parties, the insurer, the insured, and the policyholder (*BSLAMC, 2018).* Typically, the insured is the policyholder, but that it not always the case. For instance, if Jane buys life insurance policy on her own life then she is both the insured and the policyholder, but if Thomas her husband, buys the policy on her behalf then she is the insured and Thomas is the policyholder/ owner (*BSLAMC, 2018).* A beneficiary is the person who will receive all the proceeds if the insured dies. The beneficiary is not a party to the policy but is designated by the policy owner to receive the insurance proceeds when the insured dies.

A life insurance policy is a legal contract between the policy owner and the insurance company. It specifies all the terms, conditions and agreement between both parties including the monthly/ annual premium, the risks assumed, and the death benefits the insurer will pay to the beneficiary upon the insured’s death. This policy can become null in certain situations (i.e. the suicide clause). If for example the insured commits suicide in the first couple of years (usually two years) of the policy, the contract becomes null because the investment potion hasn’t had enough time to mature *(Ann et al., 2013)*.

There are several types of Life insurance policy, all consisting “death benefits” and “premium”. A premium is a payment the policyholder pays to the insurance company. This premium payment is determined by several risk factors, (including age, gender, medical history, weight, height, smoking history, dangerous hobbies, family history, occupational hazards and other risk factors.), administrative fee, and policy maintenance fee. Actuaries are professionals who calculate the cost of insurance (COI) using probability and statistics. They calculate the cost of insurance using mortality tables (statistically based tables that shows a person’s average life expectancy). The insured/ policyholder usually chooses “death benefits” amount based on the estimated future needs of their beneficiary. Death benefit is the combination of money/ benefits the insurance company agrees to pay to the beneficiary upon the death of the insured.

**Types of Life Insurance**

Generally, there are two main types of life insurance policy; Whole-life insurance, and Term life insurance. They contain subcategories including universal life insurance, variable life insurance, accidental death insurance and other umbrella terms. Which policy is best suited will depend on the type of coverage the owner seeks. For instance, 39-year-old Tom had a major health scare, and in the moment he realizes he doesn’t have a plan put in place for his family’s survival when he passes. Tom is now faced with a decision on how best to ensure his family’s safety when he his gone. He decided on life Insurance and has to determine what policy is best suited for his situation. Let's say he has a wife and a daughter and calculated how much they would need to survive. He also calculated how much mortgage they had left on the house, and other debts they had incurred. He estimated his family would need a total of $500,000 if he dies. Now what type of insurance would guarantee a death benefit of $500,000 over the course of his life? He has yet another choice to make *(Ann et al., 2013)*.

**Whole-life insurance policy:** A whole life policy combined a term policy with an investment portion *(Ann et al., 2013).* Some of the premium is put in the insurance and some goes into a savings or investment account that built value. A whole life policy is usually more expensive than a Term policy, a part of the premium goes into a investment account, and it helps people who aren’t good at saving, save money. A whole-life insurance policy does not pay an acceptable amount of death benefit until its 12th of 15th year *(Ann et al., 2013).* If the owner dies before then, the entire premium paid before then would be forfeited, because the investment portion had not had time to build. If the owner doesn’t die before then, the return is guaranteed.

**Term-life insurance policy:** This type of policy only covers the insured for a certain period of time. For example, a customer can decide to be insured for 10, 15, 20, or 30 years. Depending on how many years the customer wants coverage for. The life insurance policy expires after a specified number of years and once it expires, so does the coverage. Typically, this type of life insurance premium is usually less expensive than whole life insurance. A disadvantage to Term-life insurance policy is that the insured can outlive his/her policy and if the insured still needs coverage, he/her will have to buy another policy and would pay more depending on the insured’s current age.

**Universal-life insurance:** This type of life insurance falls under a broader category of policies sometimes referred to as cash-value or permanent insurance. Universal-life insurance policy combines a term policy with an investment portion. Some of the premium is put in the insurance and some goes into a savings or investment account that built value. It is just like whole-life insurance except that the return is not guaranteed. It combines death benefits with a saving component or cash value that is reinvested and tax deferred *(Mary, 2018)*. Buyers needed to treat the investment portion of the insurance like an investment and change the investments as they aged. Universal life had even more investment choices Universal life and variable life were choices for such special needs; but costs were higher, and returns were not guaranteed. *(Mary, 2018).*

**1.1. Motivation**

Millennial are a large, and increasingly influential, segment of the U.S population yet only account for 1 out of 5 people who own life insurance *(Rishel, 2015).* Millennial are somewhat undereducated when it comes to life insurance. According to (The Insured Retirement Institute and the Center for Generational Kinetics, 2015), 15 percent of millennial think winning the lottery is a “viable retirement strategy” and 11 percent are hoping for monetary gifts to see them through their later years. Getting Life insurance takes a long and tiring process. Could this be the main reason millennial’s are turned off? What if the process becomes easier? Would this help drive life insurance sale? These are the questions this paper seeks to analyze.

Life insurance companies like any other insurance company are not required to issue coverage to any applicant (individuals/ small groups), they are allowed to pick the risks they will insure and to deny or limit their coverage through “underwriting”. Underwriting ensures that each applicant receives fair and consistent evaluation and, if accepted, is charged an appropriate premium for the coverage offered *(Schuman, 2015).* A basic insurance principle is that every individual or small group should pay an amount of premium that is equivalent to the amount of risk the company assumes for that person. If not a lot of individuals and small group markets would pay a higher premium than they should. Life Insurance companies’ base underwriting decisions on the response to the questions they ask. The questions are designed to know the risk factors of each individual. It regards to the medical history, family history, narcotics or alcohol usage, age, weight, height, employment records, occupational hazards, travelling history, dangerous hobbies and other risk related questions. The applicant is expected to answer all these questions truthfully, so the insurer can correctly classify their risk level. Sometimes the insurer also ask personal unrelated questions such as an applicant’s income and net worth, previous purchase or cancellation of insurance coverage, previous suspension or revocation of a driver’s license, and conviction of a misdemeanor or felony *(Schuman, 2015)*. The insurer sometimes ask the applicant general questions like when they had their last physician visit and would expect them to disclose visit to a gynecologist, or optician even though this medical specialty is not listed on the application. The insurer would also require applicant to undergo numerous medical exams to check for diseases that would increase the applicant’s risk factor.

This process would be a turn-off to most people. The lengthy process will cost insurance companies to lose potential customers. For instance, a person who decided to get a life insurance policy goes to a life insurance company and is immediately bombarded with tons of paper work containing questions about his/her personal life. He/she was also informed that they will need to take some medical exams and the process will take an average of 30 days. The process sounds stressful and time consuming, so the customer might decide to invest in something else, and just like that the insurance company loses a customer. This is especially bad for the whole industry in general and is the main motivation behind this paper. The paper would look at different predictive models of analyzing the most important risk factors associated with Cost of insurance (COI). This analysis will help by significantly cutting down the lengthy process of getting life insurance policy. This paper will show a predictive model that accurately classifies risk using the CRISP-DM approach.

**1.2. Literary Review**

The life insurance business carries a lot of risk, mainly this comes in two forms: market risk factors, and the risk classification of the customer who is purchasing the life insurance policy (known as the customer for the rest of this paper). This study will be focusing on the later of the two. Correctly identifying and assigning a risk classification to the customer is incredibly important to the life insurance companies, the risk classification of the customer considers several of the customer’s unique identifiers. These identifiers could be: gender, height, weight, age, smoking status, medical history, employment information, etc. All these identifiers and more are used in a process known as underwriting to assess and classify the risk of the customer. *(Wupperman, 2017)* In general, the higher the classified risk of the customer, the higher the premium the customer must pay in order to get life insurance.

There are two different ways life insurance companies charge their premiums, these are: pooled pricing for life insurance, and risk-differentiated pricing *(Hao et al., 2017).* Different countries have different guidelines on what a life insurance company can and cannot do for risk classification of their customers. Under the Affordable Care Act, the United States only allows insurers to classify customers by: smoking status, location, family size, and age; in the EU gender classification has been banned; and several countries have restricted insurers’ use of genetic test results *(Hao et al., 2017).* Risk pooling is when life insurance companies do not charge different premium prices to their customers based upon the risk classification, typically this is used when the insurance companies are not allowed to charge risk-differentiated pricing to the customer. The pooled price is generally more expensive for the lower risk customers and cheaper for the higher risk customers, higher risk customers buy more insurance than the lower risk customers *(Hao et al., 2017).*

Adverse selection is ever present when discussing risk classification, usually it is framed negatively as a complication that should be avoided and/ or minimized *(Thomas, 2008).* Adverse selection is a concept that people with a high risk of death are more likely to own life insurance (*Hedengren et al., 2016)*. For the most part the perspective that adverse selection is a bad thing is accurate; in the case of public policy this perspective is limiting and implies that adverse selection is a complication in creating efficiency when pricing life insurance. The advent of some adverse selection in certain insurance markets may be a good thing “the right people, people more likely to suffer loss, will tend to buy (more) insurance.” *(Thomas, 2008)*. Thomas (2008) suggests that if restrictions on risk classification do increase adverse selection that this may not be a bad outcome from a public policy viewpoint *(Thomas, 2008).* The question then becomes who does adverse selection affect and how? Adverse selection when not accounted for can be adverse to insurance companies, adverse selection can also be adverse to lower risk customers (who pay more for insurance than they might have without adverse selection) *(Thomas, 2008)*. Adverse selection also affects the higher risk customers. The higher risk customers can obtain insurance that they otherwise may not and/ or even pay a lower premium for the same insurance *(Thomas, 2008)*.

Adverse selection can be created by information asymmetries, which leads to inefficient outcomes *(Wuppermann, 2017)*. In the prior paragraph it was ascertained that adverse selection could be a problem for insurance companies and the lower risk customers. This comes into the equation when insurance companies are not allowed to charge premiums based upon risk classification and are forced to pool the risk classifications together, on the other hand adverse selection can help out the higher risk customers when risk pooling is required *(Hao et al. 2017)*.

Adverse selection is a main topic of many studies and articles to life insurance and risk classifications seen in *(Hao et al., 2017, Wupperman, 2017, and Thomas, 2008)* as well as several others that are not referenced in this paper. Adverse selection is one of the main reasons for loss coverage. Loss coverage is defined as “expected losses compensated by insurance” *(Hao et al., 2017)*. Adverse selection affects the expected loss coverage that an insurance company will have to pay. This is done because, as insurance companies pool pricing for life insurance it is more likely that higher risk customers will purchase the life insurance, but as this happens the average pooled price for the life insurance will be higher than when risk – differentiated pricing is used. This is the reason a few lower risk customers will buy insurance *(Hao et al., 2017)*. As the shift towards higher risk customers has to reach a certain point, it can easily offset the fall in the number of customers insured; when this happens the loss coverage to insurers can be higher *(Hao et al., 2017).*

Hao et al. (2017) presented a simple example that represented the reality that not all high risk customers will purchase insurance when presented with a fair premium *(Hao et al., 2017).* This revelation is important to highlight as it is generally not taken into account with the typical theories of insurance demand. The typical theory is based upon the premise that all individuals will purchase insurance if offered a fair price *(Hao et al., 2017)*. From the work Hao et al. (2017) presents, it is apparent that individuals make different decisions when it comes to purchasing life insurance when all other things are considered equal. From the studies and modeling conducted by Hao et al. (2017) and Thomas, (2008), loss Coverage is one of the larger risk factors for life insurance companies. These companies combat the effects of loss coverage by raising or lowering the premiums of their product. They do this in two ways either by charging risk-differentiated pricing (where allowed to by law) or by raising the cost of the pooled price of the life insurance *(Hao et al., 2017).*

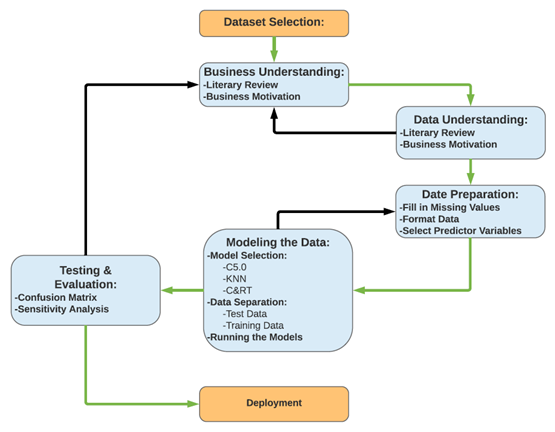
Risk classification or risk factors were discussed many times in the literature in researching this topic as well is the dependent variable in this study. Risk classification is the method used by insurance companies to rate customers based upon unique variables that are assessed during the underwriting process to calculate their premiums *(Wuppermann, 2017)*. The top two risk factors considered in calculating premiums are gender and age, although there are other factors like: smoking status, weight, and marital status that are taken into account *(Wang et al., 2016)*. In different markets for life insurance across the world there are differences in how underwriting is conducted; this is mainly due to the different legal restrictions *(Wuppermann, 2017).* The United States prohibits employers from changing the premium of life insurance based upon health related information, while in the European Union the use of gender in insurance underwriting has been banned *(Wuppermann, 2017).* Since 2010, life insurance companies in the United States can only use the following as risk factors: age, geographic region, household size, and tobacco usage *(Wuppermann, 2017).* Though smoking is considered one of the top three risk factors, there are reports out there that 5 years after you quit smoking the risk of stroke is reduced to someone who has never smoked *(Blackman, 2017).*

In places like Taiwan, the premiums of most policies depend only on age and gender *(Wang et al., 2016).* For the most part the elderly and males are charged with higher premiums, other factors like smoking, obesity, high blood pressure, and other health conditions are causes for higher insurance rates. *(Wang et al., 2016)* Due to the difficulty of obtaining health related information of its citizens, life insurance companies in Taiwan use other means to gather information to assess the risk of customer. The use of Taiwan’s population register system contains information like the marital status and the marital history of each individual *(Wang et al., 2016)*. This type of information is increasingly important to life insurance companies as the relationship between life expectancy and marriage is well known. The married tend to live longer than the unmarried and have lower mortality rates *(Wang et al., 2016).* “In addition to living longer, it is believed that married couples achieve a stable life and good physical and mental condition through the division of labor; thus, the marriage also plays a role in health protection” *(Wang et al., 2016).* Life insurance companies are looking for this type of information due to the fact that married couples are more likely to live longer and therefore pay more for over their lifetime for life insurance than unmarried individuals. Because of this, the premiums for married couples will be less for the equivalent unmarried individual *(Wang et al., 2016)*.

The needs of customers change when it comes to life insurance, in turn life insurance companies need to be able to change with the times and be flexible and adaptable to the changing needs of their customers *(Rishel, 2015)*. As the millennial generation is starting to become the dominate generation in the work force and the baby boomer generation is starting to retire, the life insurance industry is evolving to better serve the baby boomers as they are living longer than ever) *(Rishel, 2015).* At the same time life insurance companies are starting to prepare for the millennial generation with new product offerings that are more appealing to the up coming generation of millennial *(Rishel, 2015).* The fact is that many millennials have already fallen behind in planning for their financial future, about 70% of millennials don’t know how much they ought to be saving for retirement, and only 40% are putting at least 10% of their incomes aside for retirement *(Rishel, 2015).* Millennials grew up in great financial turmoil, and are also paying burdensome student loans, getting married, and taking on other financial risks; it is understandable that long term financial planning isn’t the top priority for them *(Rishel, 2015).* Coupled with the fact that Americans are also grappling with chronic illnesses that may contribute to their need for costly long-term care and baby boomers are trying to cope with more stress and health issues compared to the same age group 10 years ago, its no surprise health insurance companies are trying to evolve their products to address the baby boomer generations as well as the millennials *(Rishel, 2015)*. Two of the ways that life insurance companies are doing this is that they are partnering with medical canters on research and taking advantage of advanced analytics to better understand and price risk *(Rishel, 2015).*

**2. Methodology:**

This study will be using a version of the Cross-Industry Standard Process for Data Mining (CRISP-DM) technique for conducting research, data analysis, and reporting results. The CRISP-DM process was proposed in the mid-1990’s to be a standard methodology for data mining *(Sharda et al., 2015).* CRISP-DM is six sequential steps that allows for a lot of backtracking, making it an iterative process as the analyst becomes more familiar with the managerial needs of the company *(Sharda et al., 2015).* Each successive step in the process is a building block of the previous step giving a well-organized and thought out framework of the data mining project. The six steps are: Business Understanding, Data Understanding, Data Preparation, Model Building, Testing & Evaluation, and Deployment. Figure 1 is a graphical depiction of the CRISP-DM process used for this study.



**Figure 1.** CRISP-DM workflow used in this study.

**2.1 Business Understanding:**

Risk classification for life insurance companies is very important and can be incredibly difficult, especially when the companies are bound by law that they cannot use risk classification or that their risk factors are pre-defined to a small subset of fields. Naturally the more information about the customer, the more accurate the life insurance company can assess the risk and charge an appropriate premium for their product. Since this is a major component of the loss coverage estimates that affect a life insurance company’s profitability, it is very important to predict the risk of the customers very accurately so that they can provide the right product at the right price to their customers.

Prudential posted a data mining assessment in 2015 as they are looking for ways to accurately predict and assess risk in a much more compact and time friendly manner. Historically the way to accurately assess the risk was to ask countless questions and get a medical exam (which can take over 30 days to get). This takes too much time with today’s population used to having what they want with a few clicks of a button and delivered the next day, Prudential is looking for a more accurate way of assessing risk and to do so in a much shorter time period than before. After completing an extensive literary review into life insurance companies, it was apparent that the proper risk classification of customers was important for managing loss coverage from adverse selection and high mortality rates among individuals.

**2.1 Data Understanding:**

The data used in this study was a dataset provided by Prudential Life Insurance downloaded from Kaggle.com (2018). The original data set had over 59,000 entries with over 100 different variables. Prudential provided an explanation of what each of the variable are and their data type, this can be seen in Table 1. From the literary review it is apparent that the variables that are the most commonly used in risk classification is: age, health, smoking habits, gender, martial status, employment information, and medical history. From this it is possible to start narrowing the 100+ data variables down to just the variables that are of interest in this study.

|  |  |  |
| --- | --- | --- |
| Variable Name: | Data Type: | Description: |
| Id |  | Unique identifier for an applicant |
| Product Info: 1, 2, 3, 5, 6, 7 | Nominal | Normalized variables relating to the product applied for |
| Product Info: 4 | Continuous | Normalized variables relating to the product applied for |
| Ins\_Age | Continuous | Normalized age of applicant |
| Ht | Continuous | Normalized height of applicant |
| Wt | Continuous | Normalized weight of applicant |
| BMI | Continuous | Normalized BMI of applicant |
| Employment Info: 1, 4, 6 | Continuous | Normalized variables relating to the employment history of the applicant |
| Employment Info: 2, 3, 5 | Nominal | Normalized variables relating to the employment history of the applicant |
| Insured Info: 1-7 | Nominal | Normalized variables providing information about the applicant |
| Insurance History: 1, 2, 3, 4, 7, 8, 9 | Nominal | Normalized variables relating to the insurance history of the applicant |
| Insurance History: 5, 6 | Continuous | Normalized variables relating to the insurance history of the applicant |
| Family Hist: 1 | Nominal | Normalized variables relating to the family history of the applicant |
| Family Hist: 2-5 | Continuous | Normalized variables relating to the family history of the applicant |
| Medical History: 1, 10, 15, 24, 32 | Discrete | Normalized variables relating to the medical history of the applicant |
| Medical History:  2-9, 11-14, 16-23, 25-31, 33-41 | Nominal | Normalized variables relating to the medical history of the applicant. |
| Medical Keyword: 1-48 | Dummy Variables | Dummy variables relating to the presence of/absence of a medical keywords associated with the applicant |
| Response | Ordinal | Target Variable, a measure of risk that has 8 levels |

**Table 1. Variables in Prudential Dataset.**

**2.3 Data Preparation:**

After reviewing the dataset from Prudential, it was determined that the “Id” variable is a unique identifier for the applicant and is not suitable for datamining, therefore this variable was removed from the dataset. In the same manner this study was not to analyze the products that the customer was applying for so the variables “Product Info 1-7” were also eliminated from the data set. The final set of variables that was determined that could be eliminated were the “Medical Keyword 1-48” variables, as Prudential categorized these as dummy variables that represented the presence or absence of medical keywords associated to the customer. After this initial data preparation that still left the dataset with 59,000+ entries, 71 predictor variables, and 1 target variable.

Upon inspecting the remaining variables for data completeness, it was seen that several of the remaining variables were missing values for some of the entries. First variable to be analyzed was “Employment Info\_1”; it was found that our of 59,383 entries, 19 of them were missing data for this variable, or .00032% of the data was missing for this variable. From Table 1 it is known that this data is continuous and a normalized value (ranges from 0-1). The following approach was taken to fill in the missing values: 1) Dataset was copied and opened in Excel and the entire data set was then sorted from small to large on this column, 2) the column was then selected and all duplicate values were removed, 3) an average was taken from the values that remained, and a values of 0.098184612 was found, 4) the average value in step 3 was then added to the 19 missing entries in the original data set. After reviewing the rest of the data set there was a total of another 12 variables that had missing values, all 12 variables were reviewed and found that all of them had greater than 10% of the entries were missing data for those variables. For the purposes of this study, those 12 additional variables were also removed from the data set, along with the “Product\_Info” variables as they are not typically used in calculating the risk classification of a customer. The list of all variables that had missing entries filled in, or the variable was removed from the original data set and be found in Table 2. After preparing the dataset, there was a total of 61 variable removed from the data set, and 1 variable had missing data filled in. This left the dataset with a total of 66 variables to be used as predictors, and 1 target variable in the data set.

|  |  |
| --- | --- |
| **Variable Name:** | **Reason:** |
| Id | Unique identifier, not able to be data mined |
| Employment Info \_1 | 0.00032% of entries were missing values, a value of 0.098184612 was entered for those missing values. |
| Employment Info \_4 | 11% of entries were missing values, variable removed |
| Employment Info \_6 | 18% of entries were missing values, variable removed |
| Insurance History\_5 | 43% of entries were missing values, variable removed |
| Family Hist\_2 | 48% of entries were missing values, variable removed |
| Family Hist\_3 | 58% of entries were missing values, variable removed |
| Family Hist\_4 | 32% of entries were missing values, variable removed |
| Family Hist\_5 | 70% of entries were missing values, variable removed |
| Medical Hist\_1 | 15% of entries were missing values, variable removed |
| Medical Hist\_10 | 99% of entries were missing values, variable removed |
| Medical Hist\_15 | 75% of entries were missing values, variable removed |
| Medical Hist\_24 | 94% of entries were missing values, variable removed |
| Medical Hist\_32 | 98% of entries were missing values, variable removed |
| Medical Keyword 1-48 | Dummy variables, variables removed |
| Product\_Info | Variables are not used in calculating risk, variables removed |

**Table 2. List of Variable that were removed from the original dataset.**

**2.4 Model Building:**

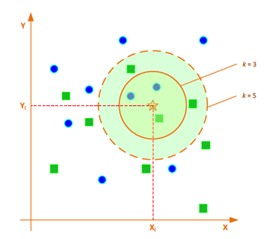
The fourth step of the CRISP-DM technique is model building, this is done after the previous steps. Now the analyst has an in-depth knowledge of the business and its needs, the data and its variables, and they have prepared the data for modeling. The data for this study was analyzed using three different modeling techniques: C5.0, K-nearest neighbor (k-NN), and Classification and Regression Trees (CART) from statistics (*Sharda et al., 2015).* Each of these techniques were used in cooperation with the k-fold cross validation method for creating training and testing data sets.

***2.4.1 C5***

The C5 algorithm is a decision tree model, more specifically it is a variant of the ID3 model which is known as the most widely known decision tree algorithm *(Sharda et al., 2015).* The ID3 model was developed by Ross Quinlan in 1986, and it uses a splitting technique known as information gain *(Sharda et al., 2015).* In place of using the Gini Index to measure the purity of a class as a result of a decision to branch along the particular attribute, ID3 and its variants use entropy to measure the extent of the uncertainty or randomness in a data set *(Sharada et al., 2015).*

***2.4.2 K-Nearest Neighbor (k-NN)***

Compared to Artificial Neural Networks (ANN) and Support Vector Models (SVM), k-NN is rather simplistic in its operation. k-NN is a instance-based learning algorithm where the function is approximated locally and all final computations are done after the actual prediction is made. This model works well for classification and regression prediction problems *(Sharda et al., 2015).* The objective of the k-NN model is assign a new object to a classification based upon its k nearest neighbors, Figure 2 has a simple illustration depicting the modeling type. The star in the illustration represents the new case, and the object is to assign the start to either the circle classification or to the squares based upon what the majority of its nearest neighbors are *(Sharda et al., 2015).* What is most importance when it comes to k-NN is choosing of the k-value. As displayed in Figure 2, a k = 3 value assigns the start to the circles, but a k = 5 value gives the start to the squares.



**Figure 2. Illustration of how k-NN works. *(Sharda et al., 2015)***

***2.4.3 Classification and Regression Tree (CART)***

The CART algorithm was created by Breiman et al. 1984 and uses the same type of greedy search found in AID and THAID *(Loh, 2014).* The main benefits of CART is that instead of using stopping rules, it allows the decision tree to keep growing then it prunes the tree to an appropriate size that has the lowest cross-validation estimate of error *(Loh, 2014).* The pruning procedure is a weakest-link cut method, and the links between the indexed values of a cost-complexity parameter. *(Loh, 2014)*

***2.4.5 k-Fold Cross Validation (Rotation Estimation)***

k-Fold Cross Validation is a method used to split the dataset into testing and training sub groups while minimizing the bias associated with the random sampling of the training and testing data samples *(Sharda et al., 2015).* Using k-Fold Cross Validation, the data set is broken into k groups of equal size, all but one of the groups is used as the training data and the remaining group is used to test the model and then the results are compared for accuracy. This method is then repeated k times switching which groups are the training data set and which is the testing data set *(Sharda et al., 2015).* Once the process is complete, the accuracy of the model is easily calculated by averaging the k individual accuracy measures. The number of folds to create in the data set has been found to be k = 10, this has been the best balance in terms of performance and being optimized for computational time *(Oztekin, 2018).*

**2.5 Testing and Evaluation**

***2.5.1 Confusion Matrix***

In order to evaluate the performance of each of the models a confusion matrix will be used. A confusion matrix is the standard method for evaluating classification-based models. The confusion matrix is valuable to the data analyst as it can derive a few key performance indicators: accuracy, sensitivity, specificity, and precision.

 Eq. 1

 Eq. 2

 Eq 3.

 Eq. 4

Where TP stands for True Positive, TN stands for True Negative, FP is False Positive, and FN is False Negative.

***2.5.1 Information Fusion-based Sensitivity Analysis***

Once the performance of a model has been evaluated and defined, it is important to know the weight of each of the predictors in the dataset to see how important they are to the model. This is done using a sensitivity analysis, the basis of sensitivity analysis is to remove a predictor variable from the model and see how the model output changes, this process is then repeated for all of the variables in the dataset. The Information Fusion-Based Sensitivity Analysis then applies a weight to those sensitivity scores for each of the models and combines them so that different types of models can be combined and compared.

**2.6 Deployment**

The last step of the CRISP-DM technique for data mining is deploying the model. This involves taking the final results of the model and coalescing the information in an appropriate manner so that management can review the data and be able to make decisions based upon the results of the data mining project.

**3.0 Results:**

In the introduction and motivation sections, the main goal of this study was to accurately predict the risk classification of an entry in the dataset, so that Prudential Life Insurance can speed up the process of quoting life insurance premiums and minimize their loss coverage. In order to do this the dataset was used to train three different models. The dataset was then fed back into the models and the model performance was evaluated to compare which model was the most accurate in predicting the risk classification. Each model was subject to a k-fold cross validation routine (10 folds is what was used for the k value in the cross validation) and then evaluated using a confusion matrix, other data such as predictor variable sensitivity was also extracted from the models. The k-Cross Validation routine averages the performance of each fold of the validation process into a single table of results for the model. The models used in this study was decision tree C5 model, the k-Nearest Neighbor (k-NN) model, and the Classification and Regression Tree (CART) model.

***3.1 C5 Model Results***

The C5 Model is a variant of the ID3 decision tree model created by Ross Quinlan in 1986 *(Sharda et al., 2015).* Using the C5 model in this study, it had an overall accuracy of 71.8% and Table 3 shows the confusion matrix for the model. Table 4 lists the accuracy, sensitivity, specificity, and the precision of each of the risk classifications.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Predicted Risk Classification | Risk Classification |  |  |  |  |  |  |  |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 3954 | 435 | 96 | 70 | 230 | 389 | 247 | 183 |
| 2 | 475 | 4099 | 75 | 65 | 317 | 396 | 203 | 202 |
| 3 | 13 | 14 | 435 | 32 | 57 | 49 | 37 | 36 |
| 4 | 18 | 15 | 21 | 683 | 36 | 82 | 61 | 77 |
| 5 | 213 | 387 | 102 | 6 | 3403 | 439 | 129 | 102 |
| 6 | 456 | 540 | 171 | 257 | 513 | 7832 | 798 | 632 |
| 7 | 220 | 261 | 30 | 35 | 194 | 556 | 4397 | 423 |
| 8 | 858 | 801 | 83 | 280 | 682 | 1490 | 2155 | 17834 |

**Table 3. C5 Confusion Matrix**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Risk Classification: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Accuracy: | 0.718024284 |  |  |  |  |  |  |  |
| Precision: | 0.637023 | 0.625611 | 0.429418 | 0.478291 | 0.626473 | 0.697231 | 0.547776 | 0.91508 |
| Sensitivity: | 0.705567 | 0.702846 | 0.64636 | 0.687815 | 0.711776 | 0.699348 | 0.718934 | 0.73746 |
| Specificity: | 0.958105 | 0.954191 | 0.990155 | 0.987241 | 0.962839 | 0.929413 | 0.93185 | 0.95298 |
| T-n | 51524 | 51096 | 58130 | 57643 | 52571 | 44781 | 49635 | 33543 |

**Table 4. Results Evaluation of the C5 Model.**

***3.2 Classification and Regression Tree (CART) Model Results***

The second model that this study used was the Classification and Regression Tree (CART) which is a statistics-based decision tree *(Sharda et al., 2015).* As with the C5 model the CART model integrated a k-fold cross validation routine and evaluated using a confusion matrix, once the model was complete a sensitivity analysis was conducted on the predictor variable importance to the model. This model was unique in that the IBM SPSS Modeler 18.1 software program did not have a cross validation routine built into the model node, thus the k-fold cross validation routine had to be created in the modeling software. Table 6 has the complete confusion matrix for the CART model, from which the overall accuracy of the CART model in predicting the risk classification was calculated to be 48.07%, Table 7 shows the accuracy, precision, sensitivity, and specificity of each risk classification of the CART Model.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Predicted Risk Classification | Risk Classification |  |  |  |  |  |  |  |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 234 | 99 | 22 | 21 | 37 | 104 | 102 | 37 |
| 2 | 443 | 914 | 21 | 0 | 340 | 70 | 2 | 3 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 424 | 653 | 185 | 2 | 1661 | 505 | 26 | 24 |
| 6 | 2208 | 2155 | 570 | 769 | 1682 | 5968 | 2052 | 1435 |
| 7 | 1104 | 1112 | 20 | 26 | 683 | 1773 | 3207 | 1427 |
| 8 | 1794 | 1619 | 195 | 610 | 1029 | 2813 | 2638 | 16563 |

**Table 6. CART Model: Confusion Matrix**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Risk Classification: |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Accuracy: |  | 0.480743 |  |  |  |  |  |  |  |
| Precision: |  | 0.037699 | 0.139499 | 0 | 0 | 0.305781 | 0.531292 | 0.399527 | 0.849864 |
| Sensitivity: |  | 0.356707 | 0.50976 | 0 | 0 | 0.477299 | 0.354415 | 0.342921 | 0.607571 |
| Specificity: |  | 0.898289 | 0.902098 | 0.982941 | 0.975952 | 0.932541 | 0.87624 | 0.903656 | 0.908904 |
| T-n |  | 52752 | 51950 | 58368 | 57953 | 52130 | 37277 | 45209 | 29194 |

**Table 7. CART Model: Results Evaluation**

***3.3 k-Nearest Neighbor (k-NN) Model Results***

The last model that was used in this study was the k-Nearest Neighbor Model or k-NN. The k-NN model is designed to assign the new object to a classification based upon the neighbors of the predictor variables of its neighbors *(Sharda et al., 2015).* The k-NN model took a very long time to compute compared to the C5 and CART models above, it took over 2 hours to obtain the analysis and confusion matrix alone, whereas the C5 and CART model took less than an hour. It proved to be difficult to get a predictor importance while using the IBM SPSS Modeler 18.1 using this model, in part due to the large dataset. The overall accuracy of the k-NN model was 54.02%. Table 8 has the complete confusion matrix for the k-NN model, and Table 9 shows the accuracy, precision, sensitivity, and specificity of each risk classification generated by the k-NN model.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Predicted Risk Classification | Risk Classification |  |  |  |  |  |  |  |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 1587 | 246 | 34 | 27 | 157 | 197 | 97 | 72 |
| 2 | 490 | 1778 | 43 | 31 | 261 | 242 | 98 | 77 |
| 3 | 13 | 29 | 115 | 9 | 27 | 37 | 16 | 18 |
| 4 | 27 | 43 | 31 | 231 | 23 | 49 | 33 | 51 |
| 5 | 218 | 312 | 52 | 32 | 1278 | 223 | 109 | 83 |
| 6 | 1274 | 1450 | 350 | 402 | 1112 | 6176 | 1260 | 764 |
| 7 | 594 | 668 | 75 | 76 | 589 | 694 | 3011 | 524 |
| 8 | 2004 | 2026 | 313 | 620 | 1985 | 3615 | 3403 | 17900 |

**Table 8. k-NN Model: Confusion Matrix**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Risk Classification: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Accuracy: | 0.540172783 |  |  |  |  |  |  |  |
| Precision: | 0.255679 | 0.271368 | 0.113524 | 0.161765 | 0.235272 | 0.549809 | 0.375109 | 0.918467 |
| Sensitivity: | 0.656599 | 0.588742 | 0.435606 | 0.473361 | 0.553966 | 0.482953 | 0.483229 | 0.561727 |
| Specificity: | 0.918896 | 0.915296 | 0.98481 | 0.979675 | 0.927217 | 0.891464 | 0.905626 | 0.94225 |
| T-n | 52344 | 51587 | 58219 | 57696 | 52920 | 41536 | 48134 | 25926 |

**Table 9. k-NN Model: Result Evaluation**

**4.0 Information Fusion-Based Sensitivity Analysis**

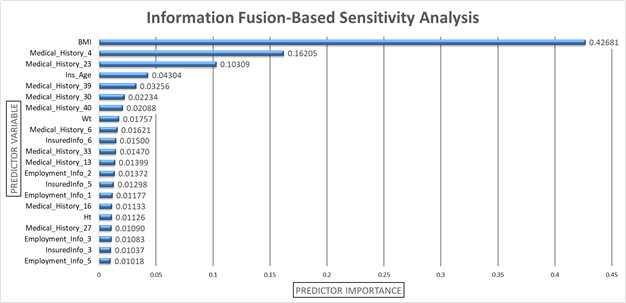
One of the final steps in a data mining model is to run a sensitivity analysis on the predictor variables of the models and compare the sensitivity analysis of each model to each other, this is known as information fusion-based sensitivity analysis. There is no single model that would work the best for every problem that exists in data mining, likewise there does not exist a single best method for representing the results of multiple models in a manner that can be easily digested, this is where information fusion-based sensitivity analysis comes important *(Oztekin et al., 2014).* Information fusion-based sensitivity analysis is a unbiased way of ranking the predictor variables across multiple models by a weighting method (*Oztekin et al., 2014).*

For this study the information fusion-based sensitivity analysis consists of gathering the individual model predictor variable importance and applying a weighting scheme of using the individual model accuracy that was calculated from the confusion matrix. An example of the fusion sensitivity for a variable can be seen in Eq. (5).

Fusion Score=(A1∗V1)+(A2∗V2)+(A3∗V3) Eq. (5)

Where A1 is the accuracy of model 1, A2 is the accuracy of model 2, A3 is the accuracy of model 3, V1 is the variable predictor importance calculated from model 1, V2 is the variable predictor importance calculated from model 2, and V3 is the variable predictor importance calculated from model 3. Eq. 5 gives a way to combine the individual model information for each predictor variable importance while applying a weighting of the model accuracy, this then gives a way to rank the information and therefore the models.

After using the IBM SPSS Modeler 18.1 software package to run all three of the models, the individual sensitivity analysis for each model was exported to an excel file and then combined to create and calculate the information fusion-based sensitivity of each predictor variable to the model. Figure 4 shows the top 20 most important predictor variables and their fusion based sensitivity scores to the models that were run in this study while Table 10 has the complete list of the variable fusion-based sensitivity analysis scores.

**Figure 4. Information Fusion-Based Sensitivity Analysis.**

|  |  |
| --- | --- |
| **Predictor Variable** | **Fused Sensitivity Score:** |
| Medical\_History\_22 | 0.000390592 |
| Family\_Hist\_1 | 0.000718 |
| Medical\_History\_26 | 0.001356181 |
| Medical\_History\_7 | 0.001807432 |
| InsuredInfo\_7 | 0.001926314 |
| Medical\_History\_31 | 0.0020104 |
| Medical\_History\_14 | 0.002644314 |
| Medical\_History\_8 | 0.0027284 |
| Medical\_History\_9 | 0.002872 |
| Medical\_History\_34 | 0.0030874 |
| InsuredInfo\_4 | 0.003125697 |
| Medical\_History\_2 | 0.003388935 |
| Medical\_History\_25 | 0.00359 |
| Medical\_History\_11 | 0.003850871 |
| Medical\_History\_41 | 0.003949 |
| Medical\_History\_35 | 0.004397329 |
| Medical\_History\_36 | 0.0048106 |
| Insurance\_History\_9 | 0.004954114 |
| Medical\_History\_17 | 0.005016605 |
| Insurance\_History\_3 | 0.005337076 |
| InsuredInfo\_1 | 0.005526611 |
| Medical\_History\_37 | 0.0056722 |
| Medical\_History\_20 | 0.006880514 |
| Medical\_History\_19 | 0.006981753 |
| Insurance\_History\_1 | 0.007168629 |
| Medical\_History\_28 | 0.007239355 |
| Insurance\_History\_8 | 0.007336105 |
| Insurance\_History\_7 | 0.007412553 |
| Medical\_History\_29 | 0.0077544 |
| InsuredInfo\_2 | 0.007908065 |
| Medical\_History\_5 | 0.008616 |
| Insurance\_History\_2 | 0.008663606 |
| Medical\_History\_21 | 0.009072811 |
| Medical\_History\_3 | 0.009168487 |
| Medical\_History\_38 | 0.009532977 |
| Medical\_History\_18 | 0.009604777 |
| Medical\_History\_12 | 0.009775011 |
| Insurance\_History\_4 | 0.010004076 |
| Employment\_Info\_5 | 0.010179177 |
| InsuredInfo\_3 | 0.010368334 |
| Employment\_Info\_3 | 0.010825377 |
| Medical\_History\_27 | 0.010897177 |
| Ht | 0.011256177 |
| Medical\_History\_16 | 0.011327977 |
| Employment\_Info\_1 | 0.011770453 |
| InsuredInfo\_5 | 0.012979377 |
| Employment\_Info\_2 | 0.01372423 |
| Medical\_History\_13 | 0.01399313 |
| Medical\_History\_33 | 0.014702577 |
| InsuredInfo\_6 | 0.015004576 |
| Medical\_History\_6 | 0.016210377 |
| Wt | 0.017574577 |
| Medical\_History\_40 | 0.020877377 |
| Medical\_History\_30 | 0.022341377 |
| Medical\_History\_39 | 0.032558572 |
| Ins\_Age | 0.043044031 |
| Medical\_History\_23 | 0.103093226 |
| Medical\_History\_4 | 0.16204673 |
| BMI | 0.426810392 |

**Table 10. Fusion Based Sensitivity Analysis**

From the information provided by each of the individual models and the fusion based sensitivity analysis, the best model performance seen in this study was the C5 decision tree model with an accuracy of 71.8%, followed by the k-NN model with an accuracy 54.02%, and last was the CART model with an accuracy of 48.07%. From the information fusion-based sensitivity analysis the top variables that has the most effect on the models and the results of this study can be seen in Figure 4, with the top three being: BMI, Medical\_History\_4, and Medical\_History\_23.

**Conclusion**

This study used a few data mining models to predict risk classification of customers based on some variables (age, weight, height, medical history, etc.). This will help life insurance companies classify customers into their appropriate risk group, thereby ensuring the appropriate premium quote is given to each customer. A life insurance company’s ability to accurately classify customers into appropriate risk group will help the company gain a competitive advantage and would also reduce the market risk the company tend to face when they incorrectly classify a customer into a wrong risk group. Life insurance companies would save a lot of money by charging the right premium for certain risk levels (I.e. the riskier you are, the more premium you pay). This way, less risky people are protected from “Risk pooling” (when life insurance companies do not charge different premium prices to their customers based upon the risk classification). It will also help reducing the lengthy process of getting life insurance policy. This will in turn, impact public perception of the industry and entice more people to get life insurance.

The models adopted in this study are Classification and Regression tree (CART), C5.0 Decision Tree, and k-Nearest Neighbor (k-NN). The C5 decision tree model yielded the highest accuracy result of 71.8% and showed the most important variables to the model are: BMI, Medical history 4 and Medical history 23. Not knowing the real identity of most of the predictor variables made it hard to provide an in-depth result analysis of the risk classification. The overall accuracy of the k-NN model was 54.02% and the overall accuracy of the CART model in predicting the risk classification was calculated to be 48.07%.

For this study the information fusion-based sensitivity analysis consisted of gathering the individual model predictor variable importance and applying a weighting scheme of using the individual model accuracy that was calculated from the confusion matrix. With approximately 72% accuracy using the C5 model, the results can be used with reasonable reliability to predict the most important variables in risk classification. This study concludes that additional research is still needed to increase the accuracy of modeling outcomes, while a 72% accuracy is not bad, it can be further improved to increase the reliability of the model.

**References:**

1. Hedengren, David & Stratmann, Thomas (2016). Is there Adverse Selection in Life Insurance Markets? *Economic Inquiry, Vol. 54, No. 1, pp. 450-463*
2. Blackman, Irv (2017). You maybe insurable after all… *Contractor Magazine, July 2017, pp. 24*
3. Hao, Minglie, Macdonald, Angus S., Tapadar, Pradip, & Thomas, R. Guy (2017) Insurance loss coverage and demand elasticities. *Insurance: Mathematics and Economics, 2018, Vol. 79, pp. 15-25*
4. Bajtelsmit, Vickie L., Villupuram, Sriram V., & Wang, Tianyang (2015) Life Insurer Cost of Equity with Asymmetric Risk Factors. *The Financial Review, 2015, Vol. 50, pp. 435-457*
5. Thomas, R. Guy (2008). Loss Coverage as a Public Policy Objective for Risk Classification Schemes. *The Journal of Risk and Insurance, 2008, Vol. 75, No. 4, pp. 997-1018*
6. Wuppermann, Amelie C. (2017). Private Information in Life Insurance, Annuity, and Health Insurance Markets\*. *The Scandinavian Journal of Economics, 2017, Vol. 119, No. 4, pp. 855-881*
7. Rishel, Rod (2015). Life Insurance Outlook: New approaches for changing realities. *National Underwriter Life & Health, December 2015, pp. 38-42*
8. Wang, Hsin Chung, Yue, Jack C., Tsai, Ti-Hsuan (2016). Martial Status as a Risk Factor in Life Insurance: An Empirical Study in Taiwan. *Astin Bulletin, 2016, Vol. 46, No. 2, pp. 487-505*
9. Hackert, A., Brookman, J (2013). My life in numbers: an insurance decision. *Journal of Critical Incidents. Oct2013, Vol. 6, p71-74. 4p*
10. Schuman, G (2015). The devil is in the details: establishing an insured’s intent to deceive in life and health insurance recession cases *FDCC Quarterly.Winter2015, Vol. 64 Issue 2, p84-113. 30p.*
11. Bajtelsmit, V., Villupuram S, Wang T (2015). Life insurer cost of equity with asymmetric risk factors. *Financial Review. Aug2015, Vol. 50 Issue 3, p435-457. 23p.*
12. Spierdijk, L., Koning, R (2014). Estimating outstanding claim liabilities: the role of unobserved risk factors. *Journal of Risk* ***&*** *Insurance. Dec2014, Vol. 81 Issue 4, p803-830. 28p.*
13. Gatzert, N., Wesker, H (2014). Mortality risk and its effect on shortfall and risk management in life insurance. *Journal of Risk* ***&*** *Insurance****.*** *Mar2014, Vol. 81 Issue 1, p57-90. 34p.*
14. Kaggle.com. (2015). Prudential life insurance assessment. Retrieved from [https://www.kaggle.com](https://www.kaggle.com/)
15. Nerdwallet. (2017) How life insurance works. Retrieved from <https://www.nerdwallet.com/blog/insurance/how-does-life-insurance-work/>
16. Sharda, Ramesh, Delen, Dursun, Turban, Efraim (2015). Business Intelligence and Analytics: Systems for Decision Support. Tenth Edition. *Pearson Eductation, Inc. Copyright 2015, 2011, 2007*
17. Loh, Wei-Yin (2014). Fifty Years of Classification and Regression Trees. *International Statistical Review ,2014, Vol. 82, No. 3, pp. 329-348*
18. Oztekin, Asil, Khan, M. Raiz (2014). A Business-Analytic Approach to Identify Critical Factors in Quantitative Disciplines. *Journal of Computer Information Systems, 2014, Vol. 54, No. 4, pp. 60-70*

**About the Authors**



In 2017 Ms. Modupe Ajala received her B.S in International business from Eastern Mediterranean University located in North Cyprus (Turkey). She started her career by working full time at a marketing firm in Nigeria, till she decided to go abroad to pursue her master's degree. She is currently pursuing a M.S degree in Business Analytics at the University of Massachusetts Lowell and is currently working part time with the marketing team at Victoria’s secret. She recently accepted a position at Digital Federal credit union (DCU) and would officially resume work there spring of 2019.