Predicting burglaries and other incidents

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Download date: 04. Oct. 2023

Predicting Burglaries and Other Incidents

Karthik Srinivasan
September **2015**

Predicting Burglaries and Other Incidents

Eindhoven University of Technology
Stan Ackermans Institute / Software Technology

Partners

Interpolis

Eindhoven University of Technology

Steering Group   Karthik Srinivasan
Sjak van Nieuwkuijk
Johan Ringeling
Joaquin Vanschoren

Date   September 2015

Document Status   Open Access

This design described in this report has been carried out in accordance with the TU/e code of scientific conduct

| Abstract | Predicting Burglaries and Other Incidents (PBOI) is a project to estimate the possibilities of burglary at a given location at a specified time duration. Crime prediction using fixed-location observation points are being used by police. Prediction using location-independent methods, which use a variety of environmental and other additional sources (data), is currently being explored by researchers.
PBOI is a pilot-project to study the feasibility of estimating burglaries using machine learning methods, which uses data from Interpolis/Achmea, Dutch police, TNO and a number of open data sources.
PBOI uses open-source tools to import, analyze, generate a model (using data), and generate predictions. The model is generated using a machine learning algorithm. The algorithm is used to model systems from historic data, which can be mapped to non-parametric functions with independent variables.
Based on the findings and a comparative study, the special algorithm performs better than others on the crime prediction domain. The predictions are populated on a map, which can help police and insurance professionals, to make informed decisions and to avoid burglaries and inform clients respectively. |
|---|---|

Foreword

Risico's, wie krijgt er niet mee te maken? Ons leven zit er vol van. Verzekeraars weten daar natuurlijk alles van, de gevolgen van risico's kunnen immers nagenoeg allen door een verzekering worden afgedekt. Overkomt je een risico en heb je een verzekering, dan ontvang je van de verzekeraar een financiële compensatie.

Interpolis is van mening dat verzekeren anders kan door klanten te helpen bij het voorkomen van risico's en daarmee de vervelende gevolgen ervan. Dit doen wij door klanten inzicht te geven in de risico's die men loopt en door hen preventieoplossingen aan te bieden. Verzekeren wordt daarmee een vangnet voor het geval je als klant toch met een schade geconfronteerd wordt.

Binnen het thema "inzichtelijk maken van risico's en het voorkomen ervan door preventieoplossingen" zijn wij continue op zoek naar innovatieve en vernieuwende oplossingen. Een van de initiatieven is het voorspellen van risico's –in het bijzonder woninginbraken- op basis van (big) data die we samen met de technische universiteit van Eindhoven opgepakt hebben.

Jaarlijks vinden er een slordige 75.000 woning inbraken in Nederland plaats. De laatste jaren zien we dit aantal fors afnemen, maar we spreken nog altijd over 200 inbraken per dag! Een inbraak heeft een gigantische impact op de gedupeerden. Los van de maatschappelijke kosten (geraamd op ruim 300 miljoen euro per jaar) levert een inbraak onrust en het gevoel van onveiligheid op, niet alleen voor gedupeerden, maar ook voor hun omgeving.

We zijn daarom blij met het model dat door Karthik Srinivasan ontwikkeld is om woninginbraken te kunnen voorspellen. Niet alleen Interpolis, maar ook de landelijke politie is benieuwd naar de uitkomsten van een aantal praktijktests die we gaan doen. Mogelijk dat de Nederlandse politie gebruik gaan maken van (delen van) het model. Op deze manier hoopt Interpolis in de nabije toekomst een bijdrage te kunnen leveren aan het verminderen van woninginbraken en daarmee het veiliger maken van Nederland.

Een speciaal woord van dank aan Karthik is op zijn plaats. Interpolis is onder de indruk van de manier waarop jij dit project opgepakt en vormgegeven hebt. Gedrevenheid, deskundigheid, enthousiasme en openheid zijn de eerste woorden die bij ons opkomen. Jouw persoonlijkheid en onvoorwaardelijke bereidheid om andere te helpen sieren je. Dank je wel Karthik!

Sjak van Nieuwkuijk
Interpolis Risico & preventive manager

Johan Ringeling
Innovatie manager Interpolis

Date: Aug 20, 2015

Testimonial

"In the research 'predicting burglaries' Karthik has shown that he has a thorough approach. The project involved many parties, such as the police, analysts and even burglars, so he used information from different angles. This has led to a good understanding of the problem.

Karthik developed a model that provides an indication of the chance of a burglary at a given spot. Hereby he uses the latest techniques in the field of machine learning. You can clearly notice that he is involved and skillful. He is an expert in his field.

In his model he uses open data and burglary rates. In addition, he has developed a module that determines the key factors that lead to an increased risk of burglary. Moreover, the model is easy to maintain and the results are presented graphically well. A very successful result!"

- Leon Vink
  Specialist Analytics / Data Scientist,
  Achmea

"Jij nam het initiatief tijdens een seminar over "Predictive Policing" op onze politieacademie om contact te leggen met politiemensen en kort iets te vertellen over je project bij Interpolis "Predicting Burglaries and Other incidents". Daaruit kwamen een aantal meetings met collega's van de nationale politie om te onderzoeken hoe we samen kunnen optrekken bij het bestrijden van dit fenomeen. We hebben daarin eerste stappen gezet, elkaars "werelden" ontdekt en gaan nu naar een fase van uitproberen!

Verkennen samen en leren van elkaars mogelijkheden tussen politie en bedrijfsleven is niet nieuw, wat bij jouw aanpak in het oog springt is dat jouw model vooral vooruit kijkt om criminaliteit te kunnen voorspellen. Je hebt een prima vergelijking gemaakt van soortgelijke modellen in de wereld en het jouwe verdient alle kans om te worden uitgeprobeerd.

Je hebt met je vriendelijke en degelijke aanpak een manier van werken die mensen aan je bindt, in combinatie met het door jou ontwikkelde model een prima combinatie.

We hopen in de 2e helft van 2015 te gaan uitproberen en hopen op mooie gezamenlijke resultaten!"

- Frank Wiewel
  Operationeel specialist C Intelligence,
  Politie, Landelijke Eenheid

"Karthik, you made it possible to identify geographic 'hot spots' of potential burglaries. With your research Interpolis has a concrete solution for customers with which they can prevent burglary.

I appreciate your excellent analytical and social skills. You fit in very well with the Interpolis workspace. Regularly you were the one who formed a big group of colleagues for lunch.

You also gave me an insight on the Indian culture and food. Especially your knowledge about the nutritional and health effects of different herbs and leafs was amazing. It was very nice to have met you and I wish you all the best in the future!"

- Joop van Leenders
  Market Intelligence Analyst
  Interpolis

"I met Karthik at an innovative project, "Predicting burglaries and Other Incidents", at Interpolis (Dutch leading insurance company). He is one of the most knowledgeable analysts that I have met and is an inspiration for the other analysts in different areas. Karthik has introduced me to a new way of data analysis with new techniques and software. He is able to inspire people because he is very knowledgeable and a very pleasant colleague.

His project is leading the way data analysis to be performed in professional organizations in the future."

- Jan-Willem Blank
Marketing Data Analyst,
Interpolis

"I have worked with and met Karthik at Interpolis because the project for predicting burglaries and other incidents is funded by our Innovation budget. It is a project that my fellow innovation manager coordinates, but I have met Karthik out of interest in the project. Karthik is very dedicated to the project and his promotion research. He is a very hard worker, is conscientious and is extremely knowledgeable about what he does. In this respect he brought a new competence to our company and the field of analysis in particular. As a data scientist he is a new breed of knowledge workers that will revolutionize many fields of business. It was an honor to be able to see a hint of that future through Karthik's eyes. But more important than any of this is the fact that I have learnt Karthik to be one of the most social, yet sometimes quietly so, and nice people I have met. He is open mindend, also has a very strong mind, and you can learn from his vision on life apart from the business side. But I think what characterizes Karthik most is the way he oozes respect for the people he comes into contact with. I enjoy the company of Karthik very much and have a great deal of respect for him as a professional and a private person."

- Bob van Leeuwen
Innovation Manager Business strategist,
Interpolis.

Preface

This document is an extended abstract of the technical report that holds the complete results of the graduation project *Predicting Burglaries and Other Incidents*, carried out by Karthik Srinivasan at Interpolis/Achmea, Tilburg, The Netherlands. The project was conducted as a full-time, nine-month assignment in the context of the technological designer program in Software Technology (OOTI Program). This PDEng programme is offered by the Stan Ackermans Institute of the Eindhoven University of Technology.

The report describes briefly the problem that is to be solved and the results. A detailed problem analysis together the complete results are documented in a separate confidential report, based on privacy/security grounds.

This report is intended for audiences who are interested in having a general overview of the project objectives and results. A glossary to support readers is provided as a part of the document.

Karthik Srinivasan

Date: Aug 20, 2015

Acknowledgements

Executive Summary

The main objective of this project is to study the feasibility of realizing the idea of Predicting Burglaries and Other Incidents. This is achieved by the performing the following tasks:

- Making an inventory of required (agreed) data sources.
- Creating a prediction model out of the data sources.

The result of the study is a proof-of-concept for building prediction systems. The prediction results can help us understand the possibility of a burglary at a given location at a specified time.

PRINCE 2 is used for project management applied over waterfall and agile practices. The project was a full time assignment, starting from January 2015 to September 2015 carried out at Achmea/Interpolis Tilburg. The project components were designed and developed as individual modules and are integrated. Four essential components of the system are:

- Data Import – to acquire data from data sources
- Feature selection – to identify the influential factors from the data
- Model Generation – to generate a model using the special algorithm
- Result Generation – to generate a report, which indicates the possibilities of burglary

The most interesting data sources are internal claims data from Achmea/Interpolis, history of Burglary information from Dutch Police, and a few other open data sources.

Information and knowledge gathered by brainstorming session, and by conducting workshops were used to identify the influencing features. Model generation uses a machine learning technique to generate the prediction. The prediction results generated using the model will be displayed on a map.

The prototype implementation was applied on generated data to test the functioning and functionality of the system. A number of design considerations were adhered to maintain quality and to improve the performance of the system. The turn-around time of the system on generated data was around 15 minutes.

As a future enhancement, the model can be hosted as a web service to perform similar predictions, broadcasting of predictions as SMS or e-mail notification.

Table of Contents

List of Figures

List of Tables

1.Introduction

Predicting Burglaries and Other Incidents (PBOI) is a project to estimate the possibilities of burglary at a given location and time. The system uses data from different sources to create a probabilistic-prediction model, which is used to predict burglaries. The system can also be used to predict other similar type of incidents.

1.1 Context

About Interpolis/Achmea

Achmea is a large insurance company that provides a wide range of financial services and products. Their brands serve about 12 million people with Health, Life and Non-life insurances. They focus on simple insurance products while keeping premiums affordable. And when the unexpected happens, they help customers manage their financial situation. Routes dating back to 1811, they have a co-operative background serving clients from Europe to Australia, with a gross written premium of euro 20 billion. Achmea has several brands as a part of the Achmea Holdings, were Interpolis is one such brand.

Interpolis is one of the largest insurance companies in the Netherlands. The company has gained wide recognition with its advertising campaign "Inter-polis.Crystal clear". Besides financial compensation, Interpolis also offers compensation in kind.

Interpolis, as an insurer in general, is working to not only pay out insurance money, but also to prevent damage altogether. In this concrete case they want to reduce burglaries by predicting them and warning the inhabitants of the property. For this Interpolis collaborates with other parties such as:

- Dutch National Police for domain knowledge and burglary data
- TNO for domain expertise
- City administration offices of Eindhoven and Tilburg
- Open data source providers namely, the ABZ, the BAG, the CBS, the KNMI, and the Bisnode.

The project's goal is to develop models and tools to predict burglaries. The project undertook a comprehensive scoping and feasibility study, analyzed data sources, made a repository containing historical data about burglaries and produced a prediction model which can work on the data to predict burglaries.

Objective

Among the broad objectives of the project, the notable ones are as follows:

- Make an inventory of possible data sources, analyze them on overlap and complementarity and support for the burglary indicators;
- Integrate related data sources (dynamically) and make them accessible to support further analysis and computation of indicators;
- Create a prediction engine for the known indicators. The engine should be extensible with new indicators and prediction models;
- Conduct a case study to validate the techniques;

Scope

- An application to predict burglary in advance, up to the post code level
- The warning messages delivery via email or social media or SMS (based on consideration for privacy, and public approval)

- Sources of information limited to Achmea, Dutch police, and a few other sources depending on need
- Application user interface in Dutch running on the windows environment
- Possibly: Assess privacy issues with the use of this data;
- Possibly: Extend the framework to predict other types of damage.
- Possibly: Add new/updated data in the future

1.2   Outline

The motivation to realize the project originated from suggestions from Interpolis, Dutch Police, Ministry of Security and Justice, The Netherlands.

PBOI is a pilot project to understand the feasibility of developing a prediction model. The security and privacy related to data usage are some of the notable constraints, in addition to the design, implementation, and analysis.

A brief discussion about stakeholder analysis is performed in Chapter 2. A detailed problem analysis is provided in Chapter 3. The feasibility study is performed in Chapter 4. A brief discussion about the system requirements is done in Chapter 5. Details about the system architecture is discussed in Chapter 6. Details regarding the system implementation is provided in Chapter 7. Information about the system verification and validation is discussed in Chapter 8. Conclusion and system results are discussed in Chapter 9.Project management information is discussed in Chapter 10. Details about the project retrospective is provided in Chapter 11.■

2.Stakeholder Analysis

In this Chapter, a rich list of stakeholders including those introduced in Chapter 1 is discussed. The process, methodologies, and techniques used to communicate, organization, and negotiations are discussed. The Chapter also discusses agreed quality controls, roles, associated interests of stakeholders, and considerations for their inputs.

Moreover, the protocols of standards, agreed in consultation with different allied stakeholders is discussed. Status reporting as per the agreed quality standards with the stakeholders is also discussed.

2.1 Introduction

Interpolis, the main stakeholder, is a part of Achmea the parent company. It supports the project by providing the necessary resources required for the execution of the project. The Dutch police, another stakeholder, is supposed to support the project by providing data about historic burglary events.

Both Interpolis and the Dutch police have varied interests as discussed in Chapter 1. Interpolis has as its objective the reduction of the number of claims made by its customers. In order to achieve this, Interpolis wishes to inform its customers about the possibilities of burglary as early as possible. The Dutch police, with the idea to understand the next possible location of crime, was interested in the project as well. The Earlier they know about a crime event the better they can plan to prevent the incident.

The next major stakeholder is the Eindhoven University of Technology. The university provides technological support to realize the idea. A PDEng trainee is deployed at the client location to design and implement the system. TU/e also supports the project by allocating an expert (Supervisor) in the field.

The different aspects of communication and other associated activities are discussed in the following sub-sections

2.2 Stakeholders Interaction
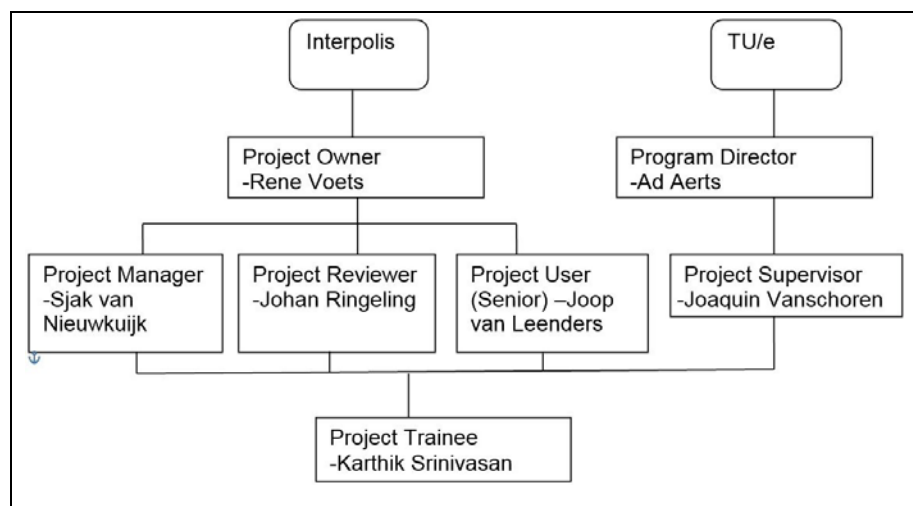
Figure 2.1 PBOI - Stakeholders Interaction

Individual stakeholders and their respective roles and interactions are represented in Figure 2.1.

- Interpolis– The industry/company with which the assignment is being done
- TU/e – The Eindhoven University of Technology, which deputes its PDEng trainee to execute his industry assignment

2.3   Profile Descriptions

- Project Owner – The person who takes care of sponsoring or financing for the project
- Project Manager – The person who takes care of arranging necessary resources to help complete the project
- Project Reviewer – The person who takes care of providing necessary domain and functional knowledge
- Project user (Senior) – The person who provides data analysis knowledge on possible usage from the user perspective, functional knowledge, and non-functional knowledge
- Project supervisor – The person who provides technological support from the university side.
- Project Trainee – The person who analyses, designs, implements, tests and generates reports of the system

Apart from the activities mentioned above, the following activities and associated members have a considerable role in the progress of the project.

Table 1 - PBOI - Stakeholders Activity and Contribution	
Activity	Contributor/Provider
Application test manager -Functional and non-functional	Johan Ringeling, Joop van Leenders, Sjak van Nieuwkuijk, Rinus Krijnen
Transition Manager	Sjak van Nieuwkuijk
System Administration	Karthik Srinivasan
DB Admin	Rinus Krijnen
Project Sponsor	Rene Voets, Interpolis
Project Steering Committee	Rene Voets, Sjak van Nieuwkuijk, Johan Ringeling, Joop van Leenders, Joaquin Vanschoren

2.4   Stakeholders Expectation
TU/e and Interpolis have different expectations from the project.
- TU/e – The University prefers the trainee to produce a working design.
- Interpolis – The industry prefers to have a prediction model.

2.5   *Communication Plan*

The mode (Table 3) and frequency (Table 2) of communication with stakeholders vary depending on the range of contribution they have on the system.

Table 2 - PBOI - Communication Frequency Plan							
Legend	**Information Type**						
A – As required/ updated D – Daily W – Weekly M – Monthly	Status Report	Issue log Report	Risk report	Change Request	Project Board Reports	Knowledge sharing Interpolis	Legal department notification report
Sjak van Nieuwkuijk	W	W	A	A	A	A	A
Johan Ringeling	W	W	A	A	A	A	A
Joop van Leenders	W	W	A	A	A	A	A
Rene Voets	A	A	A	A	A	A	A
Joaquin Vanschoren	W	W	A	A	A	A	A
Karthik Srinivasan	W	W	A	A	A	A	A
TU/e	A	A	A	A	A	A	A

Weekly and monthly reports are used to update the stakeholders about different information related to the project, which is presented in **Table 3**.

Table 3 - PBOI - Communication mode plan					
Item	Purpose	Frequency	Audience	Responsibility	Channel
Status Report	Progress notification	Weekly	Sjak van Nieuwkuijk, Johan Ringeling, Joaquin Vanschoren	Karthik Srinivasan	Status meeting
Issue Log / Report	To keep track of the issues	Weekly	Sjak van Nieuwkuijk, Johan Ringeling, Joaquin Vanschoren	Karthik Srinivasan	Status meeting

Risk report / Register	Risk analysis and mitigation	As required	Sjak van Nieuwkuijk, Johan Ringeling, Joaquin Vanschoren	Karthik Srinivasan	Person-nel meeting

Stakeholders preferred mode of communication varies based on their availability and document type.

2.6 Stakeholder priority

Based on the project controlling power and interest, the stakeholders are classified onto a grid as shown in Figure 2.2.

Figure 2.2 PBOI - Stakeholder Power-Interest Grid
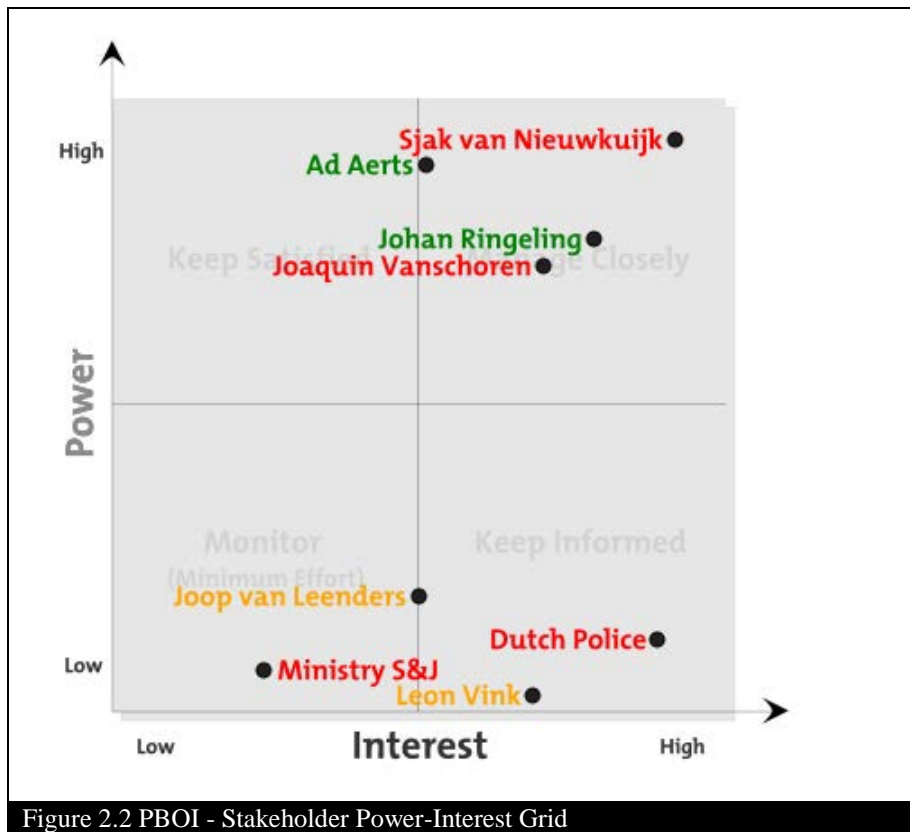
Stakeholders plotted in the high power region have higher influence on the project execution and management.

2.7 Agreed Approach

The project is executed based on methodologies adopted by Interpolis, and as per the regulations agreed upon by the stakeholders. The model was constructed in-house, using data sources from different parties as mentioned in Section 1.1 .
Different phases of the project include:
- Literature review
  - Explore existing models/ algorithms that are in use

- - Find the suitable data from among the vast pool, which could be more useful.
  - Identify appropriate tools for data analysis.
- Analysis
  - Understand and explore the available data
  - Identify features for classification
- Design
  - Develop a suitable design for the system
- Build
  - Implement the system - Prediction model
- Test
  - Verify and Validate the model
- Evaluate and Analyze
- Report the lessons learned

2.8   *Project Deliverables*

Major deliverables agreed with industrial stakeholders include,
- Project Plan
  - Project Management documents
  - Timing and resource document
- Requirements document
- Design Document
  - System design document
- Status report
  - Weekly status reports
  - Project status report
  - Test cases report
- Prediction Engine
  - Source code, application
- Documentation of operational maintenance
- Project learning document■

3.Problem Analysis

In this Chapter, a brief discussion about the problem analysis is performed. The requests, suggestions, methods proposed and agreed with stakeholders are considered. The analysis of the problem is based on the stakeholder analysis done in Chapter 2.

3.1   Context

Predicting burglaries is a new area of focus in the crime prevention world. An increase in the number of burglaries in turn is expected to increase the financial (insurance) claims by victims. This influences the budget of every insurance company. Burglary prevention can contribute to a safe living environment and reduce claims which has an effect on revenue of the insurance company.

Software tools to predict burglaries exists in a limited number. One such tool is Crime Anticipation System used by Amsterdam Police. Tools have their advantages and disadvantages, which is also linked with process, methods and techniques used.

The project needs to use data from different sources, making it necessary to identify the most suitable tools for data preprocessing and analysis. Furthermore, it is important to consider tools that can classify, create prediction models, and evaluate the model.

The basic constraints on the tools include
- Ease of use for non-programmers
- Software tools with free commercial license
- Lesser learning time than those currently used tools
- Less implementation and installation time
- Adaptability
- Ability to choose algorithms
- Ability to extend multiple other popular tools
- User interface similar to the user interface provided by tools used in Interpolis
- Ability to operate on and interact with other tools

3.2   5 step approach

In order to propose a solution for the problem, we need to understand the problem and user needs.

3.2.1. Gain Agreement

The problem statement description is show in Table 4

Table 4 - PBOI - Problem Analysis - Gain Agreement	
Element	Description
Problem	Burglaries at residential apartments and other types of residential buildings
Affects	• Customers property loss • Revenue of Interpolis due to claims • Citizens safety and wellbeing

	• Trust people have in the Police
Results in	• Increased Claims
	• Disturbs safe living of citizens
Benefits of a solution	• Increase revenue,
	• Increases mental health level of people
	• Increases public trust in the Police

3.2.2. Root Cause Analysis

One of the reasons for burglaries is the lack of a mechanism to predict burglaries in advance. This is hampered by the lack of central/distributed data about burglaries.

Table 5 - PBOI - Root Cause Analysis

Problem to solve	Technique Applied
Burglary & Related Information	Obtain & Integrate Data from different sources
Prediction System	A New software solution

3.2.3. Users

The system users vary from different level of management. The users are from both police and insurance fields as shown in Table 6

Table 6 - PBOI - System Users

Users	Field/Domain
Market Analysts	Insurance
Business Managers	Insurance
Intelligence Officer	Police
Patrolling Team	Police
Safety Watch Team	Police
Premium calculator • Statisticians • Market Preposition Manager	Insurance

3.2.4. System Boundary

The interaction of different users with the system boundary are shown in Figure 3.1.
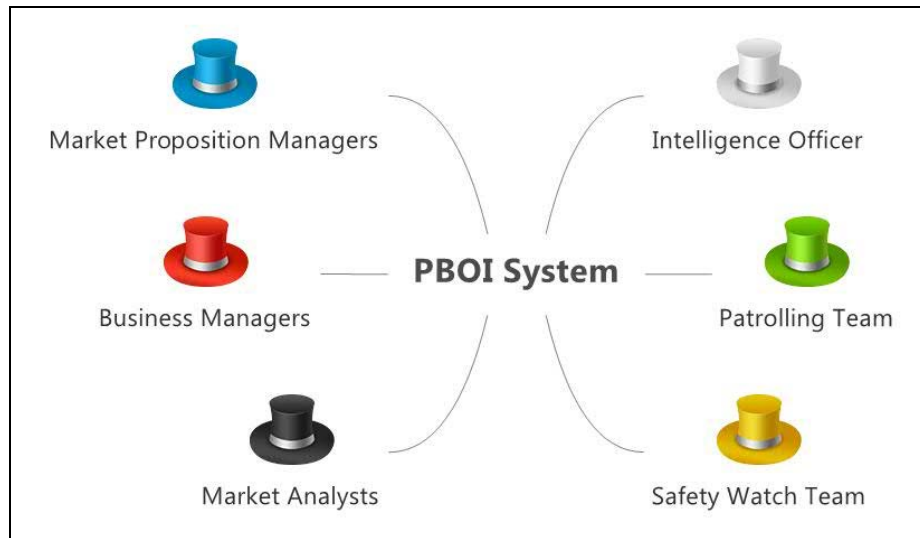
Figure 3.1 PBOI - System Boundary

Market Proposition Managers – use the system as a value addition for advertising the insurance policy.

Business Managers – use the system to make decisions, thereby deciding on the premium of the insurance for a given location.

Market Analysts – use the system for understanding the possibility of a burglary at a given location, or for verification of the claims.

Intelligence Officer – use the system to gather information while investigation.

Patrolling Team – use the system to observe burglary prone areas.

Safety Watch Team – use the system to estimate the safety of a location and to suggest new plans which could improve safety at a location.

3.2.5. System Constraints

Various key system constraints exist and they are explained in Table 7

Table 7 - PBOI - System Constraints	
Source	Consideration
Economics	• Use as little finance as possible, apart from the TU/e costs • Use free license as much as possible • Use in-house resources as much as possible
Politics	• Tackle and handle the existing interdepartmental issues in a smooth and professional way
Technology	• Use Open-Source technologies as much as possible. • Use R Programming for Model Generation
System	• The solution is a new and fresh implementation • The solution should be implemented on Mac OS X • The solution should have a GUI which must be user friendly for non-programmers
Environment	• The implementation should be on a standalone computer • Legal constraints, o Privacy of data being used.

	o Privacy of algorithm being used
	o Privacy of methods being used.
Schedule	• 9 month duration, 1,0 FTE
Resources	• No external paid resources

3.3   *Challenges and project approach*

The main challenge in this project is to generate a model which can mimic the behavior of burglars. The more knowledgeable the model is the higher is the prediction accuracy. Generating a model with knowledge depends on the amount of data and the variety in data, while maintaining the inter-relationships among the data used. Based on literature review, it is suggested to train the model with an agreed level of knowledge to predict burglaries.

Another challenge concerns data sources. When importing data from different sources, the data is mostly diverse. This necessitates the data-import implementation to be specific to each incoming data source.

A next challenge is to cope with the increasing amount of data. Due to hardware constraints and increasing memory requirements, the system is made to process data based on region level. This approach addresses the huge memory and computation requirement issues.

Domain knowledge expertise was another challenge at the start of the project. This expertise was furnished by including Police Officers and Ex-Burglars to help and validate the influential factors for our model. The workshop questionnaire obtained during the workshop conducted at Interpolis helped us to obtain their views and ideas on selecting a target.

User related challenges include a GUI based model generation tool. Most users at Interpolis are non-programmers. Based on the literature review (Chapter 4) and requirements (Chapter 5), an open-source tool which could resolve this challenge and satisfy the need was chosen.

As for the computation related challenge involving the proper utilization of computing resources, we have been using JAVA for automation, while the computation power of JAVA is good for CRUD operations. Big data computation requires capable tools which can utilize the computing resources in a best possible way. One of the open-source statistical tool is considered to be the best open source tool for data analysis and modelling. The integration of open-source tools and JAVA helped us solve this challenge.

3.4   *Design Opportunities*

The report, New criteria for assessing a technological design Functionality, authored by Kees van Hee and Kees van Overveld, April 2012 (Kees van Hee 2012) provides techniques to access a technological design. The most important design aspects or criteria's suggested in the report were chosen and are discussed based their importance to this project.

The important design criteria for this project are:
* Functionality
* Construction
* Realizability

Functionality – is the most important design criterion relevant to this project. The project focuses on a technological design of a complete artifact, where the formal

requirements are to be implemented to the satisfaction of the client. At the same time, the resulting tool should be easy for use, by non-programmers and people with limited computer literacy. Above all, individual modules of the project are supposed to be reused for other prediction systems.

Construction – the next vital design criterion relevant to this project. The system modules or components should be developed with minimum to no hierarchy, and at the same time measures are to be taken to provide a high degree of independence between components. This makes the components to remain as individual projects with provided interfaces. This represents the projects reusability for other similar or non-similar domains, which includes a prediction system development.

Realizability – another essential design criterion relevant to this project. The Implementation of the project should be done on JAVA technology utilizing the power of other statistical and modelling tools. The project will be introduced into business for real time prediction.

4.Feasibility Analysis

A mathematical model is considered to be better if it has better prediction accuracy. The prediction depends on the training data set available for learning. A model with more features tends to become complex and consume a considerable amount of resources for its computation. In contrast, feature reduction can help us identify the important features that can provide the same prediction, while utilizing less resources for computation.

Considering the literature from existing journal publications, book articles, and lecturer notes for prediction models of a similar nature, the following algorithms have been explored:

- Crime prediction classification models that take advantage of spatial and temporal data to make reliable prediction are J48, 1NN, SVM, Naïve Bayes
  - Out of the compared models, **Naïve Bayes** showed better results and outperformed the other three models (Chung-Hsien Yu 2011).
- Comparison of crime prediction techniques on the insurance domain was dominated by **Multilayer Perceptron** when compared to decision trees, logistic regression, and Naïve Bayes (Danso - 2006).
- **Radial Basis Function kernel** is considered when the requirement is a non-linear model which can learn a less known domain/area.
- Bird call recognition using ANN, SVM and KDE is used to classify recorded bird calls. Out of the three algorithms used, **SVM-FAR** performed way better than KDE and ANN (Ross 2006).
- Upon comparing the following models for time series forecasting,
  - Multilayer perceptron
  - Bayesian neural networks
  - Radial basis functions
  - Generalized regression
  - Neural networks (also called kernel regression),
  - K-nearest neighbor regression,
  - CART
  - Regression trees,
  - Support vector regression
  - Gaussian processes.
    the **Multilayer perceptron** and **Gaussian process** scored well, followed by SVM for classification (Nesreen K. Ahmed 2010).
- Comparison of Machine learning strategies for time series forecasting using:
  - Decision trees
  - SVM
  - NN Regression
    proved that the **decision trees** remained as the perfect model for time series forecasting (Gianluca Bontempi 2013).
- A combination of Bayesian variable selection methods (Scott 2013) using the following
  - Kalman filter – to control time series feature
  - Regression – to incorporate predictors via interacting with external interfaces
  - Model Averaging – to reduce the over fitting
    is also considered for generating a model
- Two level feature selection using
  - Chi square – filter based feature selection
  - Forward Feature selection and Backward Feature Elimination – wrapper based feature selection

The methods above, when used in combination with Chi Square + Backward Feature Elimination, improves the accuracy of **Naïve Bayes with KDE**

- **PredPol – Standard density estimation** replaced with **Non-Locally regularized** H1 MPLE could yield better results (J. T.Woodworth 2014)
- **PredPol** – Problems with KDE can be overcome by the use of TV and H1 regularized MPLE. The type of kernel function used is the Gaussian distribution (Laura M. Smith 2010).

From the above literature review, it was considered to choose the special algorithm for our model.

4.1   *Risks*

The risks that are maintained on a risk log which are tagged based on their impact level are represented in the following table. The corresponding mitigation plans for respective risks are also mentioned in **Table 8**

Table 8 - PBOI - System Risks		
Risks(Description)	Impact	Mitigation
Data availability	Major	Develop the system using simulated data, till the original is received
Business Knowledge Expert	Major	Approach experts, presentation & discussion
3rd party agreements	Medium	Explain business benefits to a Common gathering
Tools installation	Medium	Initiate in advance, Achmea IT helpdesk
Tool selection	Minimum	Transfer knowledge from Data Analyst, Literature Review
Feature Selection - relation/Association	Major	Discuss with project board, to arrange a workshop
Algorithm Choice	Minimum	Discuss with Project Supervisor, literature review
Data from 3rd parties	Major	Discuss with project Manager, arriving on agreement
Documenting all possible observations	Minimum	Discuss with Project Supervisor
Tools for accessing Achmea data	Minimum	Initiate in advance and track IM & IT, Shared folder
Inventory creation	Minimum	Structure Data from different sources
Data table relationship (Keys)	Minor	Device a common key - suggest the project board a list of best keys
Feature Vectors	Medium	Iterate over every feature selection, and feature elimination Workshop - invite police, ministry, ex-burglars, who can help in identifying the best features

5.System Requirements

Prediction systems in general have four basic requirements, namely, data processing, feature selection, model generation, and result generation. The PBOI project also has the same functional requirements. The business requirement depends on the type and usage of the system. The current project has crime as domain and burglary as the subdomain.

5.1 Introduction

The system basically has to import data for analysis. The imported data has to be preprocessed. The pre-processing is meant to make the data more useful for the system. Next, the prepared data is considered, where the influential factors are determined. Once the influential factors are ready, the model is generated by means of a machine learning algorithm.
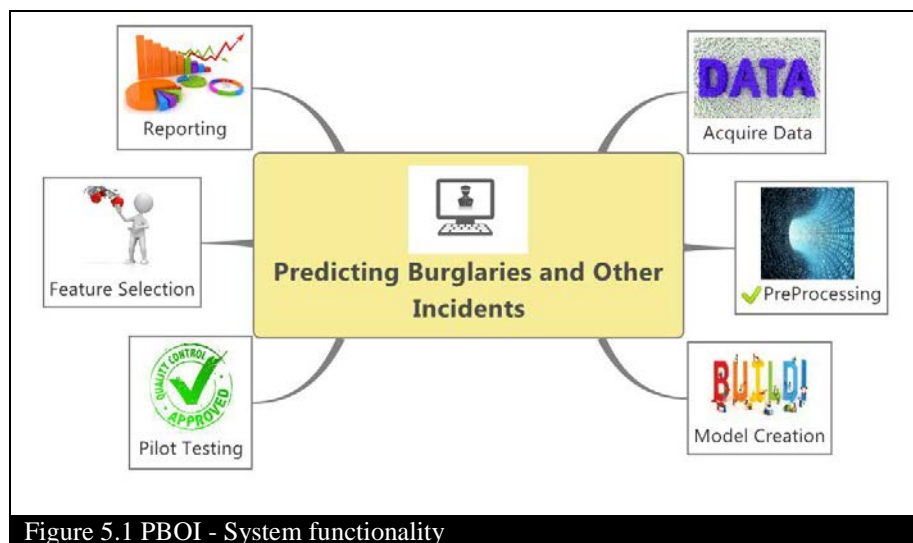
Figure 5.1 PBOI - System functionality

5.2 Requirements Management

The MoSCoW method is used to indicate priorities. This is the most widely used method in software development to arrive at a common understanding between stakeholders and to prioritize the expected deliverables.

A brief about MoSCoW method is as follows:
- MUST – a list of requirements that must be delivered as a part of the final system
- SHOULD – a list of requirements that should be a part of the final deliverable. This could also be compensated with other requirements being satisfied.
- COULD – a list of requirements that is considered desirable but depends on the time and other resource availability.
- WON'T – a list of requirements that are not a part of the system as agreed with stakeholders.

5.3   High level requirements

To understand the requirements the difference between a domain model and a mathematical model are explained as follows

- Model (Prediction/Mathematical) – a machine learning model, which is generated using an algorithm (based on the domain suitability). The model acquires knowledge from the training data.
- Domain Model – is a pictorial representation of interaction and behavior between different entities of a system

Table 9 - PBOI - System requirements prioritization	
Priority	Requirements
MUST	Develop a prediction/mathematical model for predicting burglaries
	Create and maintain inventories for different data sources
	Create a Feature selection system
	Create a decision generation system
SHOULD	Automate the tasks
	Create a GUI
COULD	Future updates
	SMS notification to clients
	e-Mail notification to clients
WONT	Predicting other types of damage

5.4   Product Description

The project scope is to build a model for predicting burglaries. The project is divided into sub-modules namely:

- Data Acquisition System
- Feature Vector Selector
- Algorithm Selector
- Decision Reporter

each module is explained in the coming sections.

Data Acquisition System
A module to acquire data by parsing the following file types:
- CSV file
- Excel file
- Text file

Quality Expectation	Data Acquisition System	Priority	M
Acceptance Criteria	ACDA1: Acquire data from database dump file		
	ACDA2: Acquire data from flat file		
	ACDA3: Acquire data from CSV file		
	ACDA4: Acquire data from Excel file		
Tolerance	None		
Acceptance method	ACDA1: Testing, Data imported from database dump file		
	ACDA2: Testing, Data imported from flat file		
	ACDA3: Testing, Data imported from CSV file		
	ACDA4: Testing, Data imported from Excel file		
Responsibilities	Karthik Srinivasan		

18

Feature Vector Selector

Module that helps the user to select features for training the machine learning model. Two possible feature vector selectors are:

- o   Manual
- o   Automatic

Quality Expectation	Feature Vector Selector	Priority	M
Acceptance Criteria	ACFS1: Select list of feature vectors		
Tolerance	None		
Acceptance method	ACFS1: Testing, List of feature vectors available for Classification Algorithm		
Responsibilities	Karthik Srinivasan		

Algorithm selector

This module presents the user with a list of algorithms, which could be used for model generation.

Quality Expectation	Classification Algorithm selector	Priority	M
Acceptance Criteria	ACCS1: Select a classification algorithm		
Tolerance	None		
Acceptance method	ACCS1: Testing, Classification algorithm selected for performing the decision		
Responsibilities	Karthik Srinivasan		

Decision Reporter

In this module, the user has the facility to generate results using the model that is generated in the model generator module.

Quality Expectation	Decision Reporter	Priority	M
Acceptance Criteria	ACDR1: Report the decision results		
Tolerance	None		
Acceptance method	ACDR1: Testing, Results are made available to intended audience		
Responsibilities	Karthik Srinivasan		

■

6.System Architecture

In this Chapter, the system architecture is discussed. Considering the requirements and the other factors, a literature review was performed and reported in Chapter 4. Taking into account the facts from the Chapter, the system architecture is devised.

6.1   Introduction

Based on the need and constraints, the prediction system was developed. The system addresses different issues ranging from modularity to extendibility.

6.2   Process View

The view represents the run time aspects of the system. The statistical computation tool's process has an increased influence on the performance of the system. The scalability and throughput partially depends on the hardware resources.

- Performance
  - o The system performance can be measured based on the load the system can respond to, which depends on the chosen algorithm and the number of training and test data sets.
- Distribution
  - o The system is assumed to be a standalone system, which can be used for generating the prediction results.

Moreover, there are possibilities to host a server.

6.3   Design Decisions

The System architecture, based on the requirements and domain, was developed while considering various decisions agreed during the course of the development.

The main reasons for those decisions are discussed as follows

> **Decision:**
> Combine data sources into a single table in a database

In order to make the data available to be used by the system, the data has be made smooth and structured. Preprocessing is essential for data analysis related tasks. During the process of preprocessing, the data to be persisted is stored in the same table by including extra fields through the use of a unique key. This eliminates the high cost join-operation during queries. To a certain extent, the data integrity is also maintained.

> **Decision:**
> Persist data into the database after every computation in each stage of the system

Considering the available computational resources, it is better to store the data after every computation. This actually lowers the memory requirements and computational capacity of hardware resources. This allows parallel-process-initiation of other suc-

cessor modules. Dividing the stages at the data-layer level provides easy analysis and future reporting possibilities. This also eliminates multi-indexing need for multiple tables on the database.

> **Decision:**
> Automate Feature Selection

Feature selection in any machine learning model is considered to be a determining factor for the accuracy of the produced results. Human error in selecting those features has a high impact on produced results. A better approach is to use an algorithm that can select feature based on agreed techniques; namely, correlation, high relevance, and low redundancy. This helps to reduce the dependence on business/domain experts on every stage of the system development.

> **Decision:**
> Analyze Region-wise data

The data is huge when considering the entire Netherlands. In order to reduce the size and computational load on the available hardware resource, the data is analyzed based on regions (e.g., Eindhoven, Tilburg). Also the crime could vary based on the location. This led to the decision of building a model per region.

> **Decision:**
> Separate implementation for data import

Data import is possible using the open-source statistical tools, but due to the size constraint and hardware availability the system had to be modified. The JAVA implementation uses a special library, which manages by streaming the data sources, thus eliminating the need for large and powerful resources.

> **Decision:**
> Save the generated model as general format file

The model once generated can be used and ported to another location or platform where the model can be reused. The idea is to generate the model, which can utilize the knowledge from different sources of available data. The model generation consumes valuable resources (e.g., time, hardware). The generated model can be used and remains the same until there is new data from agreed sources, or there is a new data source(s). The format used is platform independent and portable■

7.Implementation

In this Chapter, the realization of the requirements defined in Chapter 5. The implementation is based on the system architecture and design, discussed in Chapter 6.

7.1   Introduction

System implementation is carried out as four components. Each component is devised to have high cohesion and high independence on other.
Different modules of the system are

- Data Import
- Feature selection
- Model Generation
- Result Generation

7.1.1. Data Import

The data import module has important responsibilities. The better the data is structured, organized, and preprocessed for analysis, the better the model can learn. The data import module is implemented using JAVA as decided in Section 6.3   .
The data is imported and preprocessed for making it more organized and structured. The data is then persisted into a database as individual tables for each data source. Consider Weather data being imported, in this case the data is persisted into the database in the weather table.
The Front-End GUI developed using JAVA Swing, can help users interact with the system. The GUI interacts with JAVA applications and communicates through the open-source tool's Interface to perform data import.

7.1.2. Feature Selection

The features were selected from organized and structued data source. The selected features are persisted into a database. The features remain same until there is a new data point, or new feature to be added, or new feature selection technique has to be applied.

7.1.3. Model Generation

Model generation is achieved by using the power of an open-source tool, which is one of the best available tools on the open-source market for statistical analysis. The JAVA based GUI communicates through the interface to communicate with the server, which can perform high load computations for JAVA.

7.1.4. Result Generation

The prediction results are the essential product that stakeholders are eagerly looking for. The system generates results by utilizing the knowledge learned by the model, which was generated in the previous stage 7.1.3.

The model is imported using an open source tools package and the predictions are generated by querying the model with test data sets. The predictions are stored into a database, which will be fetched latter using the open source package. The results are populated on a map, which could give a visual picture for the interested parties to analyze burglaries. Latitude and Longitude information is used to project the results on a map.

8. Verification & Validation

In this Chapter, we discuss about the methods and techniques used for the verification and validation of the system.

8.1   Introduction

Most processes followed as a part of the project management had a high influence on verification and validation.

8.2   Verification

The communication plan discussed in Section 2.4   is a means to ensure verification of the project building process. Each step in the project building is closely watched by stakeholders.

The Project plan that was devised at the start of the project is a checklist to be closely watched, to adhere to the quality standards and agreed requirements.

Stakeholders are updated as per the communication plan, which gives them opportunity to advise of any changes that are caused due to any deviation from the projected plan.

The product description discussed in Section 5.4   provides a means to ensure that the product meets the quality expectation of the client.

Testing is used to verify the system,
- Functional testing
  - Performing a black box testing to make sure the agreed functions are present the system
- Structural Testing
  - Performing a white box testing to verify functions of individual components of the system

8.3   Validation

Agreed project deliverables discussed in Section 2.8   assures that the product developed is as per the requirements and needs of the client. This is asserted by means of user acceptance testing and those steering group meetings, where the product under development is presented.■

9.Conclusions

In this Chapter, we perform a discussion about the results and achievements of the project, followed by recommendations for future work. The project deliverables are discussed on Section 2.8

9.1   Results

The result of this project is a proof of concept, which gives a concrete indication that the prediction of residential burglaries is possible. The accuracy of the prediction depends on various factors ranging from feature selection to model generation.

The output of the system is the prediction results, which can be
- Populated on maps
- Used for analysis by statisticians

The genericity of the components developed makes them easy to be used for other prediction systems.

9.2   System Output

The following figures represents the output (prediction) generated using the system based on the simulated data. The p are predictions for a given day in the Eindhoven region.
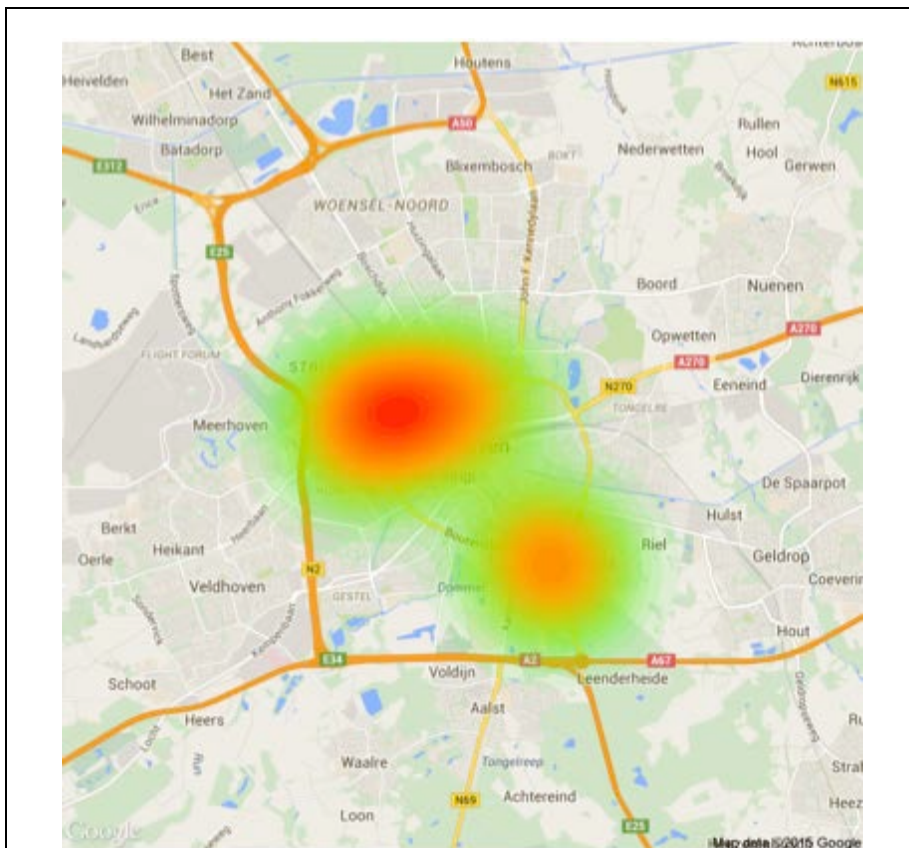

Figure 9.1 Burglary Prediction Day 1 Eindhoven
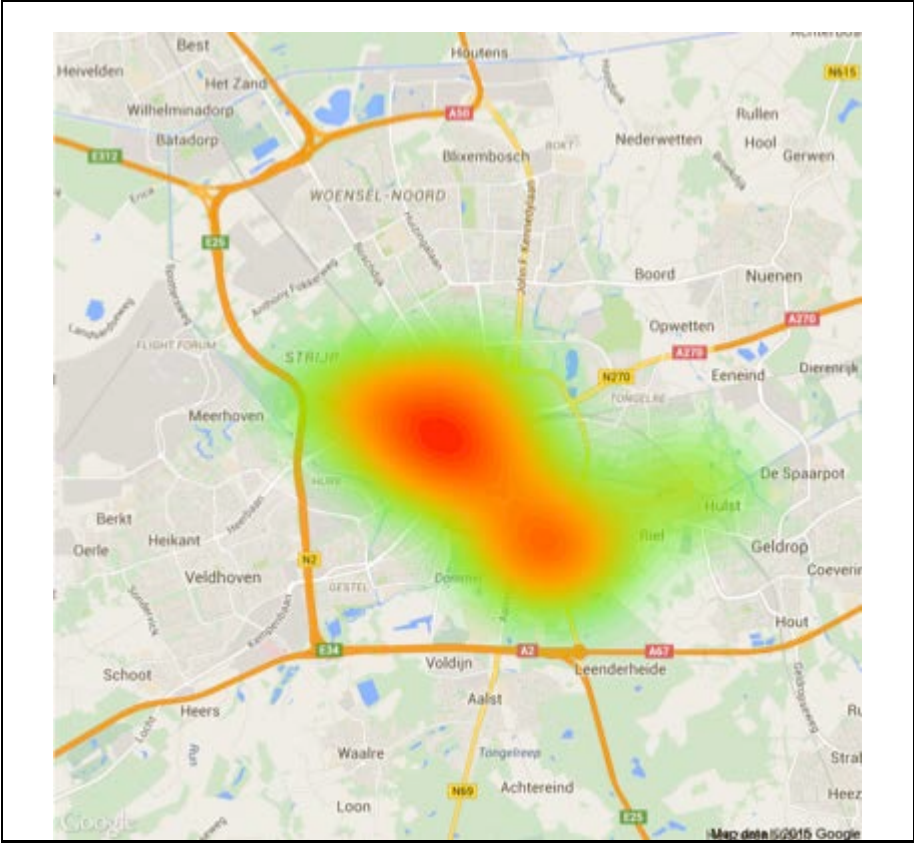
Figure 9.2 Prediction Day 2 Eindhoven
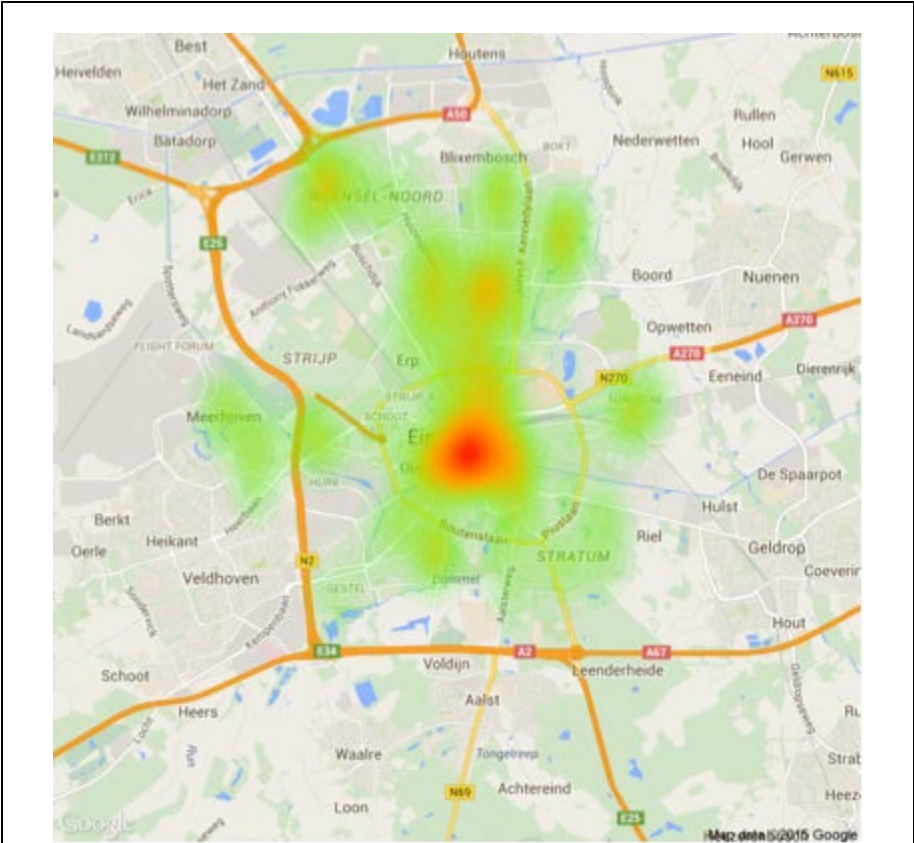
Figure 9.3 Burglary Prediction Day 3 Eindhoven

9.3 Recommendations for Future Work

Considering the results of the system, the following can enhance the business scope of the system.

- Open-source analytics server to host the project components as a web service.
- Broadcast the prediction results, to interested customers either as SMS or via e-Mail.
- Acquire more and different types of data sources to make the model better equipped with knowledge.
- Integrate algorithms better into the system and provide a hosted service for prediction■

10. Project Management

In this Chapter, we discuss various methods and techniques used for the project management.

10.1   Introduction

The project uses PRINCE 2 as project management methodology. The project tasks are divided based on PRINCE 2 stages. Each stage has its own specific achievement in advancing the progress of the project.

10.2   Project Planning and Scheduling

The project is divided into stages based on the PRINCE 2 methodology. Different stages of the project contribute towards the progress of the project. The project started its realization by means of waterfall way of working. The second part of the realization starting from Stage 4 changed to Agile. This shift in the methodology is due to the requirements.

Table 10 - PBOI - Project Schedule			
Task Name	Duration (days)	Start	Finish
Predicting Burglaries and other incidents	184	5-Jan-15	24-Sep-15
Startup and Direction	2	5-Jan-15	6-Jan-15
Initiation Process	6	7-Jan-15	14-Jan-15
Stage 1 (Planning)	10	15-Jan-15	28-Jan-15
Stage 2 (Literature Review)	11	28-Jan-15	13-Feb-15
Stage 3 (Analysis)	28	13-Feb-15	25-Mar-15
Stage 4 (Design)	31	25-Mar-15	12-May-15
Stage 5 (Build & Test)	86	12-May-15	11-Sep-15

Startup and Direction
The project started with an introduction of Interpolis/Achmea and its facilities. Stakeholders, management, and colleagues were introduced. Every new candidate entering the Achmea/Interpolis premise for executing a project is introduced to the rules and regulations of the company.

Initiation Stage
In this Stage the Project Initiation Document is prepared, this document has the requirements and scope of the project. The agreed, first version of the requirements and initiation report is presented to the project board, for approval.

Planning Stage
The Project, has to be planned to adhere to the process of PRINCE 2 management style. A number of documents and rules are established to track the quality, status, exception reporting, and progress of the project. Those documents were used in all the stages of the project.

Literature Review

Based on published proof from researchers, the software, and tools for mathematical modelling and analysis were chosen at this stage.

Analysis Stage

Available data sources and other resources that can contribute towards the feasibility study of the project are analyzed at this stage. The feature vectors which could influence the prediction capabilities of the system were also analyzed.

Design Stage (Agile)

System architecture and software design is performed at this stage. The possibility to substitute certain functions by means of a pattern is also analyzed.

Build & Test (Agile)

System Components were built and tested in this stage for the design developed in the previous stage.

Deployment

System deployment and handover to the client is performed at this stage

10.2.1. PERT Charts

Program Evaluation and Review Technique, which was developed by U.S. Navy, is one of the best project management tool used to schedule, organize, and coordinate tasks within a project. A screen shot Figure 10.1 of the MS Project PERT Chart for a portion of Stage 5 Build and Test
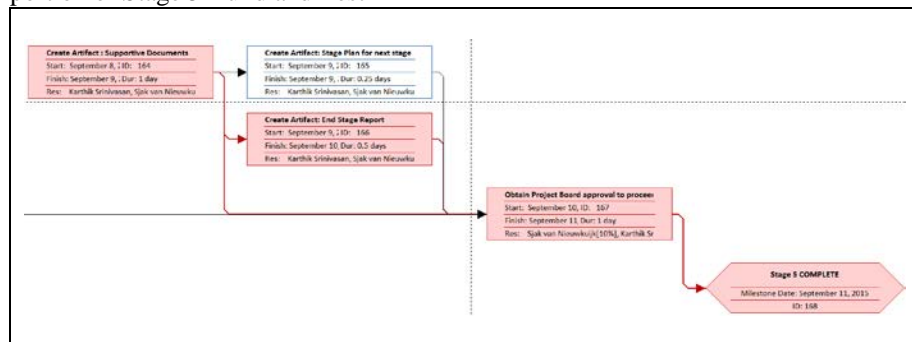

Figure 10.1 PBOI - PERT Chart - part of Stage 5 Build & Test

10.2.2. Gantt Charts & Work-Breakdown Structure

The project uses MS Project tool to track and schedule the project tasks. A Gantt chart helps us by depicting the start and end date of the terminal and summary tasks of a project on a bar like structure. Gantt chart for Planning Stage is in the Figure 10.2
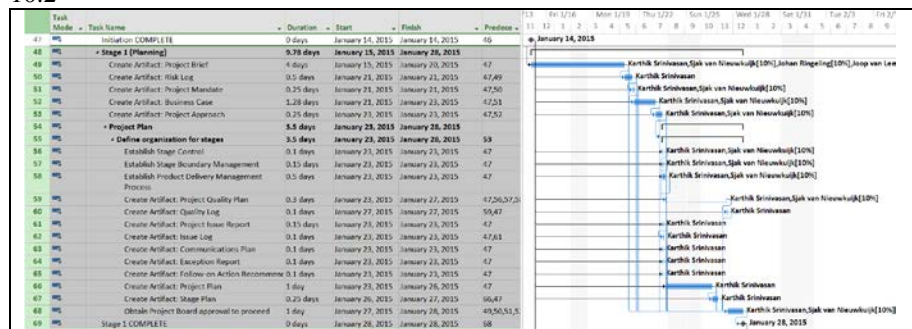

Figure 10.2 PBOI - Gantt & Work-Breakdown Structure - Stage 1 Planning

The Work-Breakdown Structure gives us an overview of sub-tasks that are present inside a given stage of the project. The Figure 10.2, represents the work-breakdown for the Planning Stage.

■

11. Project Retrospective

In this Chapter, we discuss about evaluating the project performance, identifying the lessons learned, and making recommendations (based on project experience) for future.

11.1   Introduction

Project success can be determined based on a number of factors, which include
- Time
- Cost
- Value
- Outcome
- Data availability

A project can incur differences on the criteria mentioned above, but still be considered as a success, when viewed from the overall view. ∎

11.2   Design opportunities revisited

Design criteria relevant for this project are discussed in Section 3.4
- Functionality
- Construction
- Realizability

Functionality – the requirements defined in the product description are very essential for the project. Accomplishment of formally defined requirements are vital for client satisfaction. The prototype implementation of the system represents the list of agreed functionalities realized. The prototype GUI developed for automation, and the prototype GUI for design of mathematical model developed are very close in resemblance to the proprietary tools that are used inside Achmea/Interpolis. Thus making it easy for Interpolis professionals to learn and use on a short time. The high independence and genericity followed in the design makes the system components eligible to be plugged in for other inter-domain model developments.

Construction – the high cohesion nested in each module, helps us to host each module as an individual project. High cohesion also adds to the possibility of using the system for inter-domain model building. The interest of Achmea/Interpolis to use the components for customer behavior analysis in marketing department represents their usability.

Realizability – the prototype implementation using JAVA for automation, and open-source tools for statistical computation supports the proof-of-concept.

Glossary

University	Eindhoven University of Technology
TU/e	Eindhoven University of Technology
Client	Interpolis/Achmea
Police	Dutch Police/ Politie.nl
Ministry	Ministry of Security and Justice, The Netherlands
Prediction Model	A machine learning model, that can learn from historic data
GUI	Graphical User Interface

Bibliography

References

Java   [Online].   n.d.   4   August   2015.
<https://www.java.com/en/download/help/download_options.xml>.

Kees van Hee, Kees van Overveld. *New criteria for assessing*. Eindhoven: SAI,
Eindhoven University of Technology, 2012.

PredPol. *PredPol [Online]*. n.d. 04 August 2015. <https://www.predpol.com/>.

About the Authors

Karthik Srinivasan graduated from the Department of Computer Science and Engineering, Anna University in 2007. He worked as a senior software Engineer from 2007-2010 on Travel and Transportation domain at a leading Indian IT company for projects with Unisys Inc., USA, ANA Japan and ACS [Xerox] USA, and also received his M.B.A. from the University of Madras. He received his M.Tech., from the Centre for Excellence in Computational Engineering and Networking, Amrita University in 2012. His work on Image enhancement and segmentation was published in International journals. As a student at Amrita University he was keen in establishing and developing communication facilities for the poor, living in remote and rural south India. From 2012 - 2013 he was working as an Assistant Professor. His interests include e-governance and Applications of ICT in health care and education.

From September 2013 until September 2015, Karthik Srinivasan worked at Eindhoven University of Technology, as PDEng trainee of Software Technology program from 3TU.Stan Ackermans Institute. During his industrial design project, he worked for Achmea/Interpolis in collaboration with the Dutch Police on a project to predict burglaries.

3TU.School for Technological Design, Stan Ackermans Institute offers two-year postgraduate technological designer programmes. This institute is a joint initiative of the three technological universities of the Netherlands: Delft University of Technology, Eindhoven University of Technology and University of Twente. For more information please visit: www.3tu.nl/sai.