If Only..

Data Mining 2015 Team Project

Industrial Engineering
2009170845 주경돈
2011170867 이진실
2012170721 박선민
2013170805 배지원

Contents

1. The Purpose of the Project

2. Data Description

3. Experimental Results
   - 3.1 Data Mining Algorithms
   - 3.2 Comparison and Interpretation of the results

4. Project Applications

Purpose of the Project

We predict university student`s dating period and give them advices to extend their dating period

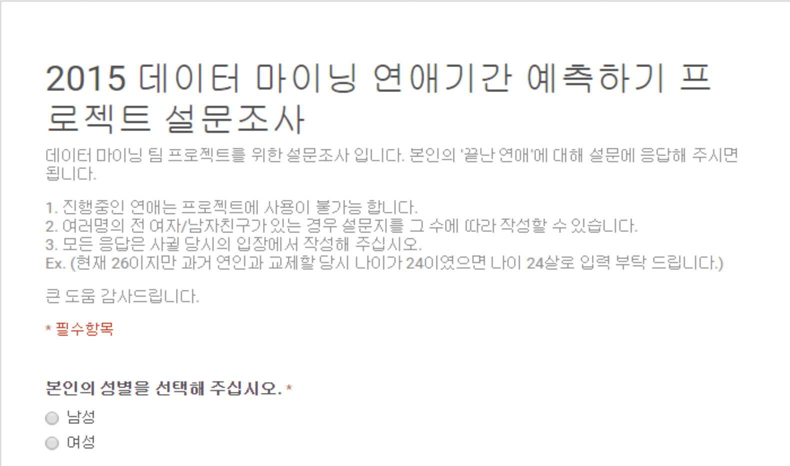
Find important factors that determine dating period

Provide the predicted dating period as a number

Data Description

Data Collection Process

- We collect data using Google Docs
- The number of questions is 21
- The respondents are University students



2015 데이터 마이닝 연애기간 예측하기 프로젝트 설문조사

데이터 마이닝 팀 프로젝트를 위한 설문조사 입니다. 본인의 '끝난 연애'에 대해 설문에 응답해 주시면 됩니다.

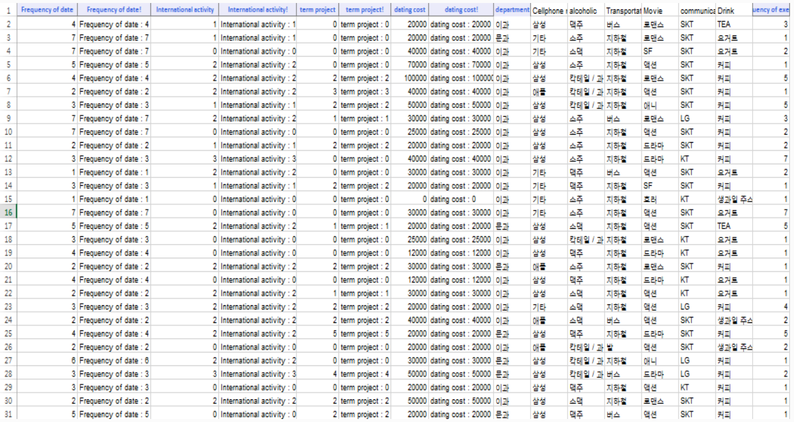
1. 진행중인 연애는 프로젝트에 사용이 불가능 합니다.
2. 여러명의 전 여자/남자친구가 있는 경우 설문지를 그 수에 따라 작성할 수 있습니다.
3. 모든 응답은 사귈 당시의 입장에서 작성해 주십시오.
Ex. (현재 26이지만 과거 연인과 교제할 당시 나이가 24이였으면 나이 24살로 입력 부탁 드립니다.)

큰 도움 감사드립니다.

* 필수항목

본인의 성별을 선택해 주십시오. *

○ 남성
○ 여성



1	Frequency of date	Frequency of date!	International activity	International activity!	term project	term project!	dating cost	dating cost!	department	Cellphone r	alcoholic	Transportat	Movie	communica	Drink	uency of exe
2	4	Frequency of date : 4	1	International activity : 1	0	term project : 0	20000	dating cost : 20000	이과	삼성	맥주	버스	로맨스	SKT	TEA	3
3	7	Frequency of date : 7	1	International activity : 1	0	term project : 0	20000	dating cost : 20000	문과	기타	소주	지하철	로맨스	SKT	요거트	1
4	7	Frequency of date : 7	0	International activity : 0	0	term project : 0	40000	dating cost : 40000	이과	기타	소맥	지하철	SF	SKT	요거트	2
5	5	Frequency of date : 5	2	International activity : 2	0	term project : 0	70000	dating cost : 70000	이과	삼성	소주	지하철	액션	SKT	커피	1
6	4	Frequency of date : 4	2	International activity : 2	2	term project : 2	100000	dating cost : 100000	이과	삼성	칵테일 / 과	지하철	로맨스	SKT	커피	5
7	2	Frequency of date : 2	2	International activity : 2	3	term project : 3	40000	dating cost : 40000	이과	애플	칵테일 / 과	지하철	액션	SKT	커피	1
8	3	Frequency of date : 3	1	International activity : 1	2	term project : 2	50000	dating cost : 50000	이과	삼성	칵테일 / 과	지하철	애니	SKT	커피	5
9	7	Frequency of date : 7	2	International activity : 2	1	term project : 1	30000	dating cost : 30000	이과	삼성	소주	버스	로맨스	LG	커피	3
10	7	Frequency of date : 7	0	International activity : 0	0	term project : 0	25000	dating cost : 25000	이과	삼성	소주	지하철	액션	SKT	커피	2
11	2	Frequency of date : 2	1	International activity : 1	2	term project : 2	20000	dating cost : 20000	이과	삼성	소주	지하철	드라마	SKT	커피	2
12	3	Frequency of date : 3	3	International activity : 3	0	term project : 0	40000	dating cost : 40000	이과	삼성	소주	지하철	드라마	KT	커피	7
13	1	Frequency of date : 1	2	International activity : 2	0	term project : 0	30000	dating cost : 30000	이과	기타	맥주	버스	액션	SKT	요거트	2
14	3	Frequency of date : 3	1	International activity : 1	2	term project : 2	20000	dating cost : 20000	이과	기타	맥주	지하철	SF	SKT	커피	1
15	1	Frequency of date : 1	0	International activity : 0	0	term project : 0	0	dating cost : 0	이과	기타	소주	지하철	호러	KT	생과일 주스	1
16	7	Frequency of date : 7	0	International activity : 0	0	term project : 0	30000	dating cost : 30000	이과	기타	소주	지하철	액션	SKT	요거트	7
17	5	Frequency of date : 5	2	International activity : 2	1	term project : 1	20000	dating cost : 20000	문과	삼성	소맥	지하철	액션	SKT	TEA	5
18	3	Frequency of date : 3	0	International activity : 0	0	term project : 0	25000	dating cost : 25000	이과	삼성	칵테일 / 과	지하철	로맨스	KT	요거트	1
19	4	Frequency of date : 4	0	International activity : 0	0	term project : 0	12000	dating cost : 12000	이과	삼성	맥주	지하철	드라마	KT	요거트	1
20	2	Frequency of date : 2	2	International activity : 2	2	term project : 2	30000	dating cost : 30000	문과	애플	소주	지하철	로맨스	SKT	커피	1
21	4	Frequency of date : 4	0	International activity : 0	0	term project : 0	12000	dating cost : 12000	이과	삼성	맥주	지하철	드라마	KT	요거트	1
22	2	Frequency of date : 2	2	International activity : 2	1	term project : 1	30000	dating cost : 30000	이과	삼성	소맥	지하철	액션	KT	요거트	1
23	3	Frequency of date : 3	2	International activity : 2	2	term project : 2	20000	dating cost : 20000	이과	기타	소맥	지하철	액션	LG	커피	4
24	2	Frequency of date : 2	2	International activity : 2	2	term project : 2	40000	dating cost : 40000	이과	애플	소맥	버스	액션	SKT	생과일 주스	2
25	4	Frequency of date : 4	2	International activity : 2	5	term project : 5	20000	dating cost : 20000	문과	삼성	맥주	지하철	드라마	SKT	커피	5
26	2	Frequency of date : 2	0	International activity : 0	0	term project : 0	20000	dating cost : 20000	이과	애플	칵테일 / 과	발	액션	SKT	생과일 주스	2
27	6	Frequency of date : 6	2	International activity : 2	0	term project : 0	30000	dating cost : 30000	문과	삼성	칵테일 / 과	지하철	애니	LG	커피	1
28	3	Frequency of date : 3	3	International activity : 3	4	term project : 4	50000	dating cost : 50000	문과	삼성	칵테일 / 과	버스	드라마	LG	커피	2
29	3	Frequency of date : 3	0	International activity : 0	0	term project : 0	20000	dating cost : 20000	이과	삼성	맥주	지하철	액션	KT	커피	1
30	2	Frequency of date : 2	2	International activity : 2	2	term project : 2	50000	dating cost : 50000	이과	삼성	소맥	지하철	로맨스	SKT	커피	1
31	5	Frequency of date : 5	0	International activity : 0	2	term project : 2	20000	dating cost : 20000	문과	삼성	맥주	버스	액션	SKT	커피	1

Variable Description

- The number of variables is 21
- Attribute characteristics are Categorical, Integer, and Real

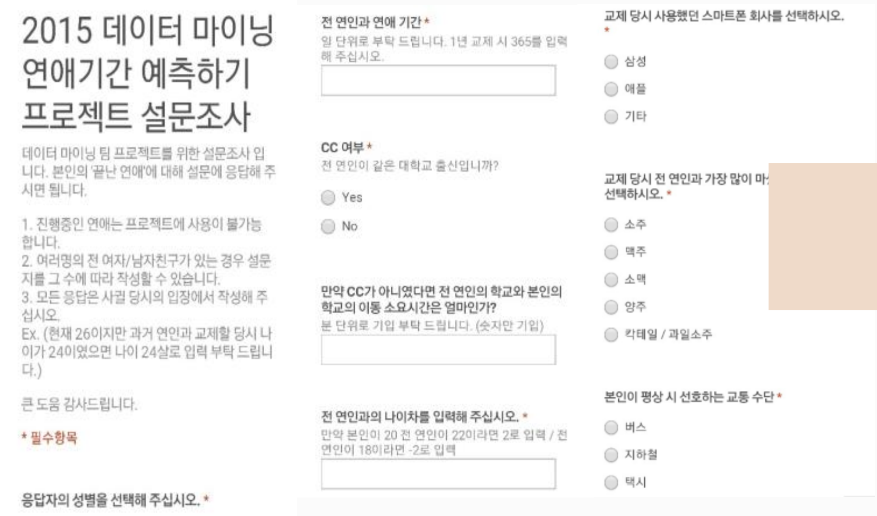
Data preprocessing results

- Data preprocessing : Normalization in K-NN prediction

2. Data Description

Survey Modification

Early survey form



2015 데이터 마이닝 연애기간 예측하기 프로젝트 설문조사

데이터 마이닝 팀 프로젝트를 위한 설문조사 입니다. 본인의 '끝난 연애'에 대해 설문에 응답해 주시면 됩니다.

1. 진행중인 연애는 프로젝트에 사용이 불가능 합니다.
2. 여러명의 전 여자/남자친구가 있는 경우 설문지를 그 수에 따라 작성할 수 있습니다.
3. 모든 응답은 사귈 당시의 입장에서 작성해 주십시오.
Ex. (현재 26이지만 과거 연인과 교제할 당시 나이가 24이였으면 나이 24살로 입력 부탁 드립니다.)

큰 도움 감사드립니다.

* 필수항목

응답자의 성별을 선택해 주십시오. *

전 연인과 연애 기간 *
일 단위로 부탁 드립니다. 1년 교제 시 365를 입력해 주십시오.

CC 여부 *
전 연인이 같은 대학교 출신입니까?
○ Yes
○ No

만약 CC가 아니였다면 전 연인의 학교와 본인의 학교의 이동 소요시간은 얼마인가?
분 단위로 기입 부탁 드립니다. (숫자만 기입)

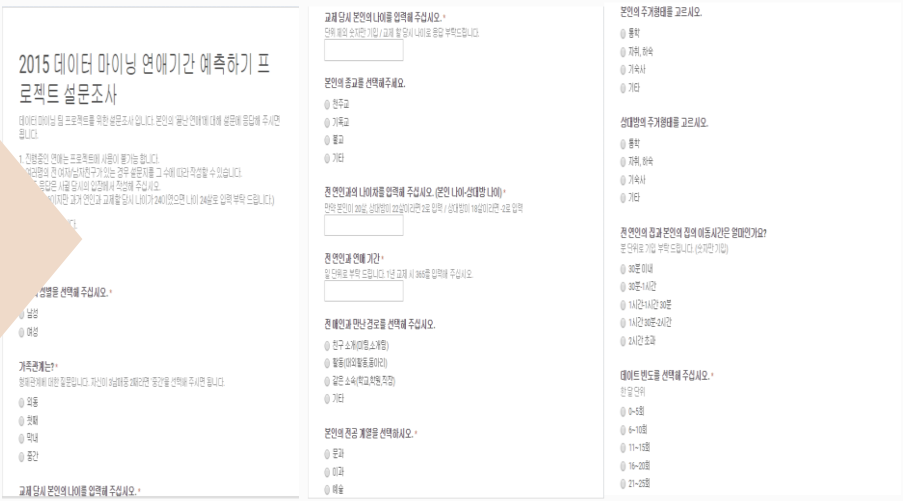
전 연인과의 나이차를 입력해 주십시오. *
만약 본인이 20 전 연인이 22이라면 2로 입력 / 전 연인이 18이라면 -2로 입력

교제 당시 사용했던 스마트폰 회사를 선택하시오. *
○ 삼성
○ 애플
○ 기타

교제 당시 전 연인과 가장 많이 마
선택하시오. *
○ 소주
○ 맥주
○ 소맥
○ 양주
○ 칵테일 / 과일소주

본인이 평상 시 선호하는 교통 수단 *
○ 버스
○ 지하철
○ 택시

Modified Survey form



2015 데이터 마이닝 연애기간 예측하기 프로젝트 설문조사

데이터 마이닝 팀 프로젝트를 위한 설문조사 입니다. 본인의 '끝난 연애'에 대해 설문에 응답해 주시면 됩니다.

1. 진행중인 연애는 프로젝트에 사용이 불가능 합니다.
여러명의 전 여자/남자친구가 있는 경우 설문지를 그 수에 따라 작성할 수 있습니다.
응답은 사귈 당시의 입장에서 작성해 주십시오.
이지만 과거 연인과 교제할 당시 나이가 24이였으면 나이 24살로 입력 부탁 드립니다.)
니다.

성별을 선택해 주십시오. *
○ 남성
○ 여성

가족관계는? *
형제관계에 대한 질문입니다. 자신이 3남매중 2째라면 '중간'을 선택해 주시면 됩니다.
○ 외동
○ 첫째
○ 막내
○ 중간

교제 당시 본인의 나이를 입력해 주십시오. *

교제 당시 본인의 나이를 입력해 주십시오. *
단위 제외 숫자만 기입 / 교제 할 당시 나이로 응답 부탁드립니다.

본인의 종교를 선택해주세요.
○ 천주교
○ 기독교
○ 불교
○ 기타

전 연인과의 나이차를 입력해 주십시오. (본인 나이-상대방 나이) *
만약 본인이 20살, 상대방이 22살이라면 2로 입력 / 상대방이 18살이라면 -2로 입력

전 연인과 연애 기간 *
일 단위로 부탁 드립니다. 1년 교제 시 365를 입력해 주십시오.

전 애인과 만난 경로를 선택해 주십시오.
○ 친구 소개(미팅,소개팅)
○ 활동(대외활동,동아리)
○ 같은 소속(학교,학원,직장)
○ 기타

본인의 전공 계열을 선택하시오. *
○ 문과
○ 이과
○ 예술

본인의 주거형태를 고르시오.
○ 통학
○ 자취, 하숙
○ 기숙사
○ 기타

상대방의 주거형태를 고르시오.
○ 통학
○ 자취, 하숙
○ 기숙사
○ 기타

전 연인의 집과 본인의 집의 이동시간은 얼마인가요?
분 단위로 기입 부탁 드립니다. (숫자만기입)
○ 30분 이내
○ 30분-1시간
○ 1시간-1시간 30분
○ 1시간 30분-2시간
○ 2시간 초과

데이트 빈도를 선택해 주십시오. *
한달 단위
○ 0~5회
○ 6~10회
○ 11~15회
○ 16~20회
○ 21~25회

What we modified

Modification of the questions

✓ Add questions : religion, major, dwelling pattern, meeting route, smoking, conflict reason

✓ Remove questions : cellphone, campus couple

Modification of the respondents

✓ The number increased from 103 to 311.

✓ More diverse respondents (ex. Major, age…)

Data Mining Algorithms

Mid Presentation	Final Presentation
MLR	MLR
K-NN	K-NN
Association Rule	CART

The reason we changed data mining algorithms

1. Association Rule : is not for predicting, only can find relations
2. K-NN Clustering : We do not need clustering because we have to predict the future value

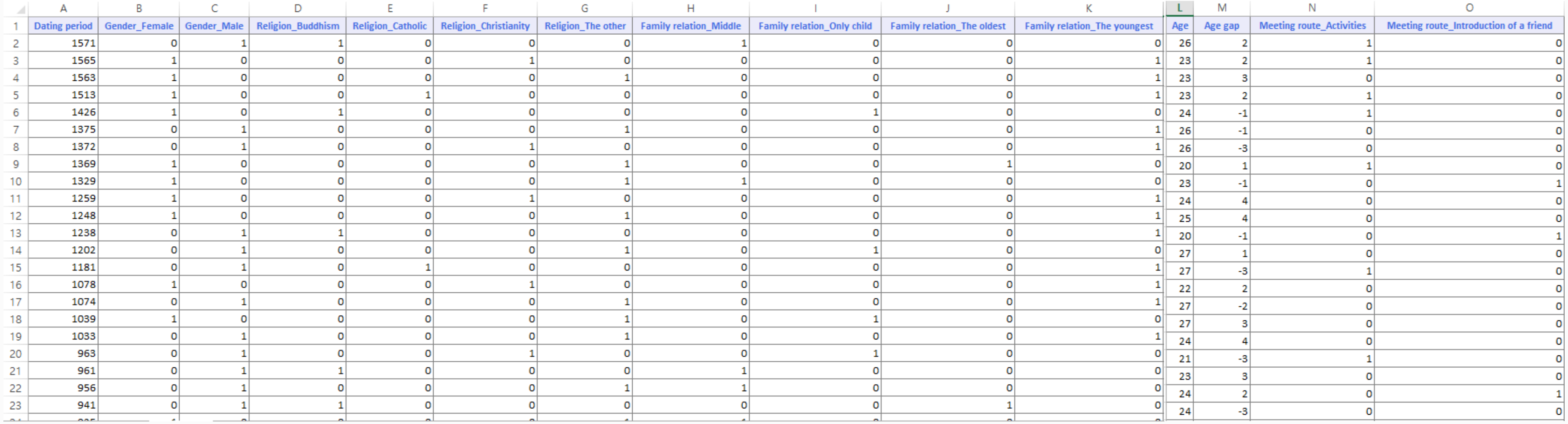
3.1 Experimental Results

Multiple Linear Regression

Information

311 people, 22 input variables, output variable : dating period

Data sets



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Dating period	Gender_Female	Gender_Male	Religion_Buddhism	Religion_Catholic	Religion_Christianity	Religion_The other	Family relation_Middle	Family relation_Only child	Family relation_The oldest	Family relation_The youngest	Age	Age gap	Meeting route_Activities	Meeting route_Introduction of a friend
2	1571	0	1	1	0	0	0	1	0	0	0	26	2	1	0
3	1565	1	0	0	0	1	0	0	0	0	1	23	2	1	0
4	1563	1	0	0	0	0	1	0	0	0	1	23	3	0	0
5	1513	1	0	0	1	0	0	0	0	0	1	23	2	1	0
6	1426	1	0	1	0	0	0	0	1	0	0	24	-1	1	0
7	1375	0	1	0	0	0	1	0	0	0	1	26	-1	0	0
8	1372	0	1	0	0	1	0	0	0	0	1	26	-3	0	0
9	1369	1	0	0	0	0	1	0	0	1	0	20	1	1	0
10	1329	1	0	0	0	0	1	1	0	0	0	23	-1	0	1
11	1259	1	0	0	0	1	0	0	0	0	1	24	4	0	0
12	1248	1	0	0	0	0	1	0	0	0	1	25	4	0	0
13	1238	0	1	1	0	0	0	0	0	0	1	20	-1	0	1
14	1202	0	1	0	0	0	1	0	1	0	0	27	1	0	0
15	1181	0	1	0	1	0	0	0	0	0	1	27	-3	1	0
16	1078	1	0	0	0	1	0	0	0	0	1	22	2	0	0
17	1074	0	1	0	0	0	1	0	0	0	1	27	-2	0	0
18	1039	1	0	0	0	0	1	0	1	0	0	27	3	0	0
19	1033	0	1	0	0	0	1	0	0	0	1	24	4	0	0
20	963	0	1	0	0	1	0	0	1	0	0	21	-3	1	0
21	961	0	1	1	0	0	0	1	0	0	0	23	3	0	0
22	956	0	1	0	0	0	1	1	0	0	0	24	2	0	1
23	941	0	1	1	0	0	0	0	0	1	0	24	-3	0	0

Process

Create dummy variables -〉 Data partition -〉 Analysis ① without variable selection
② with stepwise variable selection

3.1 Experimental Results

Multiple Linear Regression

Regression Model ①

Input Variables	Coefficient	P-Value	Input Variables	Coefficient	P-Value	Input Variables	Coefficient	P-Value	Input Variables	Coefficient	P-Value
Intercept	0	N/A	Opponent major_Science	-190.641	0.13264	Av # of international activity_1	0	N/A	Times of exercise per week_0	27.28046	0.68253
Gender_Female	0	N/A	Opponent major_art	-97.4383	0.61021	Av # of international activity_2	-101.105	0.28055	Times of exercise per week_1~2	162.0071	0.02442
Gender_Male	14.1434	0.77493	Dwelling form_Boarding house, living alone	137.2835	0.20083	Av # of international activity_3	-24.3426	0.81077	Times of exercise per week_3~4	80.20122	0.28355
Religion_Buddhism	155.128	0.07363	Dwelling form_Dormitory	45.90188	0.70088	Av # of international activity_4	18.87512	0.8573	Times of exercise per week_5~7	0	N/A
Religion_Catholic	-123.184	0.13223	Dwelling form_Living with family	40.52767	0.68813	Av # of international activity_5	-57.4488	0.55706	Favorite music_Classic	0	N/A
Religion_Christianity	0	N/A	Dwelling form_The other	0	N/A	Av # of international activity_6	-33.5284	0.74558	Favorite music_Hip hop	42.44704	0.52049
Religion_The other	80.0286	0.24161	Opponent dwelling form_Boarding house, living alone	-54.1417	0.55265	Av # of international activity_7	-187.405	0.0969	Favorite music_Jazz	136.1212	0.06769
Family relation_Middle	0	N/A	Opponent dwelling form_Dormitory	0	N/A	Favorite transportation_Bus	191.9853	0.00522	Favorite music_Rock	57.55603	0.41692
Family relation_Only child	-33.1114	0.63702	Opponent dwelling form_Living with family	-55.4125	0.49255	Favorite transportation_Subway	0	N/A	Favorite food_Chinese	0	N/A
Family relation_The oldest	-62.4591	0.40409	Opponent dwelling form_The other	-41.4851	0.69269	Favorite transportation_Taxi	16.28971	0.81775	Favorite food_Japanese	139.1853	0.12618
Family relation_The youngest	-6.47585	0.92407	Movement time_30 minutes~60 minutes	227.9566	0.00846	Favorite transportation_The other	44.23459	0.49446	Favorite food_Korean	119.2807	0.15623
Age	0.0153	0.99891	Movement time_60 minutes~90 minutes	-46.1673	0.5933	Favorite movie genre_Animation	0	N/A	Favorite food_The other	114.8016	0.15768
Age gap	9.41814	0.30837	Movement time_90 minutes~120 minutes	-75.3387	0.38229	Favorite movie genre_Art movie	184.3959	0.03679	Favorite food_Western	195.3847	0.02919
Meeting route_Activities	55.3471	0.43238	Movement time_Over 120 minutes	0	N/A	Favorite movie genre_Commercial movie	141.8344	0.17095	Smoking status_Own:Non-smoker, Opponent:Non-smoker	69.41585	0.32355
Meeting route_Introduction of a friend	30.0563	0.66482	Movement time_Within 30 minutes	92.39021	0.27991	Favorite movie genre_Documentary	148.4024	0.17454	Smoking status_Own:Non-smoker, Opponent:Smoker	-114.1905	0.15784
Meeting route_Same management	-30.0852	0.63559	Dating count_0~5	0	N/A	Favorite movie genre_Experimental film	65.77	0.43383	Smoking status_Own:Smoker, Opponent:Non-smoker	0	N/A
Meeting route_The other	0	N/A	Dating count_11~15	149.3844	0.13461	Favorite movie genre_The other	162.584	0.07933	Smoking status_Own:Smoker, Opponent:Smoker	-95.72106	0.20919
Major_Liberal	54.931	0.73971	Dating count_16~20	155.2389	0.12605	Favorite beverage_Ade	-172.39	0.73627	Conflict cause_Communication frequency	-118.2739	0.21525
Major_Physical	-115.743	0.57753	Dating count_21~25	155.209	0.13859	Favorite beverage_Juice/Smoothie	-47.6687	0.92735	Conflict cause_Difference in personality	36.56246	0.6505
Major_Science	74.312	0.65984	Dating count_26~31	-9.00234	0.93387	Favorite beverage_Tea	-182.129	0.72541	Conflict cause_Interest, gag code	-108.8552	0.18819
Major_art	0	N/A	Dating count_6~10	-51.1764	0.62064	Favorite beverage_The other	-31.5986	0.94917	Conflict cause_Physical factor	0	N/A
Opponent major_Liberal	-193.096	0.12548	Dating cost_0~10,000	-77.1149	0.29886	Favorite beverage_coffee	-138.094	0.78693	Conflict cause_Smoking, Drinking	-31.89513	0.70006
Opponent major_Physical	0	N/A	Dating cost_10,000~20,000	107.4712	0.12098				Conflict cause_The other	-112.4648	0.19458
			Dating cost_20,000~30,000	31.59956	0.69628						
			Dating cost_30,000~40,000	67.5903	0.38579						
			Dating cost_Over 40,000	0	N/A						

Residual DF	82
R?	0.7046123
Adjusted R?	0.4416451
Std. Error Estimate	224.37108
RSS	4128075.5

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
4128075.5	162.67152	-9.11863E-14

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
12563083	367.54172	108.4826908

Test Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
7181073.2	340.3289	125.8174794

3.1 Experimental Results

Multiple Linear Regression

Regression Model ①

Input Variables	Coefficient	P-Value
Movement time_30 minutes~60 minutes	227.95665	0.008459078
Favorite transportation_Bus	191.98534	0.005216039
Favorite movie genre_Art movie	184.39592	0.03678845
Times of exercise per week_1~2	162.00711	0.024417436
Favorite food_Western	195.3847	0.029192824

$$Y = 228x_1 + 192x_2 + 184x_3 + 162x_4 + 195x_5$$

x_1 : Movement time_ 30 min~60 min

x_2 : Favorite transportation_ Bus

x_3 : Favorite movie genre_ Art movie

x_4 : Times of exercise per week_ 1~2

x_5 : Favorite food_ Western

In mid-term

$$Y = 402 + 61x_1 - 306x_2 - 475 - 468x_4$$

x_1 : Number of term project

x_2 : Cellphone model _ Samsung

x_3 : Cellphone model _ Apple

x_4 : Favorite transportation _ Bus

The reason the model changed

① different respondents

② the number of respondents increased (103→311)

③ the questions were removed and added.

3.1 Experimental Results

Multiple Linear Regression

Regression Model ②

with stepwise variable selection, 13 variables selected

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	38.516509	37.10118969	1.038147552	0.30095358	-34.821123	111.85414	13745164
Religion_Catholic	-136.66671	42.728268	-3.198507973	0.001701293	-221.12734	-52.206072	192171.74
Meeting route_Activities	90.899859	41.74243184	2.177636886	0.031072874	8.3879199	173.4118	40266.893
Movement time_30 minutes~60 minutes	309.37881	39.84502985	7.764552219	1.43585E-12	230.61746	388.14017	2961446.4
Dating count_6~10	-174.67635	48.67208601	-3.588840452	0.000455479	-270.88608	-78.466618	523898.54
Dating cost_0~10,000	-104.77823	47.23020983	-2.218457769	0.028100083	-198.13781	-11.418639	485941.9
Dating cost_10,000~20,000	121.97713	40.91634281	2.981134613	0.003376438	41.09811	202.85614	538188.06
Favorite transportation_Bus	154.85734	39.71494471	3.899220952	0.000147901	76.353123	233.36157	657748.96
Favorite movie genre_Art movie	122.88162	41.34806163	2.971883349	0.003473656	41.149225	204.61401	525034.71
Times of exercise per week_1~2	135.61595	39.53319975	3.430431886	0.000787921	57.470981	213.76092	640821.17
Favorite music_Jazz	101.61259	41.23342366	2.464325832	0.014912282	20.106804	183.11838	245842.82
Smoking status_Own:Non-smoker, Opponent:Non-smoker	104.75191	35.10227488	2.984191579	0.003344869	35.36552	174.13831	372413.1
Conflict cause_Difference in personality	155.69576	45.8846976	3.393195667	0.000893921	64.995833	246.39568	506064.23

Residual DF	143
R?	0.5502525
Adjusted R?	0.5125115
Std. Error Estimate	209.64954
RSS	6285269.1

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
6285269.086	200.72416	-2.54247E-13

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
11068377.83	344.98524	0.496056782

Test Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
4671301.825	274.48785	8.725596937

3.1 Experimental Results

Multiple Linear Regression

Regression Model ②

with stepwise variable selection, 12 variables selected

Input Variables	Coefficient	P-Value
Intercept	38.516509	0.30095358
Religion_Catholic	-136.66671	0.001701293
Meeting route_Activities	90.899859	0.031072874
Movement time_30 minutes~60 minutes	309.37881	1.43585E-12
Dating count_6~10	-174.67635	0.000455479
Dating cost_0~10,000	-104.77823	0.028100083
Dating cost_10,000~20,000	121.97713	0.003376438
Favorite transportation_Bus	154.85734	0.000147901
Favorite movie genre_Art movie	122.88162	0.003473656
Times of exercise per week_1~2	135.61595	0.000787921
Favorite music_Jazz	101.61259	0.014912282
Smoking status_Own:Non-smoker, Opponent:Non-smoker	104.75191	0.003344869
Conflict cause_Difference in personality	155.69576	0.000893921

$$\boldsymbol{Y = 39 - 137x_1 + 91x_2 + 310x_3 - 175x_4 - 105x_5 + 122x_6 + 155x_7 + 123x_8}$$
$$\boldsymbol{-136x_9 + 102x_{10} - 105x_{11} - 156x_{12}}$$

x_1 : Religion_ Catholic

x_2 : Meeting route_ Activities

x_3 : Movement time_30 minutes~60 minutes

x_4 : Dating count_6~10

x_5 : Dating cost_0~10,000

x_6 : Dating cost_10,000~20,000

x_7 : Favorite transportation_ Bus

x_8 : Favorite movie genre_ Art movie

x_9 : Times of exercise per week_1~2

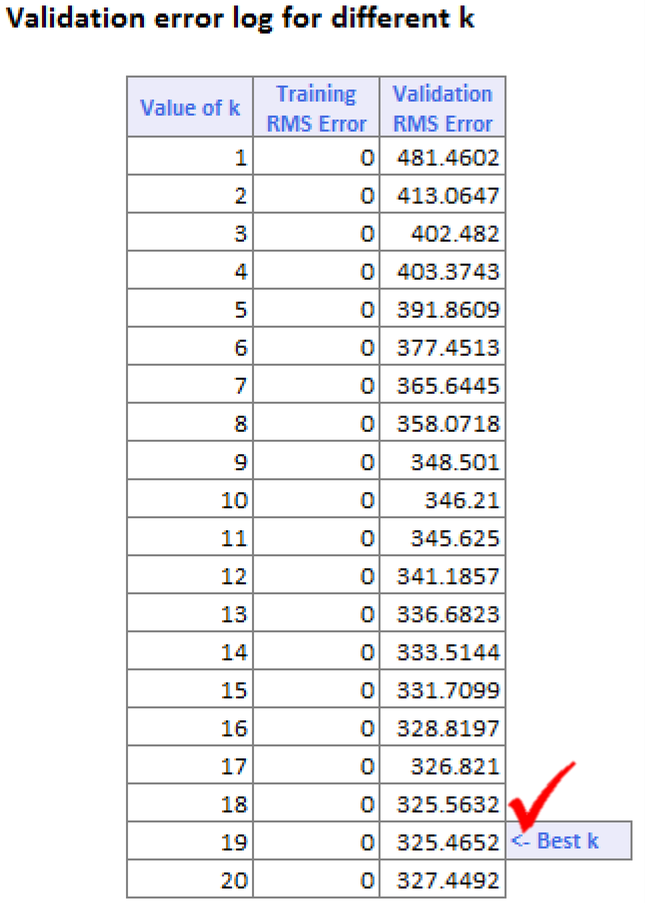
x_{10} : Favorite music_ Jazz

x_{11} : Smoking status_ Own:Non-smoker, Opponent:Non-smoker

x_{12} : Conflict cause_ Difference in personality

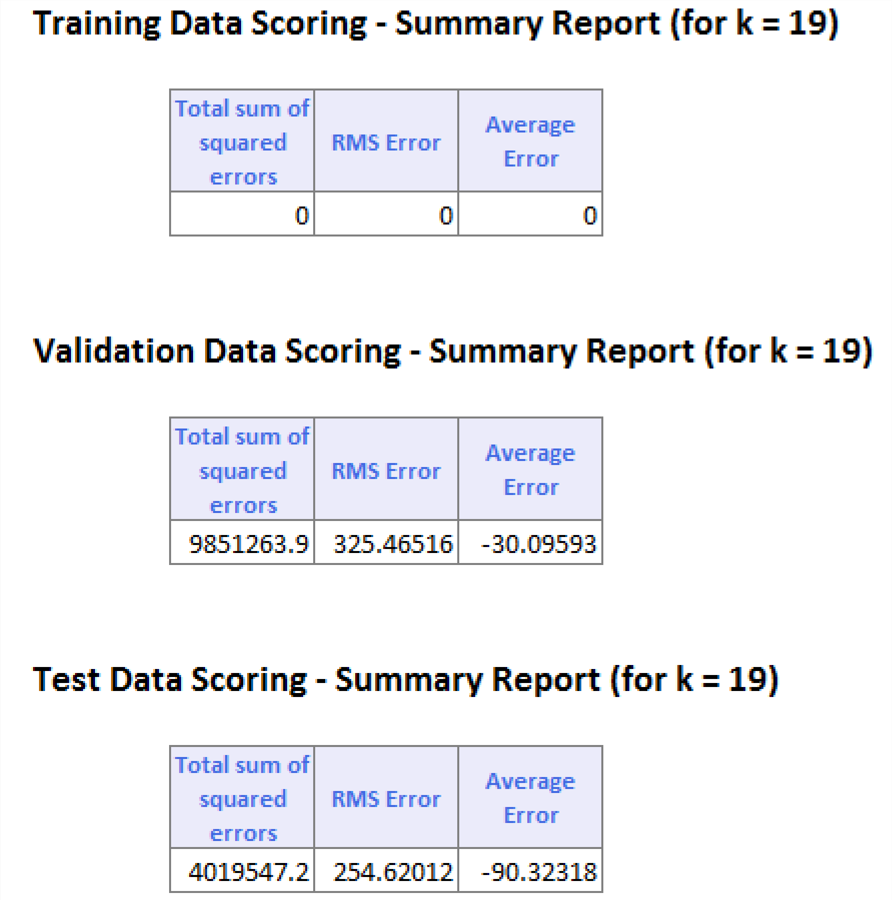
K-NN Prediction

K-NN Prediction



Validation error log for different k

Value of k	Training RMS Error	Validation RMS Error	
1	0	481.4602	
2	0	413.0647	
3	0	402.482	
4	0	403.3743	
5	0	391.8609	
6	0	377.4513	
7	0	365.6445	
8	0	358.0718	
9	0	348.501	
10	0	346.21	
11	0	345.625	
12	0	341.1857	
13	0	336.6823	
14	0	333.5144	
15	0	331.7099	
16	0	328.8197	
17	0	326.821	
18	0	325.5632	✓
19	0	325.4652	<- Best k
20	0	327.4492	



Training Data Scoring - Summary Report (for k = 19)

Total sum of squared errors	RMS Error	Average Error
0	0	0

Validation Data Scoring - Summary Report (for k = 19)

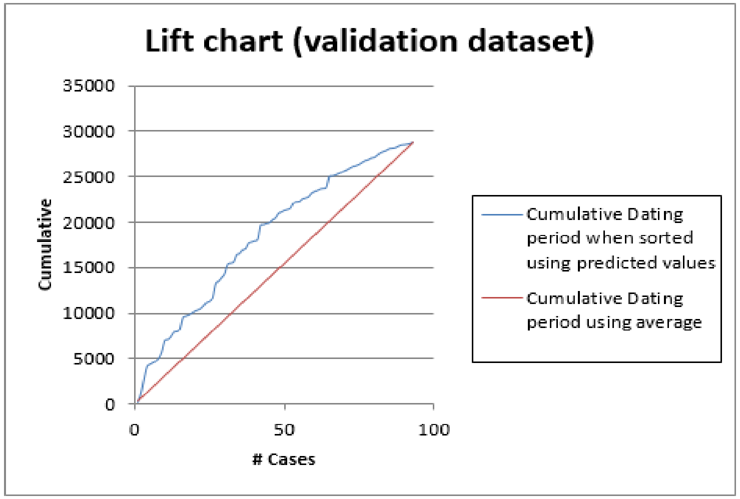
Total sum of squared errors	RMS Error	Average Error
9851263.9	325.46516	-30.09593

Test Data Scoring - Summary Report (for k = 19)

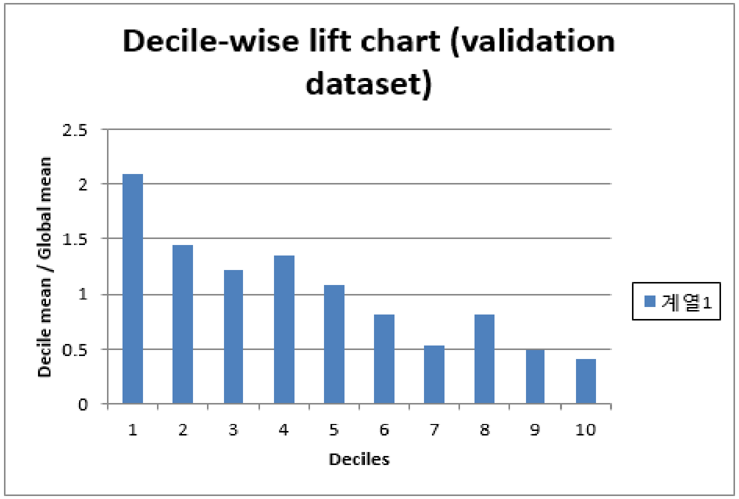
Total sum of squared errors	RMS Error	Average Error
4019547.2	254.62012	-90.32318

K-NN Prediction

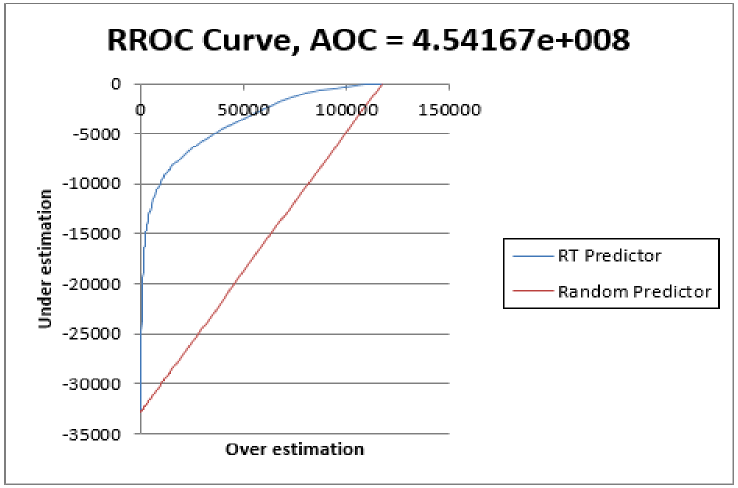
Validation Data Performance



Lift chart (validation dataset)



Decile-wise lift chart (validation dataset)



RROC Curve, AOC = 4.54167e+008

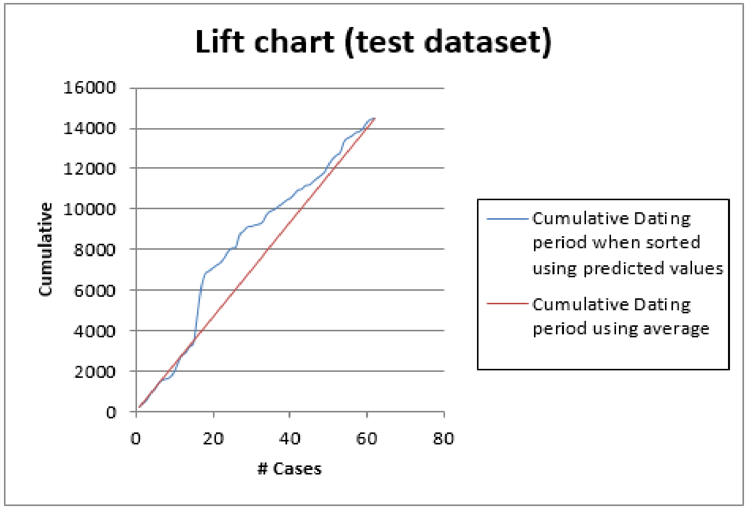
If we watch only 'AOC', K-NN`s performance is not bad

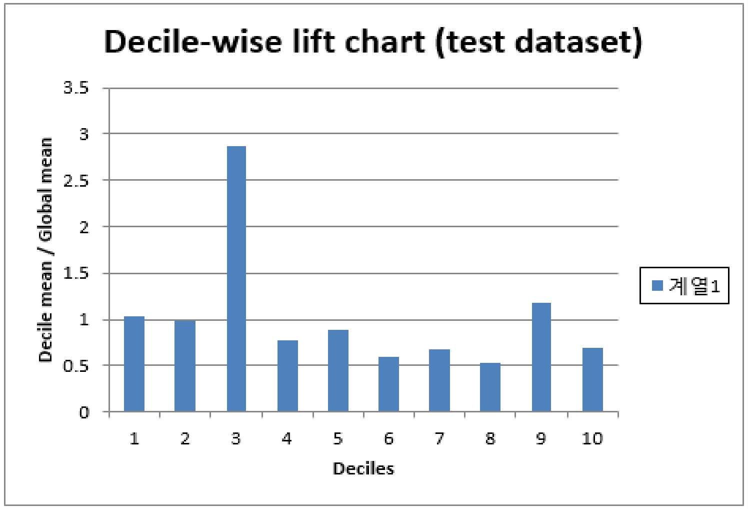
But according to Lift chart, Decile-wise lift chart, performance is very poor

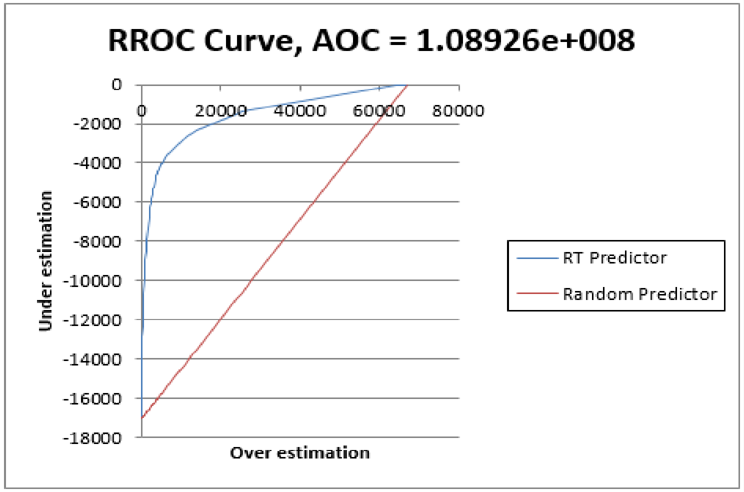
We can`t use K-NN

K-NN Prediction

Test Data Performance







If we watch only 'AOC', K-NN`s performance is not bad

But according to Lift chart, Decile-wise lift chart, performance is very poor

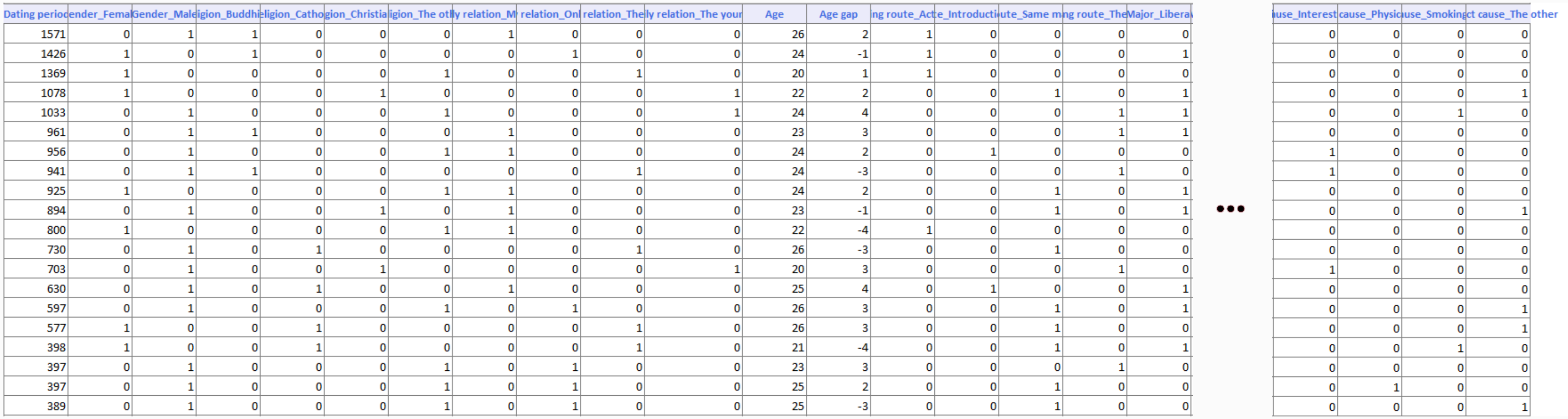
We can`t use K-NN

3.1 Experimental Results CART

Meta – Information

311 people, 22 input variables, output variable : dating period

Data sets



Dating perio	ender_Femal	Gender_Male	igion_Buddhi	eligion_Catho	gion_Christia	igion_The ot	ly relation_M	relation_Onl	relation_The	ly relation_The your	Age	Age gap	ing route_Act	e_Introducti	ute_Same m	ng route_The	Major_Libera		use_Interest	cause_Physic	use_Smoking	ct cause_The other
1571	0	1	1	0	0	0	1	0	0	0	26	2	1	0	0	0	0		0	0	0	0
1426	1	0	1	0	0	0	0	1	0	0	24	-1	1	0	0	0	1		0	0	0	0
1369	1	0	0	0	0	1	0	0	1	0	20	1	1	0	0	0	0		0	0	0	0
1078	1	0	0	0	1	0	0	0	0	1	22	2	0	0	1	0	1		0	0	0	1
1033	0	1	0	0	0	1	0	0	0	1	24	4	0	0	0	1	1		0	0	1	0
961	0	1	1	0	0	0	1	0	0	0	23	3	0	0	0	1	1		0	0	0	0
956	0	1	0	0	0	1	1	0	0	0	24	2	0	1	0	0	0		1	0	0	0
941	0	1	1	0	0	0	0	0	1	0	24	-3	0	0	0	1	0		1	0	0	0
925	1	0	0	0	0	1	1	0	0	0	24	2	0	0	1	0	1		0	0	0	0
894	0	1	0	0	1	0	1	0	0	0	23	-1	0	0	1	0	1	•••	0	0	0	1
800	1	0	0	0	0	1	1	0	0	0	22	-4	1	0	0	0	0		0	0	0	0
730	0	1	0	1	0	0	0	0	1	0	26	-3	0	0	1	0	0		0	0	0	0
703	0	1	0	0	1	0	0	0	0	1	20	3	0	0	0	1	0		1	0	0	0
630	0	1	0	1	0	0	1	0	0	0	25	4	0	1	0	0	1		0	0	0	0
597	0	1	0	0	0	1	0	1	0	0	26	3	0	0	1	0	1		0	0	0	1
577	1	0	0	1	0	0	0	0	1	0	26	3	0	0	1	0	0		0	0	0	1
398	1	0	0	1	0	0	0	0	1	0	21	-4	0	0	1	0	1		0	0	1	0
397	0	1	0	0	0	1	0	1	0	0	23	3	0	0	0	1	0		0	0	0	0
397	0	1	0	0	0	1	0	1	0	0	25	2	0	0	1	0	0		0	1	0	0
389	0	1	0	0	0	1	0	1	0	0	25	-3	0	0	1	0	0		0	0	0	1

Process

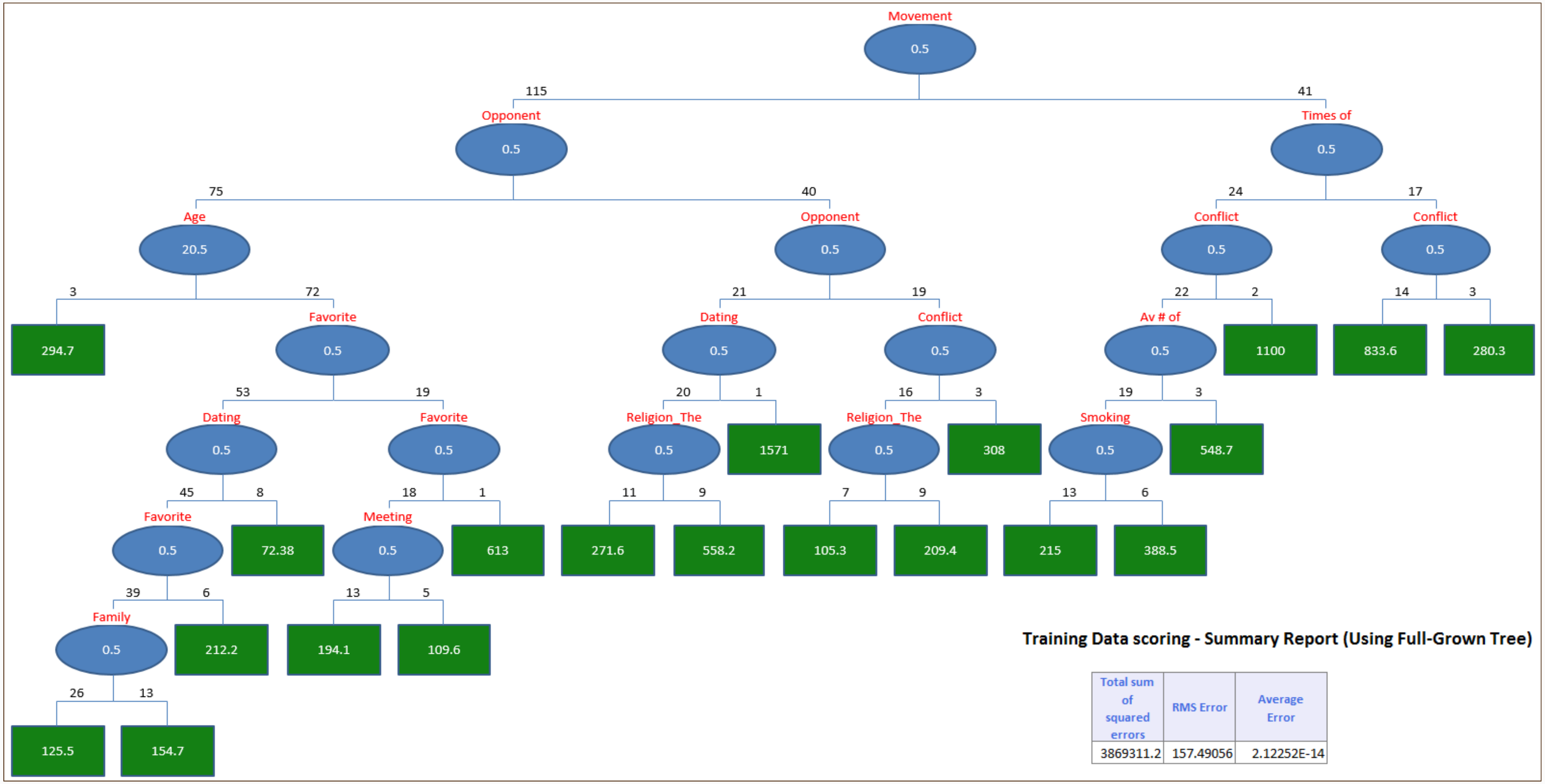
Create dummy variables -〉 Data partition -〉 Build Tree

①Full Tree
②Min-Error Tree
③Best-Pruned Tree

3.1 Experimental Results CART

① Full Tree

Using Training data
#of terminal node=20

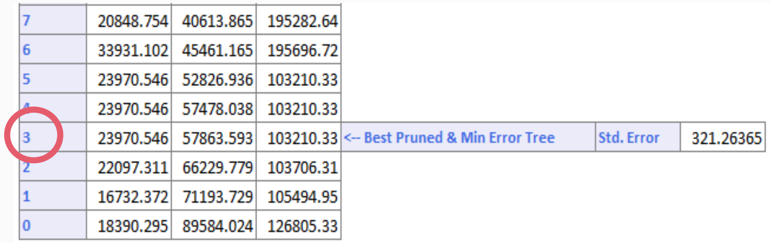


Training Data scoring - Summary Report (Using Full-Grown Tree)

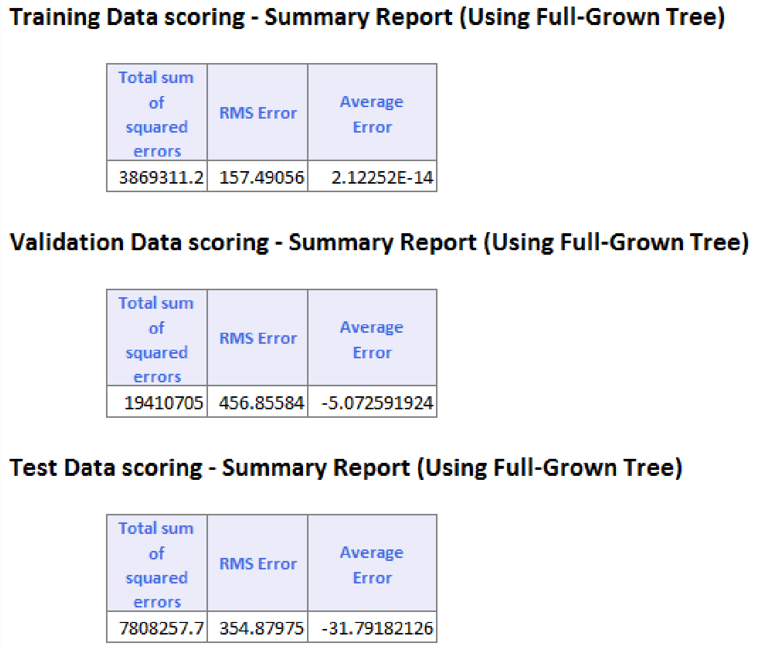
Total sum of squared errors	RMS Error	Average Error
3869311.2	157.49056	2.12252E-14

3.1 Experimental Results CART

Leaf node Penalty = a
Cost complexity= Error +a*Leaf node



7	20848.754	40613.865	195282.64			
6	33931.102	45461.165	195696.72			
5	23970.546	52826.936	103210.33			
4	23970.546	57478.038	103210.33			
3	23970.546	57863.593	103210.33	<-- Best Pruned & Min Error Tree	Std. Error	321.26365
2	22097.311	66229.779	103706.31			
1	16732.372	71193.729	105494.95			
0	18390.295	89584.024	126805.33			



Training Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
3869311.2	157.49056	2.12252E-14

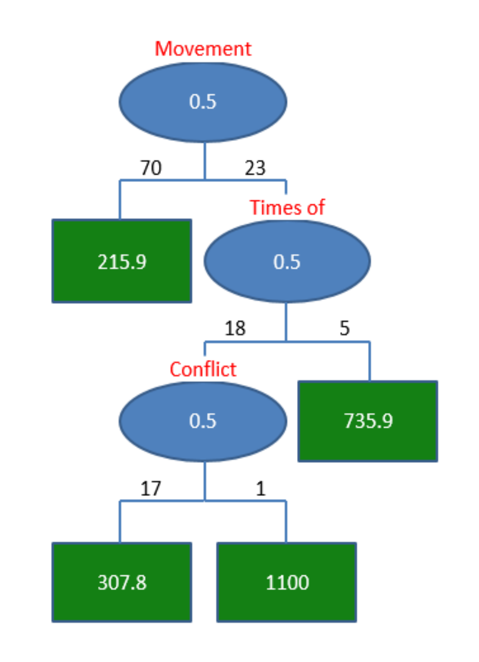
Validation Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
19410705	456.85584	-5.072591924

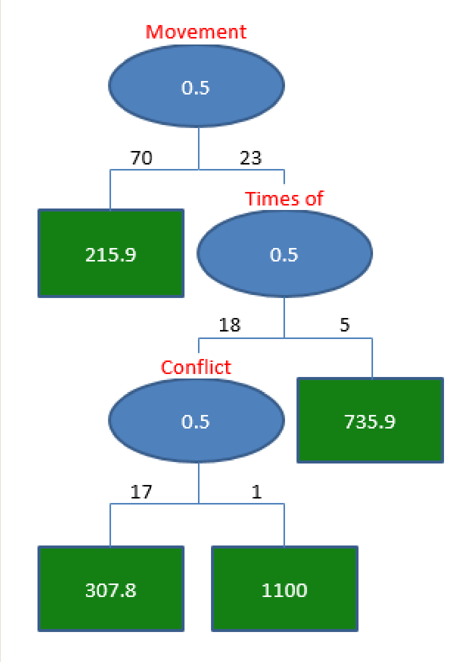
Test Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
7808257.7	354.87975	-31.79182126

② Min-Error Tree



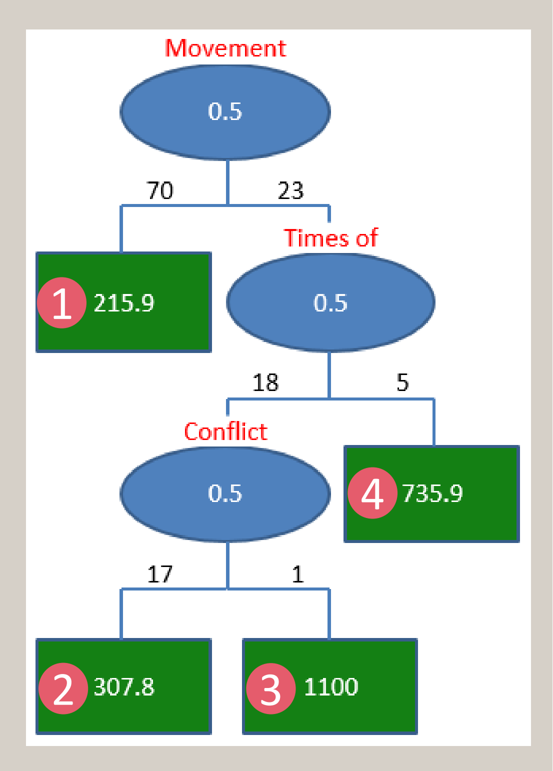
③ Best-Pruned Tree



3.1 Experimental Results CART

③ Best-Pruned Tree & Rules

Best-Pruned Tree & 4 Rules



1 $IF(Movement\ time \neq 30min{\sim}60min)$
$THEN(Dating\ period = 215.9)$

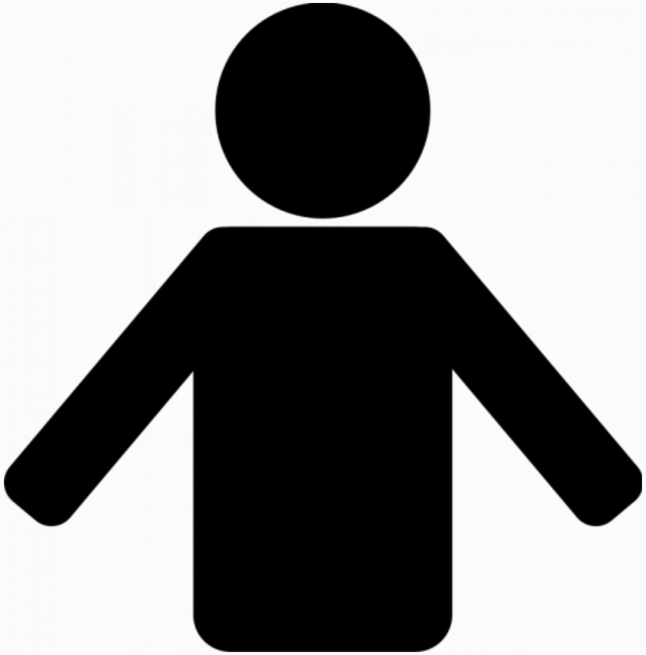
2 $IF(Movement\ time = 30min{\sim}60min)$
$ADN\ IF(Exercise/week\ \neq 1{\sim}2)$
$AND\ IF(Conflict\ cause \neq diff\ in\ personality)$
$THEN(Dating\ period = 307.8)$

3 $IF(Movement\ time = 30min{\sim}60min)$
$ADN\ IF(Exercise/week\ \neq 1{\sim}2)$
$AND\ IF(Conflict\ cause = diff\ in\ personality)$
$THEN(Dating\ period = 1100)$

4 $IF(Movement\ time = 30min{\sim}60min)$
$ADN\ IF(Exercise/week\ = 1{\sim}2)$
$THEN(Dating\ period = 735.9)$

3.2 Experimental Results

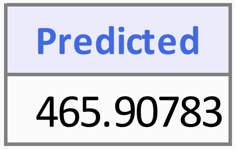
Comparison and Interpretation



We compare 3 different algorithms' results.
And we choose MLR based on test dataset RMSE

The predicted value of new data

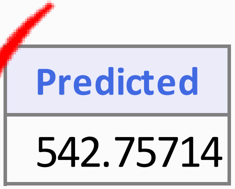
Multiple Linear Regression Model



Predicted
465.90783

Without
variable selection

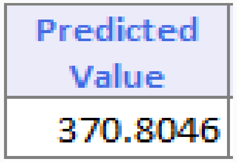
✓



Predicted
542.75714

With
variable selection

K-NN Prediction



Predicted Value
370.8046

We use ‘k=19’

Regression Tree

Rule3. Predicted dating period=1100days

4 . Project Applications

Private & Public Applications

Project Results	Possible Applications
We find important factors that determine dating period	✓ **Private** -develop their relationship -entertainment
We provide the predicted dating period as a number	✓ **Public** -service for couple-matching/ counseling ex. media program, web/application service

EOD