CRISP-DM AND ITS APPLICATION FOR INSURANCE BUSINESSES

KRISHNA KULKARNI

Contents

1) Introduction:

a) CRISP-DM:

In this report we are going to discuss about CRISP-DM and its impact on insurance industry. CRISP-DM is an approach used in the field of data science for analytical based decision making to get cost beneficial results in a company or organisation.

CRISP-DM stands for CRoss Industry Standard Process for Data Mining. It is a process model which serves as a base for data science process. A Special Investment Group (SIG) of five companies: Integral Solutions Ltd, Daimler, TeraData, NCR and OHRA in 1995 to tackle the various challenges in the data mining field. A year later CRISP-DM was unveiled. This SIG was backed by ESPIRIT funding initiative as European Union Project. The first version of this methodology/approach was released in 1999.

CRISP-DM was not the first approach to be widely used in data mining. SEMMA from SAS institute was also being used for data mining. The reason for CRISP-DM being a popular approach amongst data miners is that it is INDUSTRY, TOOL and APPLICATION NEUTRAL.

b) CRISP -DM in insurance industry:

Insurance Industry is currently going through a major transformation in terms of its working processes. The industry has a long and problematic history of dealing with issues like insurance frauds, claim efficiency. But due to changing times the industry has now started to ditch the traditional methods in favour of modern techniques such as Machine Learning. Data Analysis techniques such as ML are allowing insurance businesses to create a much more engaging one-on-one relationship with their customers. This means the customer churn can be predicted and prevented in a cost-efficient manner. CRISP-DM enabled techniques and platforms collect data and create algorithms which then help to predict the necessary outcomes.

2) CRISP-DM:

Six Stages of CRISP-DM Process

As shown in fig there are six stages of CRISP-DM namely:

i. Business Understanding
ii. Data Understanding
iii. Data Preparation
iv. Modelling
v. Evaluation
vi. Deployment

The stages in the fig are shown to be in a certain sequence however the user has to navigate back and forth between these stages to get the results.

Each stage has its own significance. Now we will go through each stage of the process in brief.

i. *Business Understanding*: In this stage the user needs to identify the targets it needs to achieve with constraints in place from a business perspective. A plan is devised to achieve data mining and business goals. Also, the user will form up a criterion to determine if the implemented plan has been successful from a business perspective. This stage is important as if the there is no clarity on project goals, then you will be wasting time on the right answers for the wrong questions.

ii. *Data Understanding*: This stage primarily focuses on data acquisition, data description and data exploration. Data description is a report on data acquired such as its format, quantity. In data exploration relationship between different variables and factors is found out. A simple statistical data analysis is performed which has an impact on direction of data mining.

iii. *Data Preparation*: This stage is the most important and also the most tedious. The user needs to select the data that it is going to worked on in the process with criteria relevant to your data mining goals and technical constraints. Then data cleaning is done on the selected data. If not performed then you'll be left with garbage-in, garbage-out.

iv. *Modelling*: A modelling technique depending on the business characteristics is to be selected in this stage. Modelling technique such as neural network may push the user/analyst towards the data preparation stage. In modelling the data is split into traindata and testdata. Traindata is used for model training and testdata for model evaluation.

v. *Evaluation*: The user/analyst builds and chooses models that appear to have high quality based on loss functions that were selected. In this stage it'll be calculated to which degree the model meets the business target/objectives. Also, any flaws/deficiencies are found out and also the business reason behind said flaws is also found out.

vi. *Deployment*: In this stage based on the evaluation results the user will form up a strategy to deploy the model in a code representative format. This code will include the mechanisms to score or categorise new unseen data as it arises. The mechanism should use the new information in the solution of the original business problem. The code must include all the steps from business understanding to modelling.

3) Application of Data Science for Insurance Businesses:

i. ***Better Consumer Targeting/Product Design***: Data Science has bought insurers a lot closer to their targeted customer base through a means of various data gathering methods. For ex. by accessing the customer's search history the insurer gets an insight on customers behavioural patterns (shopping, online purchases etc.). This allows the insurer to build products tailor-made for each type of customer.

ii. ***Fraud Prediction and Prevention:*** Fraud is one of the biggest challenges the industry faces. In USA alone around $80 Bn. is lost annually to fraudulent claims. This can be combated using predictive analytics on the data collected from the customers. A great example would be analysing the spending patterns of a customer who has filed for claims. If the there's a change in spending patterns, then that could be termed as a "Red Flag" and respective actions can be taken on such individuals.

iii. ***Better Claims Management:*** As it has become routine with all things digital insurers need to be fast and agile when it comes to claims management. The claim settlement needs to fast and accurate. Simpler cases of claims can be closed quickly while more complex and suspicious case are forwarded for a closer analysis. A closer analytical look at an individual's social media and its links to other groups can prevent a fraudulent claim.

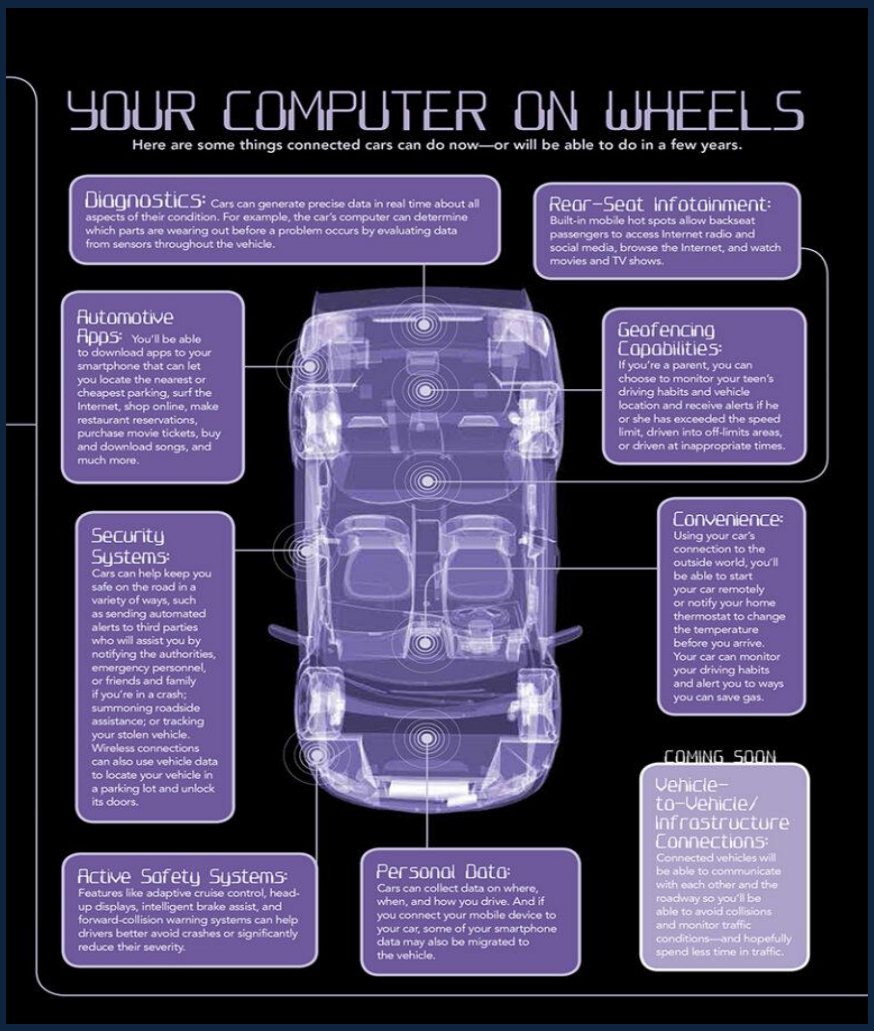
iv. ***Better Engagements with Customers:*** Customer engagement over the course of an insurance policy can be beneficial to both insurer and insured alike. If the low-risk premium policyholders are incentivised to improve their profile risks with a prize then it is massively beneficial to all the parties involved.

v. ***Customer Retention:*** Challenge of customer retention is not exclusive to insurance industry but it is a huge challenge nevertheless. Customer retention is extremely important as it costs significantly more to acquire new customers than to retain the current ones. The customer retention happens due to various factors. One of them being offered a better product from the rival company. Data science allows the insurers to have more insight about their customers such as is the insured in contact with rival company, what is the nature of products and services offered by the rival company and what are the key areas that the company needs to improve in order to retain said insured.

4) How Insurance Businesses come up with data driven solutions:

Below are examples of change in relative insurance domains due to data science,

i. ***Automotive Industry:*** Insurance industry has a significant presence in the automotive industry. It is also one of those industries in which rate of insurance related frauds is significantly higher. But like with every industry insurance in automotive sector is going through a technological renaissance.



TCU (Telematics Control Unit)

With emerging technology like "Telematics" has allowed the insurers to collect more information about their clients. Telematics is nothing but different kind of sensors placed in the vehicle. These sensors measure various parameters from vehicle performance, vehicle safety etc.

All this information is stored in a storage device which is known as TCU (Telematics Control Unit). This is akin to a "Blackbox" used in planes. This information can be used in determining the premium prices of individual policyholders. By using traditional methods, the premium price for two individuals X and Y having same car models would be same.

However, with the use of telematics you can differentiate the premium prices on the basis of car usage data. Insurer will analyse driver's X and Y car data for various factors like: Fuel Economy, Engine Diagnostics, Usage of Seat-belt, Throttle Application, Awareness of Traffic Rules etc. All of this data is stored in TCU, then it is transferred to cloud, which then is used by the insurer. In some cases, a third-party app can also be used to send user data to the insurer. A correlation model with these factors against premium prices can be plotted and the driver who satisfies these factors will have to pay a lesser insurance premium price. In 2016 British Insurance Broker's Association (BIBA) saw a 40% increase in usage of telematics in automotive insurance industry. This system promotes safer driving as the premium prices will be much lower for safe drivers. Also, with the use of an app the insured can keep track of any activity around their claims in real-time.

ii. *Health Insurance*: The recent emergence of IoT devices have completely transformed the health insurance scenario. IoT devices such as Fitbit, apple-watch can provide more information about the insured to the insurer via a first or third-party app. Information like calorie counter, steps-counter, sleep-tracking, heart-rate monitoring and recently oxygen% in blood is sent and analysed. This data then can be used to gauge individual's health status, risks, habits and quote rates accordingly. This can also be used to create incentivise program that encourage policyholders to engage in healthy behaviour in exchange for lower premiums.

Customers expect their claims to be handled in a quick and efficient manner but with the amount of data it becomes a bit complicated to deliver on time. However, with predictive analysis insurers can analyse historical data-such as customer's profile and their claim history- to identify behavioural patterns and use predictive modelling to calculate possible outcomes.

5) Conclusion: Data Science along-with CRISP-DM enabled platforms has enabled the insurance industry to tackle some of its persistent problems with keeping the core elements of it intact. It has manged to make insurance industry more cost-efficient, customer-oriented by implementing usage-based policies. With the heavy use of data driven approaches the insurers have better insights about their customers which has made possible to create programs which are individual-specific, individual-centric and engage them in healthy behaviour in exchange of lesser premiums.