River Wilkinson

Southern New Hampshire University

Professor Genon

DAT-640: Predictive Analytics

Final Project

11 July 2021

**Introduction**

Data mining allows an organization to discover hidden patterns in data by use of machine learning. Sophisticated algorithms are the mining tools used in predictive analytics to extract hidden values from newly discovered patterns. Organizations can examine and process their data to better understand their customer base, improve their operations, outperform their competitors, and better position themselves in the marketplace (Bari, A., 2017). Predictive analytic models can be categorized in various ways. These categories can be sorted by the business problems the model solves and the primary business functions they serve (such as sales, advertising, human resources, or risk management) and the mathematical implementation used in the model (such as statistics, data mining, and machine learning). Based on an organization’s business objective a clustering and classification model, decision model, or association model may be used.

Predictive analytics is the process of refining a data source using business knowledge to extract hidden value from newly discovered patterns. Organizations can use it to search for ways to better understand their customer base, improve operations, and sharpen their competitive edge by increasing their market share. By harnessing predictive analytics, organizations can manage risk, optimize operations, and maximize profits. Organizations can also use predictive analytics to search for ways to better understand their customer base; knowing the pattern associations in customer data will help an organization to shape its strategic plan, optimize resource allocation and increase customer satisfaction. An organization’s customer data can be harnessed to zero in on the customers most likely to buy and best to market to.

The hidden value in data is an organization’s most valuable asset. Organizations can gain this value using data mining, the discovery of hidden patterns of data through machine learning. Data is collected from past occurrences to try to make sense of hidden insights. Examination of data can provide potential answers to business questions. Models, mathematical representations of an object or process, are built to simulate real-world events to further investigate a problem. The value of these predictive models includes 1. How the model predictions drive improved, more effective action, and 2. How the model delivers insight which leads to an improved strategy used to guide future decisions. The development of a predictive analytic strategy helps an organization to brainstorm the data that will be collected, how that collected data is used, and how it will be analyzed.

**Organizational Background General**

T.I.C. has had previous mishaps of completing a successful data mining project to determine the attributes of the customers that would be interested in purchasing a caravan policy. These mishaps occurred due to poor implementation of the CRISP-DM framework. To better understand their customer base, improve their operations, outperform their competitors, and better position themselves in the marketplace, T.I.C would need to develop a predictive analytics plan.

**Organizational Background Potential Value**

Customer data is valuable when it is directed towards marketing results because of how marketing can be more targeted, and money can be allocated more precisely. Customer segmentation uses clustering to identify customers that are “like” each other based on attributes or behaviors. Customer segmentation has several benefits. These include the advantage of determining appropriate product pricing, developing customized marketing campaigns, designing an optimal distribution strategy, choosing specific product features for deployment, and prioritizing new product development efforts (Sagar, 2019).

With this information, T.I.C will have the opportunity to acquire the right customers and provide them with the best insurance coverage for caravans. The information derived from this dataset will uncover the attributes of the type of customer who will be interested in insurance coverage for their caravan. Having an idea of what this potential customer “looks like” will help T.I.C to better market insurance coverage and maximize its marketing strategies. Customer segmentation will have proved to be successful for T.I.C based on sales performance related to marketing campaigns.

**Business Objective**

The business objective for T.I.C is to segment its customer base by use of demographic information, geographical information, and product usage information to determine who would be interested in purchasing a caravan policy. Customer segmentation has several benefits. These include the advantage of determining appropriate product pricing, developing customized marketing campaigns, designing an optimal distribution strategy, choosing specific product features for deployment, and prioritizing new product development efforts (Sagar, 2019). Customer segmentation will have proved to be successful for T.I.C based on sales performance related to marketing campaigns.

**Data Documentation**

The data supplied by Sentient Machine Research, a Dutch data mining company, contains a training set of over 5,000 descriptions of customers including information on whether that customer has a caravan policy. A test set contains 4,000 customers whom only the organization knows if the customer has a caravan insurance policy (Putten, 2000). The data consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. There are 5,822 customer records in a tab-delimited file. Attributes 1-43 contain sociodemographic data and attributes 44-86 contain product ownership. The customers living in the same zip code have the same sociodemographic attributes. The target variable, attribute 86, is “CARAVAN: Number of mobile home policies” (Putten, 2000). The data was collected to determine if T.I.C can predict which customers will be interested in buying a caravan insurance policy.

**Predictive Algorithm Specification**

Predictive analytics is used in predicting future outcomes. Current and historic data is analyzed to make predictions by employing techniques from statistics, data mining, machine learning, and artificial intelligence (Kumar, 2018). Applying predictive analytics by use of Rattle will provide T.I.C the opportunity to bring about success in identifying the customer that will purchase a caravan policy.

Random forest algorithms belong in the category of supervised learning. Random forest algorithms entail the idea of combining multiple decision trees into a single ensemble to build a forest of trees (Williams, 2011). The random forest algorithm tends to produce quite accurate models because of the ensemble technique, which reduces the instability that is observed when single decision trees are built. Overfitting is reduced with this technique which in turn improves accuracy by a reduction in the level of variance. Having accurate models for customer segmentation will allow T.I.C to have a more accurate prediction of their customers that will be interested in purchasing a caravan policy. This accuracy will benefit T.I.C in terms of more precise marketing strategies that will not only improve sales performance but save valuable resources than using a single decision tree model that would produce less accurate results.

In comparison to decision trees, the random forest algorithm tends to be much more robust to changes that occur within the data. The need for variable selection is avoided because the algorithm does it on its own. Random forest algorithms perform well when many of the input variables have little bearing on the target variable. These algorithms are also suitable when there are many variables and not so many observations. The dataset provided contains 86 variables that include product usage data and socio-demographic data derived from zip area codes. Because random forest algorithms perform variable selection on their own, this would be greatly beneficial for T.I.C to segment their customer base.

**Reproducibility: Model Implementation**

Reproducible research is the practice of careful, annotated preservation of data, analysis code, and associated files, such as statistical procedures, output, and published results that can be directly and fully replicated (Leeper, 2014). Reproducible research involves communicating research in a way that someone else can reconstruct what has been done. During the research process, the reproducibility of research should be considered from the beginning.

Research reproducibility proves to be beneficial if the research process needs to be repeated based on having adequate documentation. Research reproducibility will save time in the long run, recording events as they happen rather than going back to recall something in less detail. If the research process is paused for any amount of time, having documentation would allow for someone to pick back up where they left off.

Text

Description automatically generated

**Figure 1: Data preparation**

Text

Description automatically generated

**Figure 2: Data preparation**

Text

Description automatically generated

**Figure 3: Data preparation**

A computer screen capture

Description automatically generated with medium confidence

**Figure 4: Data preparation**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Figure 5: Data preparation**

Graphical user interface, text

Description automatically generated

**Figure 6: Data preparation**

A screenshot of a computer

Description automatically generated

**Figure 7: Data preparation**

Graphical user interface, application, table

Description automatically generated

**Figure 8: Dataset loaded into Rattle**

Graphical user interface, application, table, Excel

Description automatically generated

**Figure 9: Exploration**

Graphical user interface

Description automatically generated with medium confidence

**Figure 10: Cluster Sizes and Data Means**

Graphical user interface, text, application, email

Description automatically generated

**Figure 11: Random Forest Model Summary**

Table

Description automatically generated

**Figure 12: Variable Importance**

**Pilot Plan: Research Techniques**

The CRISP-DM model (Cross Industry Standard Process for Data Mining) is a technique used to build data mining projects. CRISP-DM provides a framework for planning and managing a data mining project and can be implemented across any domain. This model is defined in six steps: 1. Business understanding, 2. Data understanding, 3. Data preparation, 4. Modeling, 5. Evaluation, and 6. Deployment (Manna, 2014).

The CRISP-DM methodology will be implemented to form a pilot plan for T.I.C to successfully segment its customer base to determine who would be interested in purchasing a caravan policy.

1. **Business Understanding:**

This first step focuses on the project objectives and business requirements. The business objective for T.I.C is to segment its customer base by use of demographic information, geographical information, and product usage information to determine who would be interested in purchasing a caravan policy.

Data mining techniques will be used to determine the features that identify the groups of customers interested in a caravan policy. The problem type needs to be defined (classification, prediction, or clustering).

The project plan would include an estimation of the resources needed i.e., personalized marketing campaigns, how the project will be evaluated, and what tools and techniques will be used.

1. **Data Understanding:**

This stage starts with initial data collection and follows with data familiarity by identifying data quality problems and discovering data insights. During this stage, it is important to know what information is needed, what information is available, how the required data will be collected, and what is the underlying data pattern (Srishilesh, 2021).

The data supplied by Sentient Machine Research, a Dutch data mining company, contains a training set of over 5,000 descriptions of customers including information on whether that customer has a caravan policy. A test set contains 4,000 customers whom only the organization knows if the customer has a caravan insurance policy (Putten, 2000). The data consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes.

Exploratory data analysis will need to be performed to verify the factors that will determine the customer base for caravan insurance policies. The type of data that will be used to solve the business problem would need to be determined as well. This data would be analyzed to determine the relationship between the various data types.

1. **Data Preparation:**

The steps involved with data preparation are as follows: collecting data from various sources, cleaning the data, formatting the data, blending, and integrating the data from multiple sources, and sampling the data to focus on what is important to not waste time and resources.

1. **Modeling:**

During this stage, it is important to determine what modeling techniques would fit best for the business requirements. This includes selecting the model, testing the model, creating the model, and assessing the model (CRISP-DM, 2019).

Clustering is one of the most common exploratory data analysis techniques used to gain insight into the structure of the data. It can be defined as the process of identifying subgroups in the data such that data points in the same cluster are similar, while data points in different clusters are different (Dabbura, 2018). Market segmentation uses clustering to try to find customers that are like each other based on attributes or behaviors. The business objective for T.I.C is to determine which clusters predict who will be interested in a caravan insurance policy.

1. **Evaluation:**

This step involves ensuring that the results match the intended expectations, deciding whether to proceed to the next step or return to a previous step, making note of factors that could fail, and performing testing with end-users (SPSS, 2000). Customer segmentation will have proved to be successful for T.I.C based on sales performance related to marketing campaigns. Based on a success metric, i.e., customer retention rate or monthly sales, T.I.C can gauge if the results match the intended expectations.

1. **Deployment:**

This step starts with the evaluation of the results and concludes with a strategy for deployment of the data mining results into the business. The activities associated with this stage are summarizing deployable results, developing, and evaluating alternative plans for deployment, deciding for each distinct knowledge or information result, identifying possible problems during deployment, etc. (SPSS, 2000).

**Model Optimization: Evaluation**

The random forest algorithm tends to produce quite accurate models due to the ensemble technique that reduces the instability that is observed when single decision trees are built. There is a change in the resulting decision tree by removing a small number of observations from a training dataset. Random forest algorithms are a popular tool for regression and classification and can also be used for clustering. This tool is appropriate for handling mixed datasets.

Model evaluation is the process of identifying the best amongst different models. This process allows us to understand what is to be expected from different models when scoring new observations. Evaluation of model performance also helps us to identify whether any mistakes were made in the choice of input variables (Williams, 2011). The different measures for model evaluation are as follows:

* 1. The risk variable is used as a measure of how significant each observation is for the target variable.
  2. The error rate is the measure of the performance of a model that is calculated as the proportion of observations for which the model incorrectly predicts the class regarding the actual class. The number of misclassifications is divided by the total number of observations.
  3. The confusion matrix measures the performance of a classification model on a set of data where the true values are.

Graphical user interface, text, application, email

Description automatically generated

**Figure 13: Error Matrix**

The error matrix, also known as a confusion matrix, shows the true outcomes against the predicted outcomes. By performing the error matrix in Rattle, two tables are presented: the count of observations and the proportions. The cells of the matrix are referred to as True Negatives, False Positives, True Positives, and False Negatives. In Rattle, the Confusion Matrix is the default on the Evaluate tab. Clicking the Execute button will run the selected model(s) against the chosen dataset to predict the outcomes for each of the observations in that dataset. The predictions are compared with the actual observations (Williams, 2011).

Chart, line chart

Description automatically generated

**Figure 14: Precision and Recall Plot**

The precision of a model is the ratio of the number of true positives to the total number of predicted positives (the sum of the true positives and the false positives). It is a measure of how accurate the positive predictions are, or how precise the model is in predicting.

The recall of a model is just another name for the true positive rate. It is a measure of how many of the actual positives the model can identify, or how much the model can recall. The recall is also known as the sensitivity of the model. Another measure that often arises in the context of sensitivity is specificity. This is simply another name for the true negative rate (Williams, 2011).

Text

Description automatically generated

**Figure 15: Recall Curve**

Chart, line chart

Description automatically generated

**Figure 16: ROC Curve Plot**

The Receiver Operator Characteristic (ROC) curve compares the false positive rate to the true positive rate. There is a trade-off between the number of observations that are incorrectly classified as positives against the number of observations that are correctly classified as positives. The ROC has a form and interpretation like the risk chart, though it plots different measures on the axes. ROCR is used by Rattle to generate these charts (Williams, 2011).

References

CoIL Challenge 2000 Report. (n.d.). Retrieved from http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/data.html Sentient. (n.d.). Retrieved from https://www.sentient.nl/ CoIL Challenge 2000 Report. (n.d.). Retrieved from <http://www.liacs.nl/~putten/library/cc2000/>

*Crisp-Dm | Data mining | Quick explanation*. (2019). *YouTube*. https://www.youtube.com/watch?v=A2jWg5OV0bQ.

Dabbura, Imad. 17 September 2018. Towards Data Science. *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.* Retrieved on 27 June 2021 from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Hodgson, Erin. 25 February 2020. dotActiv. *K-Means Clustering and Why It is Good for Business.* Retrieved on 27 June 2021 from https://www.dotactiv.com/blog/why-k-means-clustering-is-good-for-business.

Kumar, Vaibhav & L., M. (2018). Predictive Analytics: A Review of Trends and Techniques. International Journal of Computer Applications. 182. 31-37. 10.5120/ijca2018917434.

Leeper, T. J. (2014). Archiving Reproducible Research with R and Data verse. The R Journal, 6(1), 151-158. <http://snhu-media.snhu.edu/files/course_repository/graduate/dat/dat640/dat640_archiving_reproducible_research_with_r_and_dataverse.pdf>

Manna, M. (2014, December 18). *The data science project lifecycle*. Data Science Central. https://www.datasciencecentral.com/profiles/blogs/the-data-science-project-lifecycle.

Sagar, Abhinav. 24 August 2019. Towards Data Science. *Customer Segmentation Using K Means Clustering: To better understand your customers*. Retrieved on 27 June 2021 from https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3.

SPSS. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*.

ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

Srishilesh, P. S. (2021, February 2). *Data Mining using CRISP-DM methodology*. <https://www.section.io/engineering-education/data-mining-using-crisp-dm-methodology/>.

Williams, G. J. (2011), Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Use R!, Springer.