UNIST UNDERGRADUATE SCHOOL

RESEARCH REPORT

Student:	20151640 (Student ID)	YoungSeok (First Name)	Song (Last Name)
Track:	IE (1 Track)	CSE (2 Track)	Grade: 4th
Advisor:	IE (School)	YongJae (First Name)	Lee (Last Name)

Starting Date: 3/1/2020 (mm/dd/yyyy) Ending Date: 06/18/2020 (mm/dd/yyyy)

Research Topic: 자산관리 FinTech 핵심 기술: 개인의 생애주기 자산관리를 위한 대형

---

1. Research purpose

AI-based life cycle medical cost management is the main topic. During the first semester, we focused on collecting data on a model that predicts how much treatment will cost depending on the disease.

2.Research contents

Data collection

To understand insurance, I studied insurance in National Health Insurance and researched information about the cost of treatment for diseases. I came to know that there is a concept of a major disease(주상병) when paying the cost of treatment for a disease, and it is difficult to know exactly how much it costs when a single disease occurs due to medical treatment and various drug prices. Through research, we have learned that paying for treatment is not a simple matter, but there are several complex items.

Through the public data portal(공공 데이터 포털), we reviewed various data in the health care sector and collected data helpful in predicting disease treatment costs. Among them, we found materials classified as 3-tier disease Code and 4-tier disease Code (4th tier is more detailed than 3rd tier).

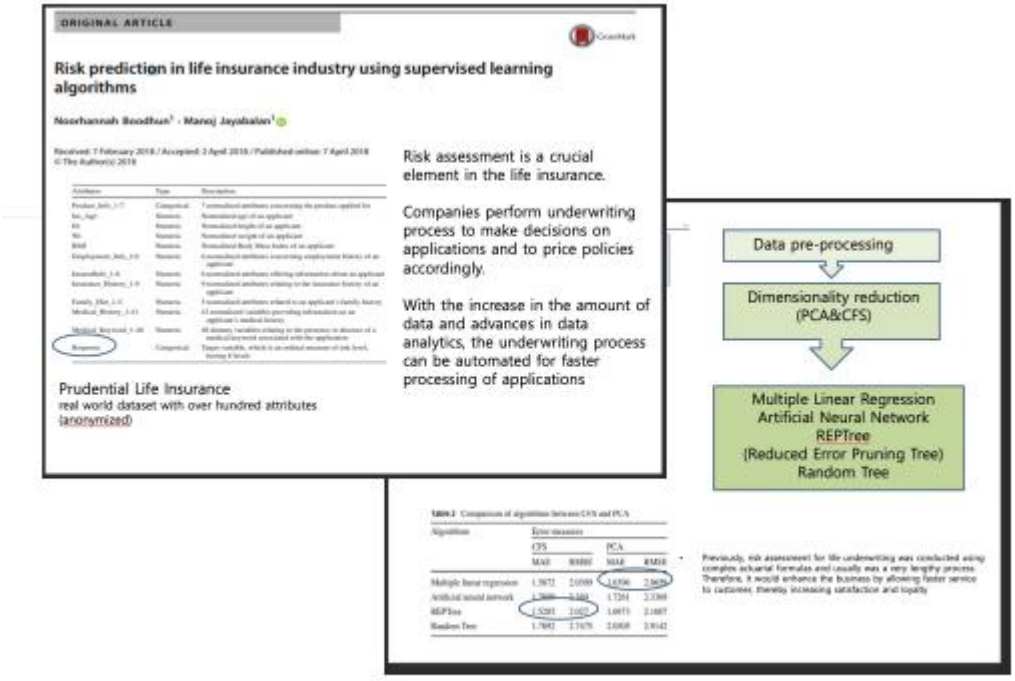
(3단 상병별 성별 연령군별 통계)

진료년도	주상병코드	성별	연령군	환자수	명세서건수	요양급여비	보험자부담금
2013	A00	남	0-4세	6	6	79730	57830
2013	A00	남	5-9세	2	2	27470	19170
2013	A00	남	10-14세	3	4	57590	43390
2013	A00	남	20-24세	1	1	177710	63010
2013	A00	남	25-29세	1	1	26420	15920

However, we found that the cost of treatment for diseases is not exactly the same by age and gender, and that the cost of treatment also differs depending on various factors such as treatment behavior and the type of hospital treated, so we investigated the data containing more detailed information. After continuous data research, we found data on the cost and drug price related to treatment along with the personal identification code of the treated subjects in the sample cohort DB of the National Health Insurance(국민건강보험) data sharing service. In particular, unlike data from other public data portals, this data is expected to be useful for predicting treatment costs because it maintains the same sample for a long period of time. We also studied machine learning methods in advance to analyze the sample cohort data of the National Health Insurance. I once again studied the concepts of Random Forest and Decision Tree, and also studied the journal of related research.

Journal study

I made a ppt material by finding papers mainly on insurance-related machine learning cases in a thesis study with a graduate student who is working on a project. The main purpose of this paper was to predict risk using Multiple Linear Regression, Random Tree, and ANN.



In insurance, risk assessment is performed when charging insurance costs or when applying for insurance, and the traditional calculation method is time-consuming and complicated. The risk calculation part was predicted using machine learning techniques using customer attributes. And it was a journal that showed how much error occurred by using the method calculated by the existing method as the correct answer. I read it because I thought it had a similar context to our study, and the paper presented models with low error among the techniques and suggested the possibility of using supervised learning in insurance cost prediction.

Future plan

In the future, I am studying R in advance to analyze using the R language in the National Health Insurance. In addition, as demo data of the sample cohort DB is provided, as in the insurance study above, the correct answer is the cost of treatment, so I will use supervised learning to find a model

with low error.

C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
RN_INST	MDCARE_	FORM_CD	MCARE_SI	SICK_SY	SICK_SYM	HSPTZ_PA	OFIJ_TYPE	OPRTN_YI	MDCARE_	VSHSP_DI	TOT_PRSC	MCARE_R	FST_HSPT	EDC_ADD_	SPCF_SYM	ED_RC_TOT_	EDC_SBA	EDC_INSU	ST
72141	20020427	03	10	A_				0	1	1	5	5		0.15		10590	3000	7590	20
82581	20020627	03	00	A_				0	1	1	1	5		0.15		12620	3000	9620	20
23181	20020806	03	23	A_		32	0	0	1	1	7	5		0.15		17130	5130	12000	20
85600	20020903	03	14	A_				0	2	2	2	5		0.15		18550	3000	15550	20
82581	20020701	03	00	A_				0	3	3	3	5		0.15		28300	9000	19300	20
82581	20020628	03	00	A_				0	2	2	3	5		0.15		17830	6000	11830	20
26501	20020322	03	14	A_				0	1	1	2	5		0.15		35920	10770	25150	20
77569	20020701	03	01	A_			0	0	4	4	3	5		0.15		40930	12000	28930	20
129487	20020826	03	10	A_			0	0	5	5	4	5		0.15		49280	15000	34280	20

The above table is a medical history table. Personal identification defense and treatment fees are listed. From the data shown above, you can see that the disease code is the same, but the price is very different(Yellow parts). The price is different even for the same disease. First of all, it seems to be necessary to calculate how much the price per disease is correct. The current priority is to create a correct answer label by analyzing the factors that change the price.