Determining Insurance Policyholder Value

Ted Zybin, Rohit Nair, Ibrahim Alwishah

2022-12-06

External Requirements

age	education_num	marital_status	occupation	cap_gain	hours_per_week	score	value_flag
39	13	Never-married	Group 2	2174	40	58.99	Low
50	13	Married-civ-spouse	Group 5	0	13	55.78	Low
38	9	Divorced	Group 1	0	40	62.75	Low
53	7	Married-civ-spouse	Group 1	0	40	60.10	Low
28	13	Married-civ-spouse	Group 5	0	40	53.31	Low
37	14	Married-civ-spouse	Group 5	0	40	54.07	Low

Data Preparation

Count and Remove Some Missing Values as Appropriate (NAs)

```
## [1] 48842     8
```

```
##       age        education_num   marital_status      occupation
##  Min.   :17.00   Min.   : 1.00   Length:48842       Length:48842
##  1st Qu.:28.00   1st Qu.: 9.00   Class :character   Class :character
##  Median :37.00   Median :10.00   Mode  :character   Mode  :character
##  Mean   :38.64   Mean   :10.08
##  3rd Qu.:48.00   3rd Qu.:12.00
##  Max.   :90.00   Max.   :16.00
##     cap_gain     hours_per_week      score        value_flag
##  Min.   :    0   Min.   : 1.00   Min.   :43.94   Length:48842
##  1st Qu.:    0   1st Qu.:40.00   1st Qu.:57.50   Class :character
##  Median :    0   Median :40.00   Median :60.24   Mode  :character
##  Mean   : 1079   Mean   :40.42   Mean   :60.23
##  3rd Qu.:    0   3rd Qu.:45.00   3rd Qu.:62.95
##  Max.   :99999   Max.   :99.00   Max.   :76.53
```

```
##            age  education_num marital_status     occupation       cap_gain
##              0              0              0              0              0
## hours_per_week          score     value_flag
##              0              0              0
```

```
## [1] 48842
```

```
## [1] 0
```

Data Transformation

```
## 'data.frame':    48842 obs. of  8 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital_status: chr  "Never-married" "Married-civ-spouse" "Divorced" "Married-civ-spouse" ...
##  $ occupation    : chr  "Group 2" "Group 5" "Group 1" "Group 1" ...
##  $ cap_gain      : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ score         : num  59 55.8 62.8 60.1 53.3 ...
##  $ value_flag    : chr  "Low" "Low" "Low" "Low" ...
```

```
## 'data.frame':    48842 obs. of  8 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital_status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation    : Factor w/ 6 levels "Group 1","Group 2",..: 2 5 1 1 5 5 1 5 5 5 ...
##  $ cap_gain      : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ score         : num  59 55.8 62.8 60.1 53.3 ...
##  $ value_flag    : Factor w/ 2 levels "High","Low": 2 2 2 2 2 2 2 1 1 1 ...
```

Research Question

How do the provided variables associate with the prediction that a customer is of high value and what model provides a feasible baseline accuracy using the most important, if not all, variables for the aforementioned prediction?

Introduction

Risk assessment is a critical aspect of the insurance industry, as it allows companies to accurately assess the likelihood of a policyholder filing a claim and to set premiums accordingly. By identifying high value policyholders, insurance companies can prioritize providing them with the best possible products and services, leading to increased customer satisfaction and loyalty. Additionally, accurately predicting which policyholders are likely to be low value can help the company make informed decisions about which potential customers to accept or reject. This can help the company avoid accepting high-risk policyholders who are likely to file costly claims, ultimately leading to improved financial performance. Marginal improvements in risk assessment techniques can therefore have a significant impact on the overall success of an insurance company.

In this study, we will explore the relationship between various predictor variables and the target variable of whether an individual is considered a "high value" or "low value" policyholder by an insurance company. We have been provided with data on past policyholders, including their age, level of education, marital status, occupation, capital gains, hours worked per week, and an insurance score. Using this data, we will construct a model that can accurately predict whether a prospective policyholder will be of high or low value to the company. While predictive accuracy is of utmost importance, we will also strive to interpret our results in a way that makes intuitive sense. For the purposes of this analysis, we will only consider individuals who are 25 years of age or older.

Based on the data provided, the factors of interest include:

- The age of the prospective policyholder
- A numerical indicator of the amount of education that a policyholder has
- Marital status of the policyholder
- Occupation of the policyholder
- Capital gains recorded on the investments of the policyholder

- The number of hours worked per week by the policyholder
- The proprietary insurance score of the policyholder

Furthermore, we can synthesize new variables from the existing ones that would prove useful in ultimately predicting that value of a potential policyholder.

- The ratio of capital gains to hours worked per week, which could provide insight into the policyholder's investment habits and financial success.
- A variable which captures the effect of increasing age on the value of education
- Use regression analysis to derive a predictive model that incorporates all of the existing predictor variables, which would allow us to identify the relative importance of each variable and potentially create a new, composite variable that represents the combined effect of all predictors.

Before deriving insights from associations implied by the models used, it is important to condense the data using useful statistical summaries. Summary statistics will provide valuable information about the distribution of the data and the overall spread of the values within each variable. For example, the mean and standard deviation can be used to identify any potential outliers in the data, while the range can provide a sense of the overall variability within the data. Additionally, comparing the summary statistics for each variable can help identify any potential relationships or patterns within the data. These summary statistics can be used to inform the development of the predictive model, as well as to interpret and understand the results of the analysis.

Data Set Description

The data set has 48842 rows and 8 columns, above it also shows the mean and max of different categories using the 'summary' function. The data set also does not have NAs or incomplete rows.

This data set contains information for an unspecified insurance provider. The data recorded is categorized by age, education number which is a score that is a numerical value given based of their level of education, marital status, occupation, capital gain on investments, and value flag which is an integer of a high value customer and low value customer.

The variables used in the data are age which is a numerical variable that is taken in years, education number is a numerical value showing the level of education of the customer, marital status is a seven level variable that tell us their relationship status, occupation is 6 split into 6 groups that there is no specific meaning to except the sixth group which indicates the occupation is not known, capital gain is a numerical value that is recording on investments, and hours per week which is a number recorded in hours of the amount of work customers work.

Exploratory Data Analysis

Descriptive Statistics

- Present the data in different types of tables, diagrams, and find all descriptive stats (mean, median, mode, std.dev, etc.,)
- 3 numerical summary tables and 3 graphs

Numerical Summary I

```
##            names    min      Q1  median         Q3      max        mean
## 1            age  25.00   36.00   44.00    53.0000    90.00    45.42797
## 2  education_num   1.00    9.00   11.00    13.0000    16.00    11.15689
## 3       cap_gain 114.00 3674.00 7298.00 14084.0000 99999.00 13460.39969
## 4 hours_per_week   1.00   40.00   40.00    50.0000    99.00    44.01833
## 5          score  43.94   57.91   60.73    63.4475    76.35    60.71336
##             sd    n missing
```

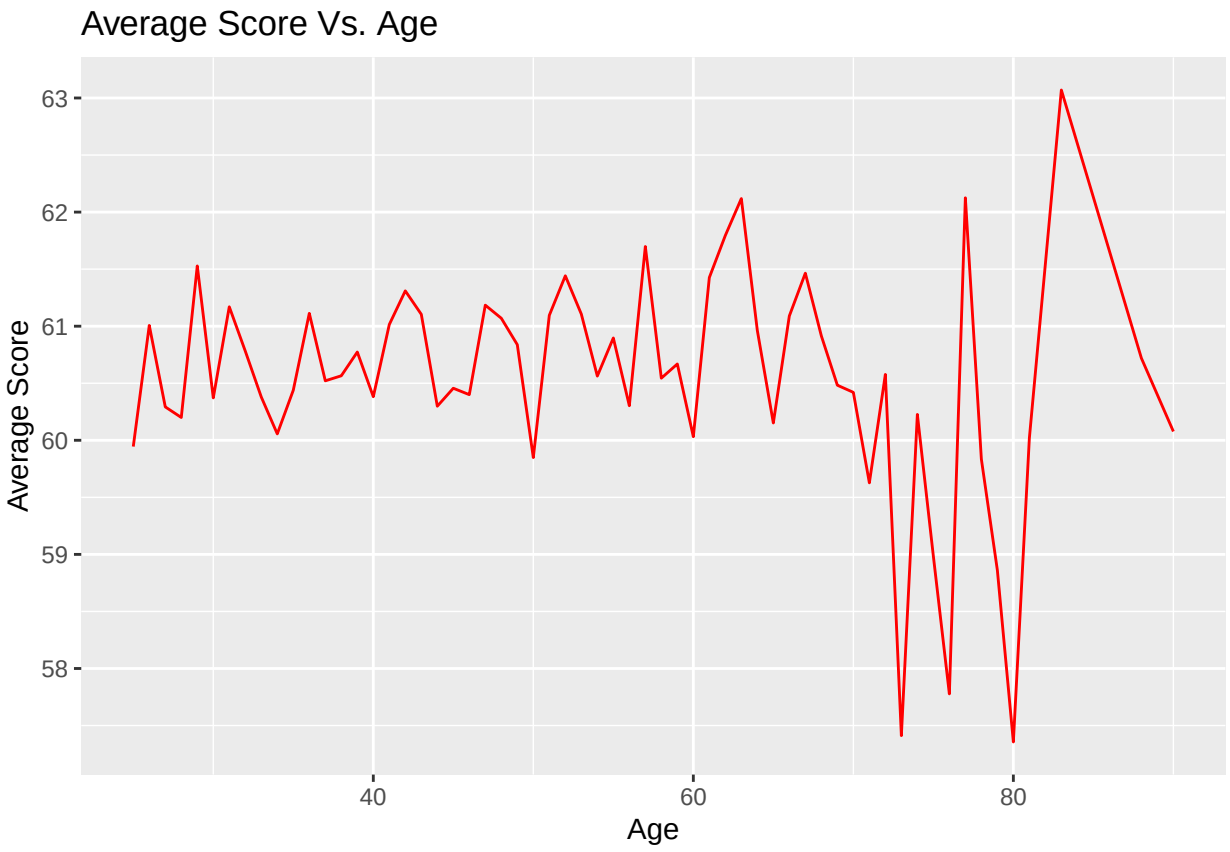
```
## 1    12.519084 3818       0
## 2     2.677252 3818       0
## 3 22991.335330 3818       0
## 4    12.201443 3818       0
## 5     4.127331 3818       0
```

Numerical Summary II

avg_education_num	avg_cap_gain
Min. :11.16	Min. :13460
1st Qu.:11.16	1st Qu.:13460
Median :11.16	Median :13460
Mean :11.16	Mean :13460
3rd Qu.:11.16	3rd Qu.:13460
Max. :11.16	Max. :13460

Visualization

Average Score vs. Age



The idea behind this graph is to show how age may play a part in someone's score. Since the insurance score is a proprietary value, it is expected to highlight policyholder trustworthiness. It was assumed that middle-aged policyholders would have a higher score. However, this was not the case. The line plays an optical illusion which makes this seem as though it is very static while if you play close attention the y-axis is ranging from 59-61.5, so the average score does not have much variance but only appeases that way because

of how the graph is skewed. To go in deeper depth taking a look at the people at the older ages where we see the variability, there rent many to record on, this causes the average to be seen as skewed. We can deduct from this that the average score when taking into account age does not play a major role in the data as there isn't any major changes in the average score

Distribution of Policyholder Scores by Occupation

The graph above explicates the association between type of occupation and score according to the value flag they were labeled with. It's shown that the median scores for all types of occupations never go above 60 for those labeled with low. As this graph is plotting the categorical variable of occupation group against numerical variable insurance score, we see that the insurance score plays an important part in choosing a high value client. The groups that we categorized by the value of high rather than low always had a insurance score higher than the low value client. This can help us infer that with a higher insurance score comes a higher likely hood that the client is a high value one.

Value Flags by Age, Based on Education

The graph above looks at the total incidence of each value_flag at all ages present in dataset, while divided into facets based on the type of occupation. For Groups 1-3, there's a higher number of people classified as low for the bulk of the recorded ages. Groups 4, 5, and NA, all do not seem to have much variablity.Group 4 shows the high and low value flagged clients mixed together, and the same can be said about group NA. We can conclude from these graphs that the age does not play a major role in the groups when observing the values flags of clients.

Synthesis of New Variables

Analysis

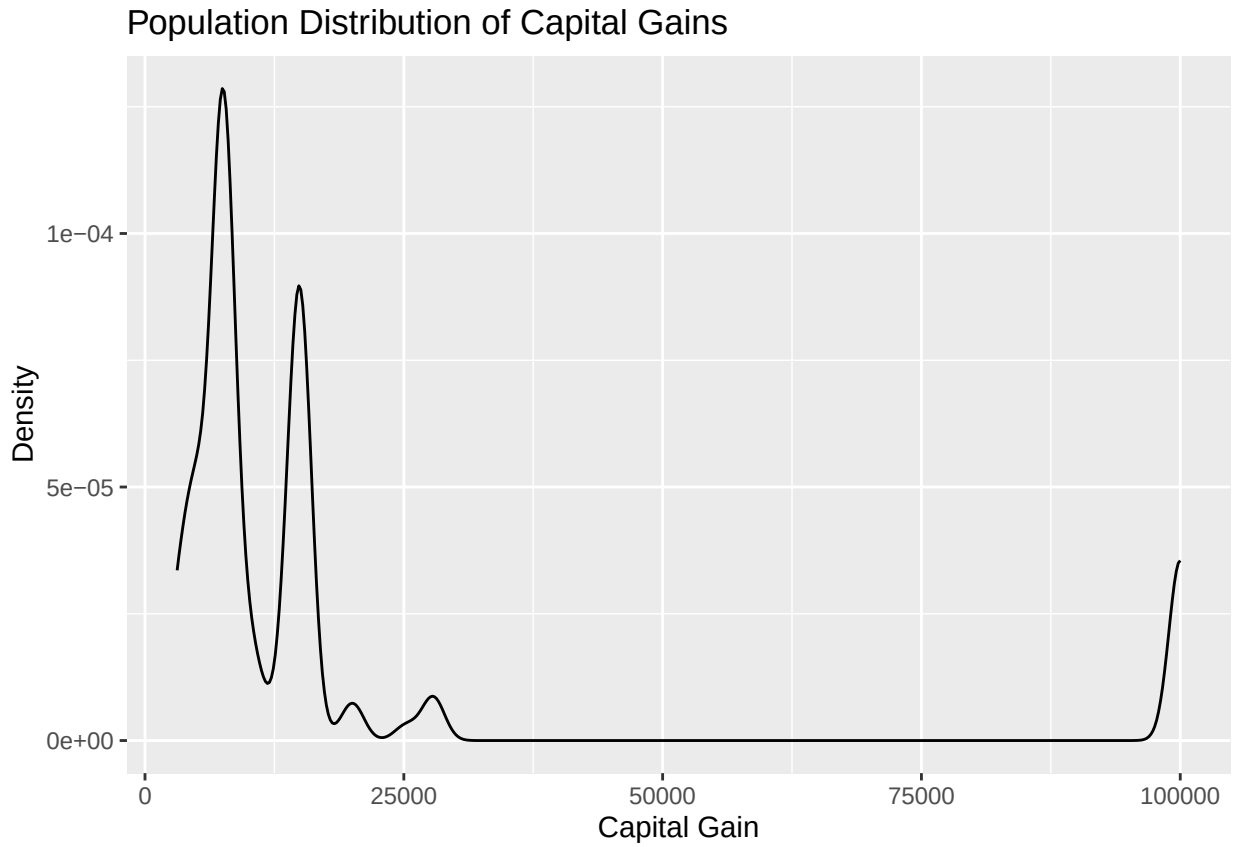
The analysis begins with a confidence interval analysis, to study the the effects that capital gains and age have on the value flag of clients. The principle components of this data set are:

1. Age
2. Education number
3. Marital status
4. occupation
5. capital gain
6. Hours per week
7. Score
8. Value Flag
9. Capital Gains to Hours Per Week Ratio

Out of these components of the data set the population of age and capital gain is focused on to conclude if they play a rule in the categorizing a client as high value.

Bootstrap Confidence Interval

Our bootstrap confidence interval is taken on individuals older than or equal to the age 25 and have a capital gain greater than 0 USD, with a high value flag. We were given a population size of 2462.



The graph shows the skew we have with the capital gain. We can infer that there is not much variability after 25000 for capital gain, although we do see a small spike when reaching 100000.

```
##   boostrap_samplemeans
## 1             21828.93
## 2             17872.95
## 3             21766.25
## 4             19839.31
## 5             19064.01
## 6             20245.12
```

In the chart above we see the estimated summary taken on the sample population of those over the age of 25 who's capital gain is greater than 0. This data can be used to calculate errors and perform hypothesis test to further support our research.

Hypothesis Testing

The average capital gains value within a sample of this population is assumed to be greater than $20000 USD. Is this supported by the data?

H0 : The average capital gains amount for an individual equal to or above the age of 25 flagged as being high value is equal to $20000.

Ha : The average capital gains amount for an individual equal to or above the age of 25 flagged as being high

value is not equal to $20000.

H0 : μ = 20000

Ha : μ != 20000

To compute the p-value, we assume that the null hypothesis is true and we'll set the significance level at 5%.

```
## [1] 16703.65
```

```
## [1] 20000
```

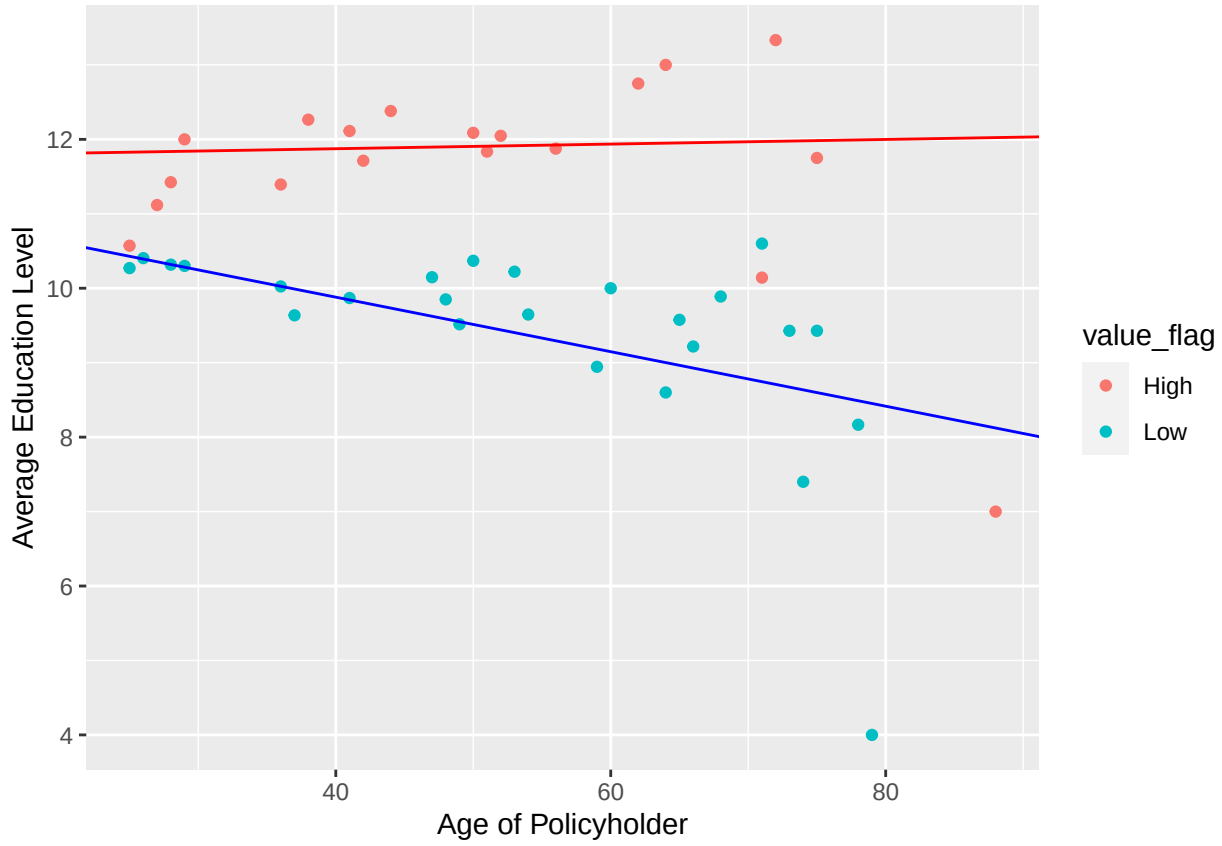
```
## p-value is equal to: 0.98
```

Since the p-value is greater than 0.05, We fail to reject the null hypothesis. The data do not provide convincing evidence at the .05 significance level that the average capital gains value is not equal to 20000. As we fail to reject the null hypothesis, it is feasible to assume that those that are older than 25 and a high value policyholder are likely to have an average capital gain of $20000. The information we have now can be used to conclude that age and capital gain do play a role in predicting high value clients. The variable that we focus on with the test would be age in years greater than 25 and a capital gain greater than 0.

Regression Models

The regression models shown are both taken on education levels and show the variability that education level gives when predicting a high value client and a low client.

This graph below shows that a higher education level is often times a high value client. We see that by taking the average education level and plot it against age we have a line of best fit that shows a linear increase for education level of high value clients. While in the same graph we a see a negative linear line that shows education level is less in low value clients



```
## [1] 1.779317
```

```
## [1] 1.61098
```

Using a Support Vector Machine Model

In order to synthesize the provided data and get a feasible predictive model, we can use an SVM model. SVM models can handle high-dimensional data, making them well-suited for situations where there are many predictor variables. In the insurance industry, policyholders may have many different characteristics that could affect their value, such as their age, education level, occupation, and insurance score. An SVM model could help to accurately predict the value of a policyholder based on these variables.

SVM models can provide high predictive accuracy, which is important for the marketing department of an insurance company. By accurately predicting the value of a policyholder, the marketing department can make more informed decisions about which prospective customers to target, leading to higher profits for the company.

SVM models are known for their ability to handle data with complex, non-linear relationships between the variables. In the insurance industry, the relationships between the predictor variables and the target variable (policyholder value) are likely to be complex and non-linear. An SVM model could help to accurately capture these relationships and make accurate predictions.

SVM models are robust to noise and outliers in the data, which can be common in real-world datasets. In the insurance industry, the data on policyholders is likely to be noisy and contain outliers, such as policyholders with unusually high or low values. An SVM model could help to accurately predict the value of policyholders even in the presence of noise and outliers.

Ultimately, using an SVM model could be beneficial for an insurance company because it can handle high-dimensional data with complex, non-linear relationships, provide high predictive accuracy, and be robust to noise and outliers in the data. These qualities could help the marketing department of the insurance company to make more informed decisions about which prospective customers to target, leading to higher profits for the company.

Discussion + Conclusion

This is a general overview of what risk assessment for an insurance policyholder would look like. With the data provided, we were able to select pertinent variables,both numerical and categorical, and identify the various associations between the variables which would contribute to the predictions of a policyholder being high or low value. Graphs were made to further analyze patterns. Bootstrap hypothesis testing was used to look at sample distributions of capital gains and affirm its connection to value flagging. The association between level of education and capital gains also was explored in order to ascertain their relative weights within a possible prediction model. In the end, based on the nature of the variables provided and the associations found, a support vector machine model was used to predict whether a potential customer would be high or low value. The most important part of our project was to trace any relationship between the flag status of a client and other variables presented in the table.
But why is it essential to identify high-value customers? Here are presented several reasons to look for high value customers:
- Knowing your high-value customers helps you to plan ahead and make better decisions about future projects.
- You can maximize their value and develop longer-lasting relationships. - You can focus your marketing and acquisition strategies on similar customers using lookalike audiences. - Delighting these customers can create a positive feedback loop, where they bring in more business for your company.

We noticed that there is a relationship between the flag status of a client and his/her education level. Clients, whose degree was higher than 10, were considered to be «high» value clients. Despite the fact that the insurance company made some exceptions, it would be right to say that the agents of the insurance company have a specific algorithm, which helps them group their clients with respect to the level of their degree. Besides, it was proven that clients with a score less than 61 tend to be considered as low value clients.

Apart from that, our team conducted a test hypothesis. Our initial hypothesis was that the average capital gains amount for an individual equal to or above the age of 25 flagged as being high value equals $20000, while the alternative hypothesis claimed that the average capital gains amount for an individual equal to or above the age of 25 flagged as being high value does not equal $20000. As we did the test, p-value turned out to be more that 0.05 and that gave us enough evidence to claim that the initial hypothesis should be accepted. In other words, we had enough statistical evidence to assume that those who are older than 25 and are considered to be high value are likely to have an average capital gain of $20000. The information we have now can be used to conclude that age and capital gain do play a role in predicting high value clients.

As an answer to our research question, we think that it is essential to say that there are different factors and variables that associate with a prediction that a random customer would be considered a high value client. Arguments like age,education level and score play a crucial role for the insurance company and clients' status. Other variables like marital status, occupation, capital gain and working hours per week are also taken into consideration by agents of the insurance company, but are considered to be less important.

Contributions

Rohit worked on the Introduction and Overview of SVMs. Ibrahim helped with the Introduction and provided the descriptions for the Exploratory Data Analysis. Ted wrote the Discussion and Conclusion.

References

- "Introduction to Machine Learning with R" by Brett Lantz
- RDocumentation - e1071 (version 1.7-12)