

An Investigation into Big Data Analytics Applied to Insurance

Rebecca Peters, University of South Wales; Penny Holborn, University of South Wales

ABSTRACT

Data is generated every second. The term Big Data refers to the volume, variety, and velocity of data that is being produced. Now woven into every sector, its size and complexity has left organizations faced with difficulties in being able to create, manipulate, and manage Big Data. This research identifies and reviews a range of Big Data techniques available through SAS, highlighting the fundamental opportunities that SAS provides for overcoming a variety of business challenges. Insurance is a data-dependent industry hence this research focuses on understanding what SAS can offer to insurance companies and how it could interact with existing customer databases and online, user-generated content. A key data source has been identified for this purpose. The research demonstrates how models can be built based on existing relationships found in past data and then used to identify prospective customers. Principal Component Analysis and Logistic Regression are considered within this research. This paper demonstrates how these techniques can be used to help capture valuable insight, create firm relationships, and support customer feedback. Whether it is prescriptive, predictive, descriptive, or diagnostic analytics, harnessing Big Data can add background and depth, providing insurance companies with a comprehensive understanding of their customers. This paper highlights how reducing the complexity and dimensionality of data can provide actionable intelligence to enable more informed business decisions.

INTRODUCTION

The age of Big Data has only just begun, but the application of advanced analytics has been long established through years of mathematical research. With vast amounts of data from a variety of fields being generated each day, companies from a range of sectors are keen to exploit their data in order to gain a competitive edge. There is nothing new about data, data has been around for decades, what's new is the volume, velocity and variety of data. Big Data definitions are constantly evolving, however (Laney, 2001) definition of the 'three V's' has become a widely known framework. (Gartner, n.d.) equally describe Big Data as "*high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.*"

From the above definitions, the first dimension volume refers to the huge range of sources that are contributing to the flood of data being created, social networking sites such as Facebook and Instagram, transactional records from shopping with our credit cards, video rendering, mobile sensors, geographical exploration and even medical images which are generated as part of our 'Electronic Health Record' HER. The second dimension, velocity refers to the continually increasing rate in which data is being streamed at; video surveillance data requires real time analysis to effectively and efficiently react, whether it be for example, crime related or traffic congestion. The third dimension, variation refers to the distinct range of data types; structured, semi-structured, quasi-structured and unstructured. According to (Dietrich, 2015) data of a defined format and structure such as, online analytical processing (OLAP) data cubes or numerical data stored in conventional databases are examples of structured data. Text files with a clear pattern that can be defined by Extensible Mark-up Language (XML) are an example of semi-structured data. Text files with an unclear pattern that require a substantial amount of effort and time to format are an example of quasi-structured data. Emails, PDFs, pictures and videos are all examples of unstructured data, where the data is un-organized and has no apparent structure.

Big Data stretches far and wide, from social data to machine data to transactional data and even search data. For example, according to (Dietrich, 2015) social networking sites such as Twitter and Facebook are enabling companies to better understand their customer's behaviour, with 230 million tweets posted on Twitter and 2.7 billion likes and comments added to Facebook each day, the potential customer insight is endless. With the wealth of information now available online, (Google, 2012) search statistics have

reported that in excess of 40,000 web searches are generated every second, which constitutes to roughly 1.2 trillion searches across the globe each year.

The amount of data that is being produced and stored globally is constantly growing and according to (IBM, 2013), the two years prior to 2013 saw the generation of 90% of the world's data. However, with companies' ability to outsource their data storage to low cost storage locations such as Iceland (Mayor, 2015), the vast volume of data can be more easily stored and therefore much of today's focus is around dealing with the variety and velocity of data. According to (Economist, 2016) the last fifty years has seen computing power double every eighteen months and it is estimated to continually increase at this rate for at least another decade, this is known as 'Moore's Law'. The challenge of understanding this data has led to the development of new tools and techniques in the fields of Computing, Mathematics and Business, driving the recent expanse in Data Mining, Statistical Learning and Business Intelligence.

Big Data analytics encompasses a range of techniques that can be used to uncover hidden patterns, discover unknown correlations, highlight market trends and reveal customer insight from the data. The results can lead to more effective marketing, boost in revenue, improved customer service, increased operational efficiency and a competitive edge over rival companies. The primary focus of Big Data analytics is to provide companies and organizations with the necessary information to make more knowledgeable decisions, through analysing customer databases and records, Web server logs, Social Media content, text from customer emails, survey responses, mobile-phone call detail records and many more. This research, with the aid of SAS technologies, will explore the capabilities of Big Data Analytics when applied to an Insurance scenario.

The rest of this paper is organized as follows. [THE INSURANCE COMPANY DATA](#) section outlines the data used within this research. The [DATA PRE-PROCESSING STAGE](#) section focuses on understanding and reducing the complexity of the data. [THE PREDICTION STAGE](#) centres on generating prediction models from historical data and identified significant characteristics. The [CONCLUSION](#) provides a summary the results found in this research, highlights the issues to be addresses in future research and presents the conclusions.

THE INSURANCE COMPANY DATA

The Insurance Industry is inundated with huge amounts of data about their customers, their relations with customers and their interactions. Customer relations and interactions can be analysed using a variety of analytical techniques to boost revenue, reduce costs and improve customer service. Financial, actuarial, claims, risk, consumer and many other forms of data underpin practically every decision an insurer makes. Advanced analytics has become integral to providing solutions for insurers; aiding decisions through capturing and analysing structured and unstructured data associated with their policyholders. Hence there is huge magnitude of potential insight to gain from modelling past data and according to (Marr, 2015) the insurance (predictability) industry is keen to explore further.

Predicting the profiles of potential customers from a set of given information based on existing customers is a well-known data mining problem that is extremely important, especially in terms of direct marketing. Directly mailing a company's list of prospective customers can be an effective approach when it comes marketing a product or service. However, more often than not, this mail is regarded by many as 'junk mail' and thrown away by people who have no interest in this product or service and is therefore neither cost effective nor environmentally friendly. This research will consider a specifically chosen insurance scenario to highlight a more effective way to predict prospective customers.

The Insurance Company data was created especially for the CoIL Challenge 2000 Data Mining Competition by (Challenge, 2000), the data is representative of the characteristics featured in a typical Big Data dataset. The challenge attracted a large variety of solutions, both in terms of approaches and performance. Essentially the goal of the competition was to predict prospective customers who would be interested in purchasing a specific insurance product, a caravan policy and why. The CoIL Challenge 2000 is a prediction problem designed to represent properties that often appear in real world problems. The data is noisy, skewed, correlated and is of high dimension with a weak relationship between input and target.

From a marketing perspective the goal of the prediction task is to rank the prospective customers of the Insurance Company according to probability that they will buy a caravan policy, so that the highest ranked customers can be contacted via mailing. A Marketing Analyst would determine an optimal number of customers in which to mail where the cost of mailing and profit of response are known. For the purpose of the competition a predetermined subset of 20% of the most likely customers should be targeted for mailing, a select 800 out of the 4000 instances in the test set.

The data as mentioned is of high dimension, a total of 5822 training instances and 4000 test instances. The data entails 86 variables of which, 83 are numeric, 2 are symbolic input features and the 86th binary target variable caravan policy. As the data is highly noisy the key features to explain policy insurance are not present. The input can be divided into two categories, the first 43 features are based on sociodemographic data and the following 42 features are based on product ownership data. Each continuous variable has been discretized into ranges, more information can be found online (Putter, 2000).

The sociodemographic data is linked to the postal code of the customer rather than to the individual customer, which adds further to the complexity of this data. For example, a value of 4 for the variable 'Household with children' means that an estimated 37- 49% of people living in the same postal code area have children that live at home. Due to these features all being linked to the hidden variable location, some of the features are as a result highly correlated.

A range of statistical techniques could be used to approach this problem, this research is split into two main sections; Data Pre-processing and the Prediction task. The Data Pre-processing stage will focus on reducing the amount of information present and selecting the most relevant attributes to improve the prediction performance of the predictors and the interpretability and generalization of the data. Principal Component Analysis has been identified to perform the pre-processing task of customer database attributes selection. In addition, Logistic Regression has been identified to improve the prediction of prospective customers based on the attributes of existing customers. These techniques appear to be an important tool in increasing the effectiveness of direct marketing a service or product.

DATA PRE-PROCESSING STAGE

Linear dimensionality reduction methods such as Principal Component Analysis (PCA) have been widely used across domains such as engineering, astronomy, biology, economics, and many more (Fodor, 2002). PCA provides a means to systematically reduce the dimensionality of the data. According to (Darbyshire, 2016), PCA has been successfully applied in finance to the risk management of interest rate derivatives portfolios; reducing trading multiple swap instruments from a function of 30-500 to just 3 or 4 principle components. PCA produces a smaller more coherent set of variables whose 'principle components' are a linear combination of the original variables. According to (Guyon, 2003), PCA is a statistical technique that linearly transforms the original correlated variables into a smaller subset of uncorrelated variables. This is considered influential, as a small and uncorrelated variable is considerably easier to interpret and far more useful for further analysis in comparison to a large set of correlated variables.

PRINCIPLE COMPONENT ANALYSIS FOR INSURANCE COMPANY DATA

A PCA will be performed in SAS® Enterprise Guide® to determine key features present amongst the data. Prior to performing the initial PCA, the Insurance Company data was pre-examined for redundancies, this is important as these should be dealt with prior to the analysis. By means of a basic summary statistics carried out in SAS® Enterprise Guide® no redundancies were found and therefore no further investigation was required, hence it was possible to proceed with the initial PCA.

The initial PCA carried out used the SAS® Enterprise Guide® default setting in order to determine the number of factors to retain. According to (Jolliffe, 2002) by Kaiser's rule (1960), components should only be included in the analysis if their eigenvalue exceeds unity. Thus, the principal component must account for at least as much variation as one of the original variables used in the analysis. The Eigenvalues of the

Correlation Matrix revealed a total of 32 factors which all had an Eigenvalues greater than 1, therefore by Kaiser's rule this would suggest keeping a total of 32 factors. However, based on the information displayed in Table 1. Eigenvalues of the Correlation Matrixit can be seen that there is also 1 factor whose eigenvalue is close to 1, which is important to examine further.

Eigenvalues of the correlation Matrix: Total = 86 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	9.36719380	4.45530092	0.1089	0.1089
2	4.91189288	0.92261884	0.0571	0.1660
3	3.98927406	0.58374223	0.0464	0.2124
:	:	:	:	:
31	1.13906848	0.07686012	0.0132	0.8115
32	1.06220837	0.07812146	0.0124	0.8239
33	0.98408691	0.03641980	0.0114	0.8353

Table 1. Eigenvalues of the Correlation Matrix

A satisfactory degree of the proportion of variance is met by retaining 32 factors, this comfortably exceeds the required 70% variation from the original variables. By retaining at least 70% of the original data it ensures that a sufficient proportion of variation has been accounted for in the new factors. By examining the scree plot in Figure 1. Scree Plot for Insurance Company Datafor any 'elbows', two can be identified. The most extreme elbow illustrates that there is potential to only include the 6 factors above this point for further analysis. However, by retaining only 6 factors, insufficient variation would be retained and too much of the original data would be compromised. Thus, the results of the scree plot are inconclusive, the number of factors to retain is not clear.

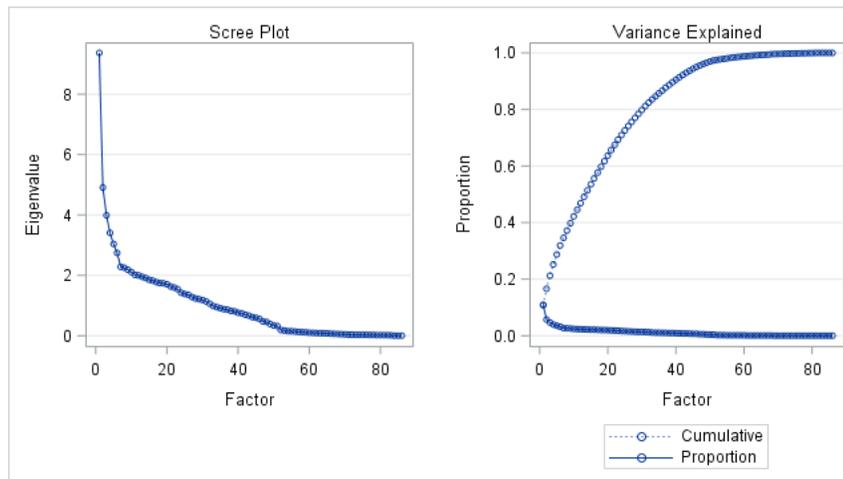


Figure 1. Scree Plot for Insurance Company Data

As the results for determining the number of factors were not clear, the factor loadings matrix will be investigated to gain further insight. A factor loading represents the extent in which a factor explains a variable. Factor Loadings are given between -1 to 1, the closer a loading is to -1 or 1 the stronger the affect that factor has on the variable. The closer the factor loading is to 0 the weaker the affect that factor has on the variable. In general, a variable is considered to be highly loaded on a factor if the value of the loading exceeds -0.6/0.6. As the SAS® Enterprise Guide®'s recommendation was to retain 32 factors

from the original 86 variables, the outputted factor pattern Matrix provides the loadings on each of these 32 suggested new factors. By examining the strength of the factor loadings on each of the new factors against each of the new variables it was possible to remove any of the new factors that were redundant (i.e. not significantly loaded).

From this analysis it was discovered that factors 21 to 32 were redundant as none of the original variables loaded highly on them and could not be incorporated into the model. Removing these Factors leaves 20 factors for further analysis and thus PCA is re-run with the number of factors to retain specified to 20. The factor pattern matrix was investigated in order to see which out of the 86 original variables were loading highly on the 20 newly generated factors. It was concluded that 41 of the original variables did not load on any of the factors and hence could not be incorporated into the new model.

Again a PCA was performed, including the specified 45 original variables which had loadings that were significant and the number of factors to retain was set to 20. The factor pattern matrix revealed that the number of factors could be reduced to 15 and an additional PCA was again run on the selected 45 original variables, this time set to 15 specified factors.

By retaining 15 factors the required degree of variation was met. However, the factor pattern matrix presented a number of cross loadings (variables loading on more than one factor) and the majority of the original variables were loading on the first factor. Due to a number of variables being highly loaded on more than one factor and multiple cross loadings present, a rotation will need to be considered. The chosen method used for this analysis is an Orthogonal Varimax. This rotation is suitable as it simplifies the principle components by maximizing the variance of the loading within the principle components and across variables. The rotated factor pattern is shown below.

		Rotated Factor Pattern														
		Status	Home Environment	Farming	Fire & Private	Family Accident	Trailer	Disability	Bicycle	Customer type	Accident	Property	Surfing	Income	Home Owner	Econo Deprived
F1	1 MOSTYPE Customer Subtype see L0	-0.36972	-0.03192	0.07258	-0.01398	-0.00882	0.01358	-0.00109	0.00034	0.8962	0.01341	-0.00683	0.01088	-0.06754	0.01329	0.09544
F3	3 MQEMOMV Avg size household 1 ♦ 6	-0.02666	0.80044	0.04598	-0.01148	-0.00301	-0.00186	0.02136	0.03163	-0.05638	-0.00125	0.02181	0.009	0.18029	-0.12012	0.14625
F5	5 MOSHOOFD Customer main type see L2	-0.36362	0.0086	0.08513	-0.01763	-0.00762	0.01594	-0.00157	-0.00281	0.838	0.01518	-0.00692	0.01052	-0.0648	-0.00601	0.09335
F10	10 MRELGE Married	0.14353	0.83773	-0.00962	-0.00955	0.00764	0.01521	-0.00099	0.0004	0.05476	0.01271	-0.02108	-0.00513	-0.01966	-0.10347	-0.23765
F12	12 MRELOV Other relation	-0.09872	-0.826	0.01184	0.01364	-0.0092	-0.01029	0.00847	0.01467	-0.01883	-0.00684	0.0159	0.00997	-0.04214	0.06517	0.32676
F13	13 MFALLEEN Singles	0.01069	-0.85549	-0.01954	0.01388	-0.01669	-0.00472	0.00105	0.00444	-0.0097	0.00308	0.00049	-0.00428	-0.17309	0.07298	0.07859
F15	15 MFWEKIND Household with children	-0.08068	0.80308	0.02157	-0.00925	-0.00179	-0.01833	0.01169	0.00978	-0.03159	0.0006	0.01678	0.00987	0.11619	-0.04134	0.2373
F16	16 MOPLHOOG High level education	0.69776	-0.04566	-0.06147	0.02182	-0.00001	-0.02869	-0.00371	0.02341	-0.25713	-0.01412	0.00933	0.006	0.27925	0.04548	-0.10412
F18	18 MOPLLAAG Lower level education	-0.62127	0.04561	0.04259	-0.00689	0.00042	0.01616	-0.00111	-0.01681	0.34304	0.01123	-0.01678	0.0071	-0.28805	0.00085	0.31157
F19	19 MBERHOOG High status	0.78025	0.04183	-0.01291	0.0267	-0.00354	-0.00443	0.00213	0.01651	-0.13683	0.00233	0.00911	-0.01429	0.0404	-0.12154	-0.09914
F24	24 MBERARBO Unskilled labourers	-0.33526	-0.15819	0.00212	-0.01414	-0.00238	0.01824	-0.00728	-0.02612	0.12095	-0.01292	-0.01599	-0.00325	0.00248	0.16982	0.6417
F25	25 MSKA Social class A	0.78187	0.04436	0.01295	0.00088	-0.00685	-0.00994	-0.00018	0.00884	-0.11381	0.00639	0.01373	0.00279	0.20294	-0.04213	-0.1204
F28	28 MSKC Social class C	-0.71821	0.07265	-0.02699	-0.00911	0.00319	-0.00605	-0.0227	-0.00646	0.20292	0.00026	0.00535	-0.00107	-0.12234	0.13812	-0.12697
F29	29 MSKD Social class D	-0.12899	-0.1537	-0.03169	-0.00266	-0.00688	-0.00656	-0.02219	-0.00588	0.08181	0.01243	-0.00123	-0.15056	-0.2567	0.05567	0.17367
F30	30 MHHUUR Rented house	-0.26904	-0.2827	-0.03115	-0.05228	-0.00719	-0.01221	-0.01627	-0.01322	0.00164	0.00456	-0.00142	-0.01016	-0.1865	0.87278	0.13005
F31	31 MHKOOP Home owners	0.26899	0.28296	0.03128	0.05224	0.00894	0.01202	0.01602	0.01268	-0.00228	-0.00476	0.00111	0.01003	0.18438	-0.873	-0.13161
F34	34 MAUTO No car	-0.20682	-0.62075	-0.02006	-0.01962	-0.02195	-0.00719	0.00805	-0.0015	-0.00189	0.00974	-0.02209	-0.0145	-0.08281	0.11129	0.41933
F35	35 MZFONDS National Health Service	-0.88664	-0.13323	-0.02404	-0.01436	-0.00184	-0.01357	0.00793	0.00276	-0.0402	-0.01977	0.00381	-0.02768	0.0024	0.09773	0.08487
F36	36 MZPART Private health insurance	0.88678	0.13493	0.02507	0.01369	0.00257	0.01408	-0.00822	-0.00243	0.04	0.01168	-0.00279	0.02742	-0.00364	-0.09654	-0.08761
F37	37 MINKM30 Income < 30.000	-0.19653	-0.40543	-0.00131	-0.01647	-0.00248	-0.01349	0.01055	-0.00349	0.03482	-0.00678	0.00104	0.00472	-0.66083	0.25239	0.14612
F39	39 MINK4575 Income 45-75.000	0.30866	0.18374	0.0195	-0.00074	0.02713	-0.01346	-0.01779	0.00477	-0.05315	-0.00368	-0.01192	0.01325	0.73412	-0.08066	-0.0842
F42	42 MINKGEM Average income	0.37279	0.28791	-0.00402	0.02371	0.00843	0.0214	-0.01482	0.00869	-0.0877	-0.01371	-0.00146	0.00477	0.64485	-0.13958	-0.12985
F44	44 PWAPART Contribution private third party insurance see L4	0.03434	-0.02722	-0.16052	0.8915	0.03048	-0.02998	-0.00668	0.01317	0.0133	-0.00867	0.02157	0.01097	0.03402	0.09546	-0.00755
F46	46 PWALAND Contribution third party insuranc (agriculture)	-0.03117	0.00469	0.88773	0.00887	-0.00111	0.03717	0.0012	-0.01012	-0.02409	0.01561	0.0024	0.00027	-0.007	-0.04938	-0.01072
F51	51 PAANHANG Contribution trailer policies	-0.00211	0.00512	0.06297	0.00448	-0.00136	0.99799	0.02483	-0.00196	0.01398	0.00552	0.00821	0.02929	0.00118	-0.01024	0.00583
F52	52 PTRACTOR Contribution tractor policies	0.02979	0.02687	0.85904	0.02269	0.01505	0.01698	0.05261	0.00075	0.08574	0.02896	-0.00456	-0.00251	0.00627	0.01985	0.00475
F56	56 PPERSONS Contribution private accident insurance policies	0.0052	0.00618	0.04333	-0.00537	-0.0026	0.00091	-0.0032	-0.00467	0.0133	0.97279	-0.00227	-0.00057	-0.00907	0.00037	0.00431
F57	57 PGEZONG Contribution family accidents insurance policies	0.00043	0.01656	0.01443	0.03682	0.99369	0.00084	-0.00291	0.00877	-0.00715	-0.00295	0.00471	-0.00103	0.0166	-0.00374	-0.00537
F58	58 PWAOREG Contribution disability insurance policies	0.00456	0.00738	0.04873	0.01174	-0.00302	0.03173	0.98445	-0.00504	0.0033	-0.0035	-0.0032	-0.00136	-0.01493	-0.01105	-0.01072
F59	59 PBRAND Contribution fire policies	0.02229	0.04282	0.27449	0.79968	0.0203	0.04374	0.04558	-0.02868	-0.02174	0.00762	0.01398	0.00131	-0.003	-0.19058	-0.1604
F60	60 PZELPL Contribution surfboard policies	0.0145	0.0087	-0.00222	0.01076	-0.00091	0.00248	-0.00067	-0.00166	0.0107	-0.00076	0.00666	0.96678	-0.00981	-0.00676	-0.00745
F62	62 PFIETS Contribution bicycle policies	0.019	0.0099	-0.00795	-0.02087	0.00043	-0.00221	-0.00511	0.98294	-0.00142	-0.00582	0.00481	-0.00211	0.00423	-0.00881	-0.01235
F63	63 PINBOED Contribution property insurance policies	0.01765	0.00177	-0.00559	0.02298	0.00584	0.00815	-0.00224	-0.00096	-0.00079	-0.00243	0.96706	0.00283	0.00062	0.00017	-0.00843
F65	65 AWAPART Number of private third party insurance 1 - 12	0.02735	-0.02891	-0.16085	0.89479	0.02805	-0.03037	-0.01389	0.00996	0.01753	-0.01084	0.01746	0.01	0.01312	0.00045	0.00028
F67	67 AWALAND Number of third party insurance (agriculture)	-0.03166	0.0062	0.88844	0.00959	-0.00122	0.04345	-0.00753	-0.00949	-0.0205	0.00777	0.00069	-0.00056	-0.00677	-0.0492	-0.01204
F72	72 AANHANG Number of trailer policies	-0.00702	0.00584	0.04678	0.00435	0.005	0.98895	0.02648	-0.0038	0.01138	0.00317	0.01879	0.02023	0.0058	-0.00795	0.00011
F73	73 ATRACTOR Number of tractor policies	0.0346	0.02657	0.84603	0.02328	0.01636	0.01136	0.05024	0.00174	0.08677	0.02383	-0.0049	-0.00238	0.01896	0.02943	-0.00592
F77	77 APERSONS Number of private accident insurance policies	0.00752	0.00021	0.03211	0.00208	-0.00336	0.00758	-0.00327	-0.006	0.00962	0.97335	-0.00337	-0.00099	0.00084	0.0056	-0.00099
F78	78 AGEZONG Number of family accidents insurance policies	-0.00228	0.01682	0.01578	0.03676	0.99362	0.00281	-0.00343	0.0151	-0.0055	-0.00314	0.00843	-0.00124	0.01045	-0.00803	-0.01604
F79	79 AWAOREG Number of disability insurance policies	-0.00161	0.00677	0.04928	0.01619	-0.00329	0.01926	0.9845	-0.00457	-0.00577	-0.00309	-0.00275	-0.00085	-0.01227	-0.01203	-0.01167
F80	80 ABRAND Number of fire policies	-0.0043	-0.01712	0.16784	0.82859	0.00255	0.02969	0.01079	-0.03076	-0.0394	0.00889	0.00858	-0.0045	-0.03506	-0.11627	0.00207
F81	81 AZELPL Number of surfboard policies	0.0205	0.00523	-0.00371	0.00493	-0.0013	0.04074	-0.00149	-0.00233	0.00578	-0.00079	0.00311	0.96511	0.0001	0.00707	-0.00904
F83	83 AFIETS Number of bicycle policies	0.02401	0.01092	-0.00922	-0.01063	0.02315	-0.00349	-0.00443	0.98254	-0.00309	-0.00494	-0.00386	-0.00193	0.00846	-0.0095	-0.01178
F84	84 ANBOED Number of property insurance policies	0.0038	0.01194	-0.00047	0.0269	0.00695	0.01854	-0.0036	0.00536	-0.01135	-0.00318	0.9657	0.04201	0.00952	-0.00235	-0.00721

Table 2. Rotated Factor Pattern for Insurance Company Data

Table 2. Rotated Factor Pattern for Insurance Company Data allows the reader to visualise the magnitude of the data being investigated. However, to investigate this further a subset of the Rotated Factor Pattern will be considered.

	Status	Home Environment
Customer Subtype	-0.36972	-0.03192
Avg Size Household	-0.02666	0.80044
Customer Main Type	-0.36362	0.0086
Married	0.14353	0.83773
Other Relation	-0.09872	-0.826

Table 3. Subset of Rotated Factor Pattern for Insurance Company Data

Table 3. Subset of Rotated Factor Pattern for Insurance Company Data reveals that the original variable 'Average Size Household' is loaded highly with a value of 0.8004 on the factor 'Home Environment'. In summary, this table illustrates which of the original variables are highly loaded on which of the newly generated factors. For example, Factor 1 'Status' has in total 7 associated highly loaded original variables. No variable is loaded on more than one factor and each factors has at least one highly loaded variables associated.

The factors are renamed based on the corresponding loading variables associated. The Names of newly generated components for the Insurance Company Data are as follows:

- Factor 1 - This factor loads on variables relating to education and social class and is therefore renamed as **Status**.
- Factor 2 - This factor loads on variables relating to house size, marital status, house with children and house with no children and is therefore renamed **Home Environment**.
- Factor 3 - This factor loads on variables relating to tractors and agricultural policies and is therefore renamed **Farming policies**.
- Factor 4 - This factor loads on variables relating to fire policies and private third party policies and is therefore renamed **Private and Fire policies**.
- Factor 5 - This factor loads on variables relating to family accident policies and is therefore renamed **Family Accident Policies**.
- Factor 6 - This factors loads on the two variables relating to trailer policies and is therefore renamed **Trailer Policies**.
- Factor 7 - This factor loads on two variables relating to disability insurance and is therefore renamed **Disability Cover**.
- Factor 8 - This factor loads on the two variables relating to bicycle policies and is therefore renamed **Bicycle Policies**.
- Factor 9 - This factor loads on variables relating to the type of customer and is therefore renamed **Customer Type**.
- Factor 10 - This factor loads on the two variables relating to accident policies and is therefore renamed **Accident Policies**.
- Factor 11 - This factor loads on the two variables relating to property insurance and is therefore renamed **Property Cover**.
- Factor 12 - This factor loads on the two variables relating to surfing polices and is therefore renamed **Surfing Policies**.
- Factor 13 - This factor loads on variables that relating to income ranges and average income and is therefore renamed **Income**.
- Factor 14 - This factor loads on variables relating to home owners and rented homes and is therefore renamed **House Owner/Rented**.

- Factor 15 - This factor loads on variables relating to low social class and includes unskilled labours and is therefore renamed as **Economically Deprived**.

In summary, it has been found 15 newly generated factors which essentially explain the key characteristics of a potential customer's interest in a caravan policy. For example, it is not uncommon for people who own caravans to also own bicycles and surfboards. It is also highly likely that customers who own a caravan would have purchased a fire and accident policy to cover the gas fires on board their caravan. By successfully reducing the dimensionality of the data, a concise subset of factors have been generated which as a result are easier to interpret for further analysis.

The results generated by PCA in this case match similarly to the results found by (Ramavajjala, 2012) however the strongest indicating feature, 'Number of Boat Policies' found by (Ramavajjala, 2012) was not incorporated into any of the factors generated by the PCA. 'Number of Boat Policies' along with a few others original variables will be investigated in more detail for [THE PREDICTION STAGE](#).

Boxplots were produced between the original variables and the newly formed factors. These highlighted a key advantage to PCA; multiple relationships could be analysed on a single graph. A number of interesting relationships were found, in particular Figure 2 illustrates the relationship between the original variable 'Average House Hold Size' and the newly formed factor 'Home Environment'.

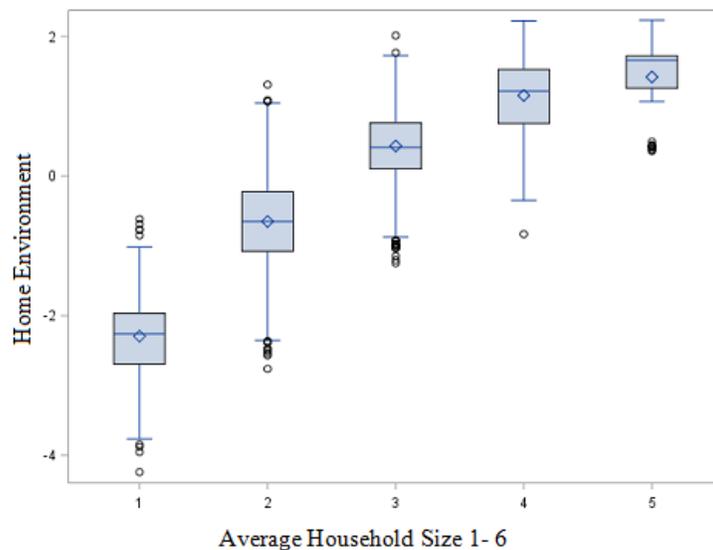


Figure 2. Box Plot Home Environment against Average Household Size 1-6

This suggests that the 'Average Household Size' and the 'Home Environment' scores increase simultaneously. This seems plausible, as you would expect a large 5-bedroom house to have a greater home environment score than a smaller 2-bedroom house.

Additional statistical testing was interpreted. A Spearman's correlation coefficient was used to interpret the relationship between the variable 'Average Size Household' and the factor 'Home Environment'.

The hypotheses in this case are:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

The test carried out in SAS® Enterprise Guide® revealed a p -value of <0.0001 . The p -value is significant at the 1% level of significance. Therefore, there is sufficient evidence to reject the null hypothesis and conclude that there is a significant relationship between the factor 'Home Environment' and the original variable 'Average Household Size'.

Spearman Correlation Coefficients, N = 5822 Prob > r under $H_0: \rho = 0$	
	Average Household Size 1-6
Home Environment	0.62631 < .0001

Table 4. Spearman Correlation for Home Environment with Average Household Size 1-6

The strength of the relationship is determined by the correlation coefficient, which here is 0.62631, indicating a moderate to strong positive relationship.

In summary, the PCA has proven to be an efficient technique to deal with Big Data for Insurance Companies. PCA has been used successfully to reduce the dimensionality of the data to a subset of more coherent set of principle components, in a way that has minimized the volume of data to be managed by the Insurer with no data compromise. The supplementary statistical tests highlighted a key advantage of PCA, successfully identifying a range of underlying relationships and correlations between the variables amongst existing customers. This also provides the ability to compare multiple relationships on a single graph. The next section of this paper [THE PREDICTION STAGE](#), will focus on identifying prospective caravan policy customers by means of Logistic Regression.

THE PREDICTION STAGE

LOGISTIC REGRESSION FOR INSURANCE COMPANY DATA

Logistic regression is a widely used statistical technique. According to (Hardin, 2017) Logistic Regression is applied to data where there is a binary (success-failure) outcome or response variable, or in some cases, where the result follows the structure of a binomial proportion. An estimate of the relationship between a predictor variable and an outcome variable is computed, similar to multiple linear regression. However according to (King, 2008), Logistic Regression takes into consideration the fact that the dependent variable is categorical. This paper will focus on modeling caravan policy ownership applying a Binary Logistic Regression using Base SAS® 9.4. The outcome variable will be parameterized in terms of the logit of caravan ownership = 1 versus no caravan ownership = 0.

MODEL 1 CODE – INCLUDING ALL FACTORS GENERATED BY PCA

The initial Logistic Regression model focuses on predicting caravan policy ownership as an outcome with the 15 newly generated factors from the PCA, performed in the Data Pre-processing section.

The hypothesis for the binary logistic regression is as follows:

$$H_0 : \beta_1 = \dots = \beta_k = 0$$

$$H_A : \text{Not all } \beta_i = 0$$

The code for the logistic regression is as follows:

```
proc logistic data=MS4S05.FACTOR_SCORES_CARAVAN;
model CARAVAN_OWNERSHIP (event="1")= STATUS HOME_ENVIRONMENT FARMING_POLICIES
FIRE_PRIVARE_POLICIES FAMILY_ACCIDNET_POLICIES TRAILER_POLICIES
DISABILITY_COVER BICYCLE_POLICIES CUSTOMER_TYPE ACCIDENT_POLICIES
PROPERTY_COVER SURFING_POLICIES INCOME HOME_OWNER_RENTED
ECONOMICALLY_DEPRIVED /clodds=p1;
run;
quit;
```

The `model` statement is similar to that of a `PROC REG`. To the left of the equals sign, is the identify outcome variable `CARAVAN_OWNERSHIP` and to the right of the equals sign is the 15 factors used as the predictor variables in this model. The `(event="1")` option on the model statements specifies that the probability of purchasing a caravan ownership policy will be predicted. The `clodds=pl` option is used to indicate to SAS that the Confidence Limits on the odds ratio using the method of profile likelihood are required. The `oddsratio` statement is used to calculate the odds ration as this is not a default when interactions are present.

MODEL 1 OUTPUT – INCLUDING ALL FACTORS GENERATED BY PCA

Response Profile		
Ordered Value	CARAVAN_OWNERSHIP	Total Frequency
1	0	5474
2	1	348

Table 5. Logistic Procedure: Response Profile

Probability modelled is
CARAVAN_OWNERSHIP = '1'.

Firstly, the output from Table 5 shows that all 5822 observations in the data set have been included in the analysis. It can also be seen that SAS is modelling **CARAVAN_OWNERSHIP** using a binary logit model with the probability that of **CARAVAN_OWNERSHIP = 1**.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	146.7832	15	<.0001
Score	152.8221	15	<.0001
Wald	141.1014	15	<.0001

Table 6. Logistic Procedure: Testing Global Null Hypothesis

Table 6. Logistic Procedure: Testing Global Null Hypothesis6 gives the likelihood ratio chi-square test statistic equal to 146.7832 with a p -value of <.0001 so there is significant evidence to reject the null hypothesis. Hence, Caravan Ownership can be modelled by at least one of the factors. It can be seen by Table 6 that the model as a whole is a significantly better fit than an empty model. The Score 152.8221 and Wald tests 141.1014 are equivalent tests of the same hypothesis tested by the likelihood ratio test and thus also indicate that the model is statistically significant.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9527	0.0651	2055.0425	<.0001
STATUS	1	0.1987	0.0504	15.5108	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
HOME_ENVIRONMENT	1	0.2206	0.0611	13.0282	0.0003
FARMING	1	-0.1281	0.0814	2.4742	0.1157
FIRE_PRIVATE_POLICIES	1	0.4050	0.0563	51.8126	<.0001
FAMILY_ACCIDENT_POLICIES	1	-0.1925	0.0537	12.8676	0.0003
TRAILER_POLICIES	1	0.0726	0.0360	4.0659	0.0438
DISABILITY_POLICIES	1	0.0405	0.0458	0.7821	0.3765
BICYCLE_POLICIES	1	0.0649	0.0365	3.1607	0.0754
CUSTOMER_TYPE	1	0.0931	0.0429	4.7060	0.0301
ACCIDENT_POLICES	1	-0.0510	0.0830	0.3773	0.5390
PROPERTY_COVER	1	0.1351	0.0581	5.4077	0.0200
SURFING_POLICIES	1	0.0249	0.0456	0.2981	0.5851
INCOME	1	0.0320	0.0295	1.1752	0.2783
HOUSE_OWNER_RENTED	1	0.2158	0.0569	14.4075	0.0001
ECONOMICALLY_DEPRIVED	1	-0.2360	0.0645	13.4051	0.0003

Table7. Logistic Procedure: Analysis of Maximum Likelihood Estimates

The output of Table 7 shows the hypothesis tests for each of the variables in the model individually. The chi-square test statistics and associated p -values shown in the table indicate the variables in the model which significantly improve the model fit. Here at the 1% level of significance STATUS, HOME_ENVIRONMENT, FIRE_PRIVATE_POLICIES, FAMILY_ACCIDENT_POLICIES, CUSTOMER_TYPE, PROPERTY_COVER HOUSE_OWNER_RENTED and ECONOMICALLY_DEPRIVED are statistically significant.

The log odds of purchasing a caravan ownership policy is equal to $-2.9527 + 0.1987 \times \text{STATUS}$. Hence, for every increase in 1 in Status, this increases the likelihood of taking out a caravan ownership policy by 19.87%.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	68.2	Somers' D	0.365
Percent Discordant	31.8	Gamma	0.365
Percent Tied	0.0	Tau-a	0.041
Pairs	1904952	c	0.682

Table 8. Logistic Procedure: Association of Predicted Probabilities and Observed Responses

Examining Table 8. Logistic Procedure: Association of Predicted Probabilities and Observed Responses it can be deduced how well the model is doing. The values of the left of the table essentially

identify how well the model is doing. The pairs are calculated by taking all possible pairs of customers in which one has the outcome in question (owns a caravan policy) and the other does not. There were 348 who had a caravan policy and 5474 who did not, the total number of pairs is therefore $348 \times 5474 = 1904952$. The values to the right of the table are all measures of rank correlation. The higher the value of these measures of association the better the predictive value. Here the statistical probability c is 0.682, this means that a customer who has purchased a caravan policy has a higher predicted probability of purchasing a caravan policy than a customer who has not purchased a caravan policy.

The odds ratios estimates are displayed in Table 9 along with the plot of the Confidence Limits in Figure 3.

Odds Ratio Estimates and Wald Confidence Intervals			
Odds Ratio	Estimate	95% Confidence Limits	
STATUS	1.220	1.105	1.347
HOME_ENVIRONMENT	1.247	1.106	1.405
FARMING	0.880	0.750	1.032
FIRE_PRIVATE_POLICIES	1.499	1.343	1.674
FAMILY_ACCIDENT_POLICIES	0.825	0.742	0.916
TRAILER_POLICIES	1.075	1.002	1.154
DISABILITY_POLICIES	1.041	0.952	1.139
BICYCLE_POLICIES	1.067	0.993	1.146
CUSTOMER_TYPE	1.098	1.009	1.194
ACCIDENT_POLICES	0.950	0.808	1.118
PROPERTY_COVER	1.145	1.021	1.283
SURFING_POLICIES	1.025	0.937	1.121
INCOME	1.033	0.974	1.094
HOUSE_OWNER_RENTED	1.241	1.110	1.387
ECONOMICALLY_DEPRIVED	0.790	0.696	0.896

Table 9. Logistic Procedure: Odds Ratio Estimates and Wald Confidence Intervals

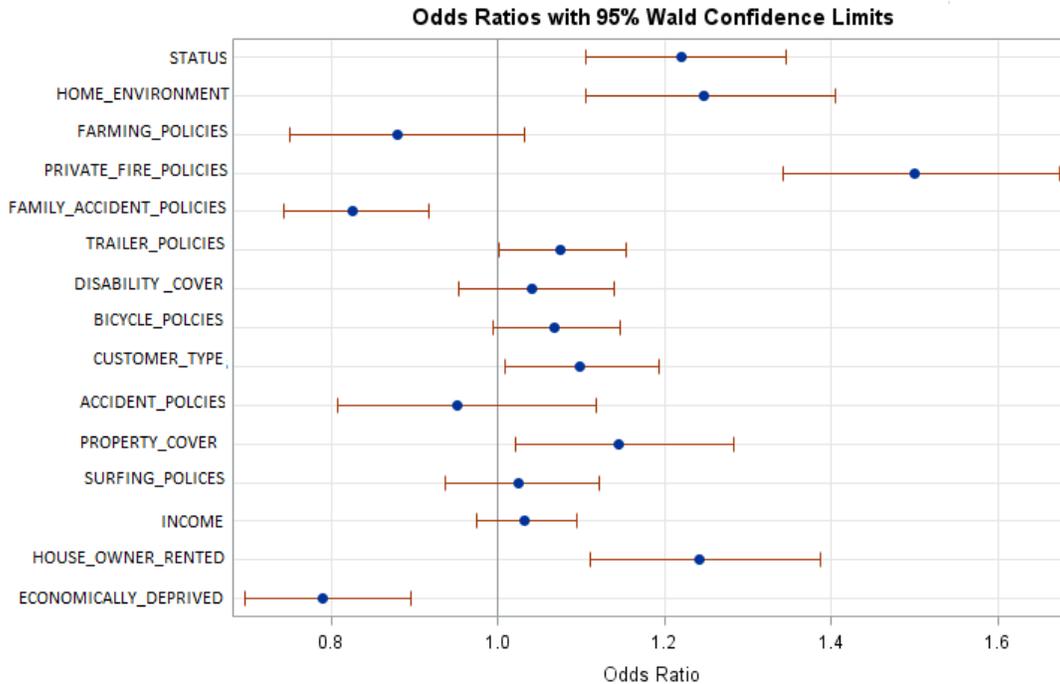


Figure 3. Logistic Procedure: Odds Ratio with 95% Wald Confidence Limits

The odds ratio values for each of the factors are stated under the estimates column in Table 9. This shows customers who hold a PRIVATE_FIRE_POLICIES are the most likely to purchase a caravan ownership policy. In fact, customers who hold a PRIVATE_FIRE_POLICIES are 1.499 times more likely to purchase a caravan ownership policy than customer who are ECONOMICALLY_DEPRIVED, who are only 0.790 times likely. Here, FARMING_POLICIES, TRAILER_POLICIES, DISABILITY_COVER, BICYCLE_POLICIE, ACCIDNET_POLICIES, SURFING_POLICIES and INCOME should be removed from the model as they include 1 in their confidence limits. As mentioned STATUS, HOME_ENVIRONMENT, FIRE_PRIVATE_POLICIES, FAMILY_ACCIDENT_POLICES, CUSTOMER_TYPE, PROPERTY_COVER HOUSE_OWNER_RENTED and ECONOMICALL_DEPRIVED were the only factors that were statistically significant. The logistic regression model will be re-run excluding the non-significant factors.

MODEL 2 CODE – INCLUDING ONLY SIGNIFICANT FACTORS

The second logistic regression model focuses on predicting caravan policy ownership as an outcome with only the significant factors from the initial model, STATUS, HOME_ENVIRONMENT, FIRE_PRIVATE_POLICIES, FAMILY_ACCIDENT_POLICES, CUSTOMER_TYPE, PROPERTY_COVER HOUSE_OWNER_RENTED and ECONOMICALL_DEPRIVED. The hypothesis for the second binary logistic regression model are the same as the previous.

The code for the logistic regression model 2 is as follows:

```
proc logistic data= MS4S05.FACTOR_SCORES_CARAVAN;
model CARAVAN_OWNERSHIP (event="1")= STATUS HOME_ENVIRONMENT
FIRE_PRIVATE_POLICIES FAMILY_ACCIDENT_POLICES CUSTOMER_TYPE PROPERTY_COVER
HOUSE_OWNER_RENTED ECONOMICALL_DEPRIVED /clodds=p1;
run;
quit;
```

The code explanation remains the same as the previous model.

MODEL 2 OUTPUT– INCLUDING ONLY SIGNIFICANT FACTORS

The output relating to the model information states the file being analysed and the number of observations included remains the same as that of the previous model. The model convergence and model fit statistics also remain the same as that of the previous model with p-values of <0.001.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9398	0.0645	2074.5806	<.0001
STATUS	1	0.1984	0.0504	15.5181	<.0001
HOME_ENVIRONMENT	1	0.2231	0.0611	13.3488	0.0003
FIRE_PRIVATE_POLICIES	1	0.4097	0.0562	53.1557	<.0001
FAMILY_ACCIDENT_POLICIES	1	-0.1922	0.0534	12.9552	0.0003
CUSTOMER_TYPE	1	0.0947	0.0430	4.8574	0.0275
PROPERTY_COVER	1	0.1367	0.0579	5.5818	0.0181
HOUSE_OWNER_RENTED	1	0.2160	0.0566	14.5780	0.0001
ECONOMICALLY_DEPRIVED	1	-0.2365	0.0643	13.5366	0.0002

Table 10. Analysis of Maximum Likelihood Estimates (Model 2)

Table 10. Analysis of Maximum Likelihood Estimates (Model 2), contains the Estimates chi-square test statistics and associated p -values suggests all factors are significant at the 5% level of significance. However, CUSTOMER_TYPE and PROPERTY_COVER are not significant at the 1% level of significance.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	67.6	Somers' D	0.352
Percent Discordant	32.4	Gamma	0.352
Percent Tied	0.0	Tau-a	0.040
Pairs	1904952	c	0.676

Table 5. Association of Predicted Probabilities and Observed Responses

Table 11 shows that the statistical probability c is 0.676, this value is only 0.06 less powerful at predicting caravan policy ownership than that of the previous model.

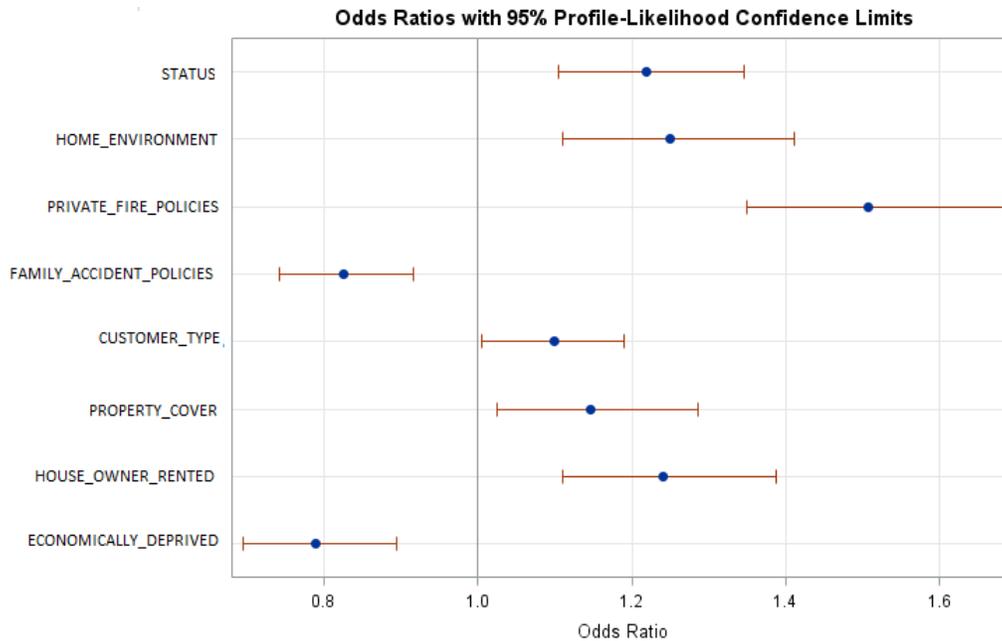


Figure 3. Odds Ratio with 95% Profile-Likelihood Confidence Limits

The confidence intervals for the factors illustrated in Figure 4 satisfied the requirements as they do not contain 1.

In summary, the second logistic model has been simplified considerably in comparison to the previous model. Despite reducing the number of factors included, the model was only 6% less powerful at predicting caravan policy ownership than that of the previous model. The next model will look to investigate the predictive abilities of a model which includes a combination of factors and original variables (Not accounted for in the PCA Factors).

MODEL 3 CODE – INCLUDING SIGNIFICANT FACTORS AND ORIGINAL VARIABLES

The final logistic regression model focuses on predicting caravan policy ownership as an outcome with the significant factors, STATUS, HOME_ENVIRONMENT, FIRE_PRIVATE_POLICIES, FAMILY_ACCIDENT_POLICES, CUSTOMER_TYPE, PROPERTY_COVER HOUSE_OWNER_RENTED and ECONOMICALL_DEPRIVED from the initial model. The model also includes the original variables from the (Challenge, 2000) data FARMER, NUMBER_MOTORCYCLCLE_POLICIES and NUMBER_BOAT_POLICIES, as these were not accounted for in the factors generated by the PCA but were found to be significant. The hypothesis for this binary logistic regression model are the same as the first model.

The code for the logistic regression model 2 is as follows:

```
proc logistic data=MS4S05.FACTOR_SCORES_CARAVAN plots(only)=(roc
oddsratio);
model CARAVAN_OWNERSHIP (event="1")= STATUS HOME_ENVIRONMENT
FIRE_PRIVATE_POLICIES FAMILY_ACCIDENT_POLICIES
FIRE_PRIVATE_POLICIES*FAMILY_ACCIDENT_POLICIES CUSTOMER_TYPE PROPERTY_COVER
HOUSE_OWNER_RENTED ECONOMICALLY_DEPRIVED FARMER
NUMBER_MOTORCYCLE_SCOOTER_POLICIES
FIRE_PRIVATE_POLICIES*NUMBER_MOTORCYCLE_SCOOTER_POLICIES NUMBER_BOAT_POLICIES
@2/ selection=backward slstay=0.05 clodds=pl;
```

```

oddsratio STATUS;
oddsratio HOME_ENVIRONMENT;
oddsratio FIRE_PRIVATE_POLICIES;
oddsratio FAMILY_ACCIDENT_POLICIES;
oddsratio CUSTOMER_TYPE;
oddsratio PROPERTY_COVER;
oddsratio HOUSE_OWNER_RENTED;
oddsratio ECONOMICALLY_DEPRIVED;
oddsratio FARMER;
oddsratio NUMBER_MOTORCYCLE_SCOOTER_POLICIES;
oddsratio NUMBER_BOAT_POLICIES;
run;
quit;

```

The basic structure of the logistic model code remains the same as the initial model. However, some additions have been made to compute a more detailed model. When performing a `proc logistic` you can specify effects and interactions. This model will include all possible combinations of interactions amongst the factors and original variables. The model also uses a `selection=backward`, essentially this is backwards elimination techniques used to build the model. This method stops main effects from being removed from the model if they are involved in an interaction. This method is referred to by (Mac Nally, 2000) as a hierarchical model. In addition the `plots(only)=(roc oddsratio)` option has been used to output a ROC curve showing the odds ratio. The `slstay=0.05` specifies the significance level used for removing effects.

MODEL 3 OUTPUT – INCLUDING SIGNIFICANT FACTORS AND ORIGINAL VARIABLES

The output relating to the model information that states the file being analysed and the number of observations included remains the same as that of the first model. The model convergence and model fit statistics also remain the same as that of the first model.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	70.6	Somers' D	0.411
Percent Discordant	29.4	Gamma	0.411
Percent Tied	0.0	Tau-a	0.046
Pairs	1904952	c	0.706

Table 12. Association of Predicted Probabilities and Observed Responses (Model 3)

Here, with the interactions all measures are improved, the statistical probability c is 0.706, this value is 0.024 more powerful at predicting caravan policy ownership than that of the first model. However, the improvements are small and arguably the simpler model could be chosen.

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
STATUS	1.0000	1.221	1.105	1.346
HOME_ENVIRONMENT	1.0000	1.245	1.106	1.406
CUSTOMER_TYPE	1.0000	1.103	1.008	1.195
PROPERTY	1.0000	1.165	1.038	1.308

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
HOUSE_OWNER_RENTED	1.0000	1.244	1.110	1.395
ECONOMICALLY_DEPRIVED	1.0000	0.807	0.711	0.912
FARMER	1.0000	0.770	0.657	0.887
NUMBER_BOAT_POLICIES	1.0000	9.619	4.514	20.195

Table 13. Odds Ratio Estimates and Profile-Likelihood Confidence Intervals (Model 3)

The confidence intervals displayed suggest that for the factors are satisfied as they do not contain 1.

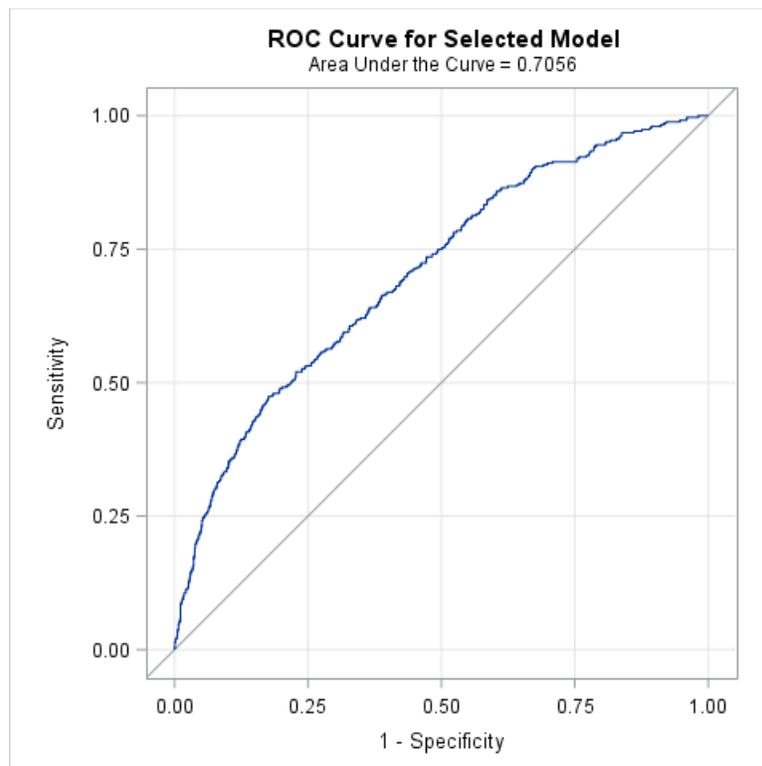


Figure 4. ROC Curve for Selected Model

Figure 4. ROC Curve for Selected Model illustrates the final Receiver Operating Characteristics (ROC) curve, this illustrates the relationship between a false-positive rate and the sensitivity of the test and are useful for determining the accuracy of prediction. A false-positive rate is the proportion of time an observation without the outcome is predicted by the model to have the outcome (it is also equal to one minus the specificity). Sensitivity is the proportion of observations that have the outcome that is predicted by the model to have the outcome. According to (Hanley, 1982) the greater the area under the curve (where the model has the highest sensitivity for a given false-positive) the better the model. Amongst the three models there is little difference between the ROC curves.

Summarizing the results of the final model, it was found that by including the statistically significant factors STATUS, HOME_ENVIRONMENT, FIRE_PRIVATE_POLICIES, FAMILY_ACCIDENT_POLICIES,

CUSTOMER_TYPE, PROPERTY_COVER HOUSE_OWNER_RENTED, ECONOMICALL_DEPRIVED and the original variables FARMER, NUMBER_MOTORCYCLCLE_POLICIES and NUMBER_BOAT_POLICIES improved the models predictability by 0.024. As mentioned earlier this was only a small increase and therefore there could be an argument for the simpler model to be chosen .From an insurance perspective a simpler model would be more practical as it would be easier to interpret and implement.

CONCLUSION

In conclusion, it has become increasingly clear that insurance companies must harness their information assets to gain critical insight and more in depth knowledge of markets, customers, products, competitors and employees. Insurance companies are still yet to discover the potential of capturing and analysing the rapidly increasing volume, velocity and variety of new and existing data. Use of appropriate techniques and SAS software can enable insurers to gain a better understanding of their operations, customers and new markets, providing insurance companies with a competitive edge necessary to thrive in this global, dynamic marketplace.

This paper has highlighted the benefits of performing a Principle Component Analysis at a pre-processing stage. The implementation of PCA proved a useful tool enabling insurers to collapse the dimensionality of their data. Furthermore, PCA also aided in assisting data analysis, for example, the output enabled multiple relationship to be explored on a single graph. The Binary Logistic Regression models demonstrated how insurers can identify potential customers by predicting their behaviour using past customer data. The model derived in this research forms the basis of identifying a selection of customers who would be interested in a particular policy and provides insurers with an efficient way to target prospective customers for direct mailing. It was found that the model outlined in [MODEL 2 – INCLUDING ONLY SIGNIFICANT FACTORS](#) provided the best predictive results based on its simplicity for insurance purposes. Hence, it can be concluded that for the (Challenge, 2000) data the factors STATUS, HOME_ENVIRONMENT, FIRE_PRIVATE_POLICIES, FAMILY_ACCIDENT_POLICES, CUSTOMER_TYPE, PROPERTY_COVER HOUSE_OWNER_RENTED and ECONOMICALL_DEPRIVED are most efficient at predicting caravan policy ownership.

This research has ultimately highlighted that the combination of applying Principle Component Analysis and Binary Logistic Regression provided a successful solution when dealing with big data sets and are therefore identified as useful tools for insurers. Further work will look to explore, review and compare alternative statistical and Big Data techniques for dimensionality reduction and predictive modelling. In particular, future work will look to compare the differences and similarities in applying a Partial Least Squares Regression (PLSR) as an alternative to PCA and an Artificial Neural Network (ANN) as an alternative to Logistic Regression to improve the reliability of the model. These techniques could be applied to other areas of the Insurance sector.

REFERENCES

- Challenge, C. (2000). *The Insurance Company Case*. Amsterdam: Sentient Machine Research. Also a Leiden Institute of Advanced Computer Science.
- Darbyshire, J. H. (2016). *The PRICING and TRADING of Interest Rate Derivatives*.
- Dietrich, D. (2015). *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. EMC Education Services.
- Economist. (2016, March 12). *Technology Quarterly*. Retrieved from After Moore's law: <http://www.economist.com/technology-quarterly/2016-03-12/after-moores-law>
- Elkan, C. (2001). Magical thinking in data mining: lessons from CoLL challenge 2000. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 426-431).
- Fodor, I. K. (2002). A survey of dimension reduction techniques. *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory 9*, 1-18.
- Gartner, I. (n.d.). *IT Glossary: Big Data*. Retrieved from <http://www.gartner.com/it-glossary/big-data/>

- Google. (2012). *internet live stats*. Retrieved from Google Search Statistics: <http://www.internetlivestats.com/google-search-statistics/>
- Guyon, I. a. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 1157-1182.
- Hanley, J. A. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. In *Radiology* 143.1 (pp. 29-36).
- Hardin, D. J. (2017). *Logistic Regression*. Retrieved from The Institute for Statistics Education : <http://www.statistics.com/logistic-regression/>
- IBM. (2013). *Bringing big data to the enterprise*. Retrieved from What is big data?: <https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>
- Jolliffe, I. (2002). *Principle Component Analysis*. John Wiley & Sons, Ltd.
- King, J. E. (2008). Binary logistic regression. In *Best practices in quantitative methods* (pp. 358-384). SAGE.
- Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. pp. META Group Research Note, 6, 70.
- Mac Nally, R. (2000). Regression and model-building in conservation biology, biogeography and ecology: the distinction between—and reconciliation of—'predictive' and 'explanatory' models. In *Biodiversity and Conservation* 9.5 (pp. 655-671).
- Marr, B. (2015, December 15). *How Big Data is Changing Insurance Forever*. Retrieved from Forbes: <http://www.forbes.com/sites/bernardmarr/2015/12/16/how-big-data-is-changing-the-insurance-industry-forever/#9e78bce435e8>
- Mayor, T. (2015, March 20). *Data centers in Iceland? Yes, really!* Retrieved from COMPUTERWORLD: <http://www.computerworld.com/article/2899654/data-centers-in-iceland-yes-really.html>
- Putter, P. V. (2000). *Insurance Company Benchmark (COIL, 2000) Data Set*. Retrieved from UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+%28COIL+2000%29>
- Ramavajjala, V. &. (2012). Policy iteration based on a learned transition model. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* , 211-226.
- Rouse, M. (2014). Big data. *Big Data and Cloud Business Intelligence*, Tech Target,30. Retrieved from Big Data and Cloud Business Intelligence.
- Smith, H. (2012, March 23). *Big Data FAQs*. Retrieved from ARC Community, arcplan, Inc: <https://community.arcplan.com/blogs/communityannouncement/big-data-faqs>
- TechAmerica. (2012). *Demystifying big data: A practical guide to transforming the business of Government*. Retrieved from <http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-bigdatareport-final.pdf>

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Rebecca Peters
 Faculty of Computing, Engineering and Science, University of South Wales
 Treforest, Pontypridd CF37 1DL
rebecca.peters@southwales.ac.uk
 @Rebeccapeters94

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.