Proceedings of the Federated Conference on Computer Science and Information Systems pp. 725–732
DOI: 10.15439/2019F318
ISSN 2300-5963 ACSIS, Vol. 18

Analysis of Relationship between Personal Factors and Visiting Places using Random Forest Technique

Young Myung Kim
Department of Computer Engineering, Hongik University
Seoul, Republic of Korea
Email: dudaud0205@gmail.com

Ha Yoon Song
Department of Computer Engineering, Hongik University
Seoul, Republic of Korea
Email: hayoon@hongik.ac.kr

***Abstract*—There has been research regarding relationship between human personalities and visiting places using Big Five Factor (BFF). However, other factors such as Social media usage, Hobby, Gender, Age, and Religion and so on are regarded as also major factors which effects the choice of visiting place of a person. Using questionnaire designed by authors, these factors as well as BFF were prepared for this research. The visiting places were collected by a smartphone app called SWARM and classified in 10 categories. In sum, personal data of 34 participants had been collected for several months. To figure out the relationship between these factors and visiting places, random forest technique of ensemble method was used.**

I. Introduction

PRIOR researches show that human personality and favorite visiting place have considerable relationship [1] [2] [3] [4]. However, there has been long belief that other than personalities, personal factors effect the selection of visiting location. To prove this belief, we collected personal factors other than personality from survey. Gender, Age, Marital Status, Religion, Salary, Vehicles, usage of SNS, Job, Educational Level, Frequency of travel for a year, Time spent on SNS per day, sort of hobby. Using Big Five Inventory (BFI), we collected person's Big Five Factor (BFF) Total 34 participants provided their personal data and location data. To collect location data, a smartphone application called SWARM is used and the duration of collection was up to six months. The method to analyze these data is Random Forest which is ensemble learning.

A. Random Forest

Random Forest is suggested by Leo Breiman in 2001 [5]. Random Forest shows good performance and high accuracy in general and without overfitting. It can handle many input features and resistant to noise. In addition, the degree of effect of input feature can be numerically represented as importance value. We considered Random Forest as a suitable method for our research.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2019R1F1A1056123).

B. Big Five Factors (BFF)

BFF is a factor of personality suggested by P.T. Costa and R.R. McCrae in 1992 [6]. It has five factors of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Set of questionnaires is answered by participants and each five factors will be valued as a score from 0 to 5 points. Since BFF can numerically represent conceptual human personality, many of research adopts BFF [7] [8] [9] [10] [11] [12].

C. SWARM Application

SWARM application is used to collect geo-positioning data installed on smartphones [13]. Users actively check in visited places with SWARM. These actively collected location data are used as part of our analysis.

In section II, we will discuss random forest and our purpose in this research. Section III will show details of the data. The handling of personal factors and location categories will be discussed. Section IV will show results of analysis by Random Forest and evaluate the results. Section V will conclude this research with future works.

II. Random Forest Technique

A. Ensemble

Ensemble is a technique which combines various machine learning models to generate powerful model. Random Forest is a sort of ensemble technique and has decision tree as its base model. Especially, Random Forest and gradient boosting have proven as useful method for classification and regression of various data set. These two distinguished models have base element of decision tree.

B. Decision Tree

Decision tree is a widely used model for classification and regression. Basically, decision tree is a consequence of yes-no question of leaning process toward the final decision. For example, to classify bear, pigeon, penguin, dolphin with the smallest number of questions, several sequences of question are introduced. The first question to classify two animals is: "Does it have wings? " Then the second question is: "can

it fly?" Then pigeon and penguin can be classified. In case there is no wing, the following question will be:"Does it have fin", and dolphin and bear can be separated. These questions are called as test in machine learning. And decision tree is consisted as nodes for test and edge connected to the following test. In case of machine learning, continuous values can be used instead of yes-no question. In this case, test can be in a form that is feature i bigger than value a.

C. Bootstrap Aggregating (Bagging)

Random Forest creates bootstrap samples of data to create several independent decision trees. Bootstrap samples are random choices of data by allowing redundancy. The size of the dataset is the same as the original dataset. Some data will be missing from the bootstrap sample and some data may be duplicated [14].

The disadvantage of the decision tree is that it can be overfitted to the training data whereas Random Forest can handle this problem. Random Forest is a bundle of different decision trees. Each decision tree is relatively good at prediction but can be overfitted in the training data. However, if we create many of decision trees and average its results, the prediction performance of the decision tree can be enhanced by reducing the overfitting. In addition, each branch of the decision tree uses a subset of different features because only part of the features is used in each node. This method makes all the decision trees in the Random Forest different from each other. Random Forest predicts with results from each decision tree. For the regressions used in this study, average of each result is used to make the final prediction.

Random Forest is one of widely used machine learning algorithm with excellent performance. It is strong in noise, works well even without much hyperparameter tuning, and does not need to scale data. It also works well on very large datasets and can parallelize the train simply. It is also appropriate to deal with many input features [15]. We can also know the value importance of the input value that affects the result. Due to this advantage and performance, a Random Forest was used for this study.

III. Personal Factors and Location Categories

A. Personal Factors

Many of research adopts BFF as a measure of personality suggested by McCrae and Costa. [6] The five factors are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each factor are measured as numerical numbers so that factors can be easily applied to training process. Table I shows BFF of several participants. We can figure out personality of a person through these values. Person with high Openness is creative, emotional and interested in arts. Person with high Conscientiousness is responsible, achieving, and restraint. Person with high Agreeableness is agreeable to other person, altruistic, thoughtfulness and modesty. While person with high Neuroticism is sensitive to stress, impulsive, hostile and depressed. For example, as shown in table I, person 4 is creative, emotional, responsible, restraint. Also considering person 4's Neuroticism, person 4 is not impulsive and resistant to stress. The personality shown in table I will be used our experimental basis with other personal factors.

TABLE I: BFF of Participants

	O	C	E	A	N
Person 1	3.3	3.9	3.3	3.7	2.6
Person 2	2.7	3.2	3.2	2.7	2.8
Person 3	4.3	3.1	2.3	3.2	2.9
Person 4	4.2	4.3	3.5	3.6	2.6
Person 5	4	3.7	4	3.9	2.8
Person 6	3.8	4	3.1	3.8	2.3
Person 7	3.2	3.2	3.5	3.3	3.5
Person 8	2.8	3.8	3.8	3.3	2.3
Person 9	3.4	3.6	3.5	3.6	3.1
Person 10	3	3.6	2.5	3	3
Person 11	4.1	3.8	3.8	2.8	3
Person 12	3.1	3	2.8	3	2.8
Person 13	3.3	3.2	3.5	2.6	2.6
Person 14	3.7	3.3	3.6	3.8	3.5
Person 15	2.4	3.7	3	2.8	2.6
Person 16	3.4	3.2	3.0	3	2.6
Person 17	3.9	3.3	3.5	2.9	2.8

TABLE II: Personal Factors: Person 1

Personal Factors	Value
Age	2
Job	1
Marriage	2
The highest level of education	2
Major	4
Religion	1
Salary	2
Vehicles	4
Commute time	3
the frequency of a year's journey	2
SNS usage status	1
Time spent on SNS per day	3
cultural life	3
Openness	3.3
Conscientiousness	3.9
Extraversion	3.3
Agreeableness	3.7
Neuroticism	2.6

In the table II, the number corresponding to the response is as follows:

Age

1: 10s, 2: 20s, 3: 30s, 4: over 40s

Job

1: students, 2: administrative position, 3: expert, 4: an engineer, 5: office job, 6: service, sales position, 7: a functional worker, 8: equipment maneuvering and assembly engineer, 9: simple laborer

cf. Occupational classifications include the International Standard Classification of Occupation (ISCO) [16].

Marriage
1: married, 2: single
The highest level of education
1: middle school graduate, 2: high school graduate, 3: college graduate, 4: master, 5: doctor
Major
1: humanities, 2: sociology, 3: pedagogy, 4: engineering, 5: nature, 6: medicine and pharmacology, 7: art, music and physical education
Religion
1: no religion, 2: Christianity, 3: Catholic, 4: Buddhism
Salary
1: Less than 500 USD, 2: 500 USD to 1,000 USD, 3: 1,000 USD to 2,000 USD, 4: 2,000 USD to 3,000 USD, 5: over 3,000 USD
Vehicles
1: walking, 2: bicycle, 3: car, 4: public transport
Commute time
1:less than 30mins, 2: 30mins to 1h, 3: 1h to 2h, 4: over 2 hours
The frequency of a year's journey
1: less than one time, 2: 2 to 3 times, 3: 4 to 5 times, 4: over six times
SNS usage status
1: Use, 2: Not use
Time spent on SNS per day
1: less than 30 mins, 2: 30 mins to 1 hour, 3: 1 hour to 3 hours, 4: over 3 hours
Cultural life
1: static activity, 2: dynamic activity, 3: both

In case of Person 1, a number of personal factors are coming from 20s, such as students, single, a high school graduate, engineering, no religion, income in 500USD to 1000USD, public transport, commute in 1 to 2hours, two or three travels per year, one to three hours spent for social media per day, and both dynamic and static cultural life.

B. Location Category Data

Location Category data was used as Label (target data) for the supervised learning, Random Forest. The location category data is checked in to the visiting places using the SWARM application. Afterwards, the number of visits and visiting places were identified from web page of SWARM. Part of the location data of person 16 is shown in the table III.

TABLE III: Sample Location Data: Person 16

location	Count of Visit
Hongik Univ. Wowkwan	19
Hongik Univ. IT Center	7
Kanemaya noodle Restaurant	3
Starbucks	3
Hongik Univ. Central Library	8
Coffesmith	2
Daiso	3

The data collected were classified into 10 categories. Table IV shows the classification of person 16's location data into a category.

TABLE IV: Sample Classification of Locations: Person 16

Category	location	Visiting Ratio
Foreign Institutions	0	0
Retail Business	6	0.04
Service industry	6	0.04
Restaurant	29	0.1933
Pub	2	0.0133
Beverage Store	26	0.1733
Theater and Concert Hall	4	0.0267
Institutions of Education	62	0.4133
Hospital	6	0.04
Museum, Gallery, a historical site, tourist spots	9	0.06

To input categorized location data to Random Forest, visiting ratio of location categories are used as labels. The formula is as follows.

$$Visiting\_Ratio = \frac{count\_of\_visit\_to\_location}{total\_count\_of\_visits}$$

IV. Analysis of Results

By analyzing data using random forest, you can see value importance, which is the degree of how each feature affects the prediction. Table V shows summary of result for each Location Category such as Symmetric Mean Absolute Percentage Error (SMAPE), Accuracy and the top five feature's importance values with the most impact. Table V abbreviated location category.

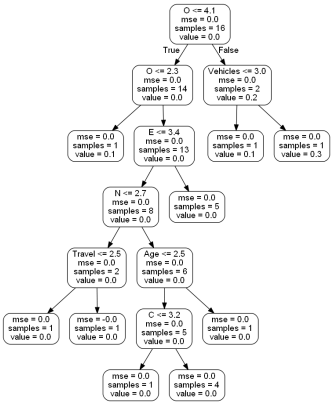
FI: Foreign Institutions
RB: Retail Business
SI: Service Industry
BS: Beverage Store
TC: Theater and Concert Hall
IE: Institutions of Education
MG: Museum, Gallery, historical sites and tourist spots

The result of the experiment randomly selected one of the decision trees is present in Fig. 1. The unbiased and well-made decision tree is found as shown in Fig. 2 when it has label of Restaurant. Several significant value importance graphs with meaningful accuracy are also shown Fig. 3.

A. Discussion about Low Prediction Accuracy

First, we analyzed the reason of very low accuracy, especially for foreign institutions and hospital. This is just because of shortage of data, since most of participants rarely went to foreign instruments. A handful of people have visited international airport only once or twice while traveling abroad.

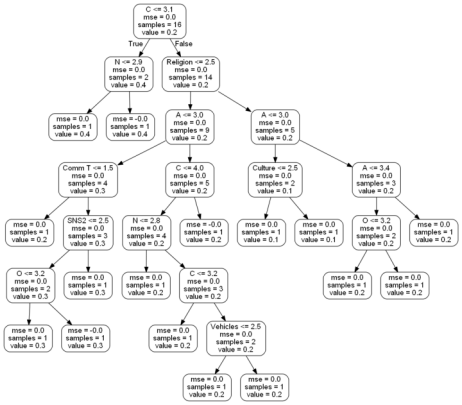
It would have been difficult to predict because the person who went to foreign institutions lacked data. Hospital shows similar situation. Hospital is not a place to go by a person's



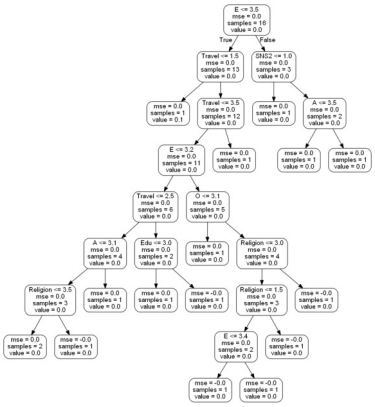
(a) label=Foreign Institutions

(b) label=Retail Business

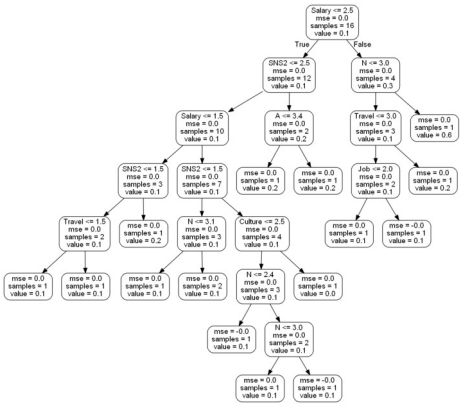
(c) label=Service industry



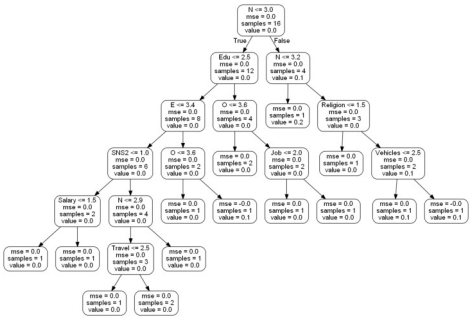
(d) label=Restaurant



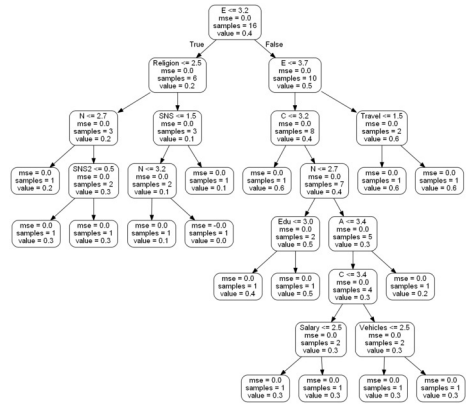
(e) label=Pub



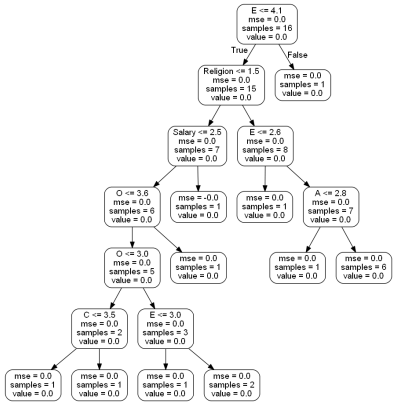
(f) label= Beverage Store



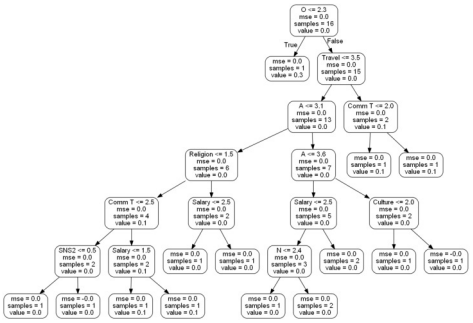
(g) label=Theater and Concert Hall



(h) label= Institutions of Education



(i) label=Hospital



(j) label=Museum, Gallery, historical sites and tourist spots

Fig. 1: Decision Tree

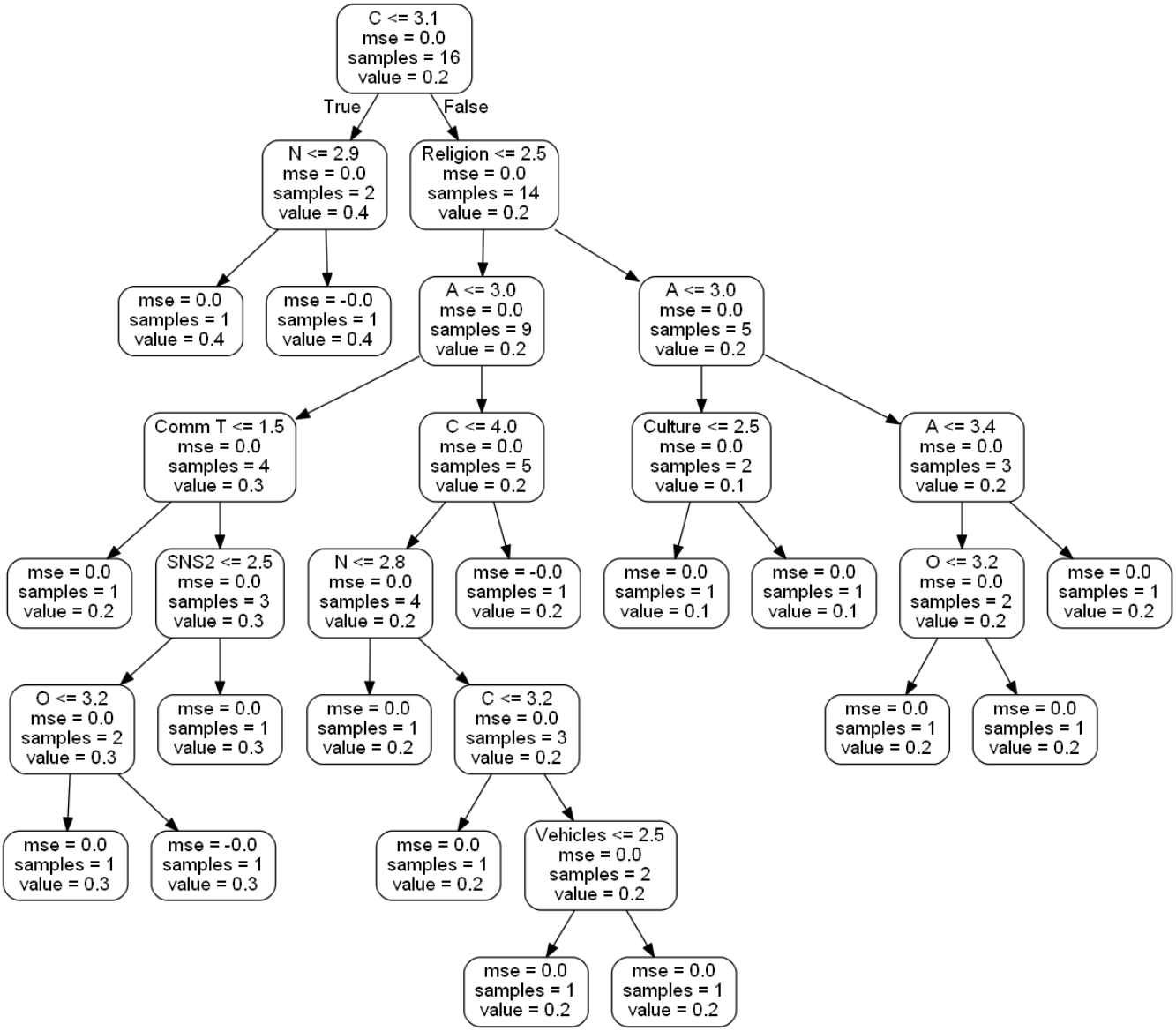


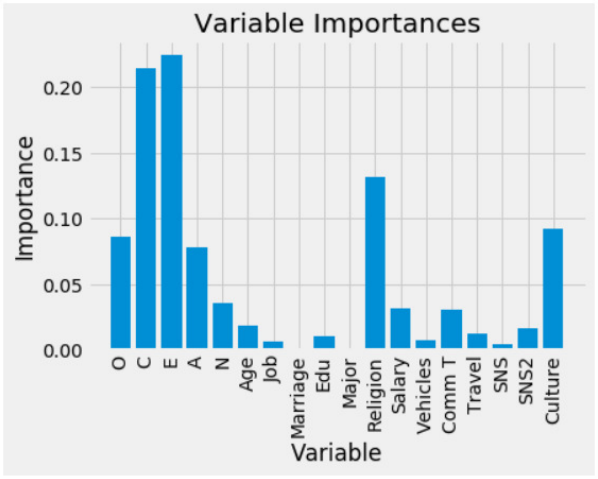
Fig. 2: Decision Tree for Restaurant

personality or preference. In general, once someone had disease or accident, visit to the hospital will be taken.

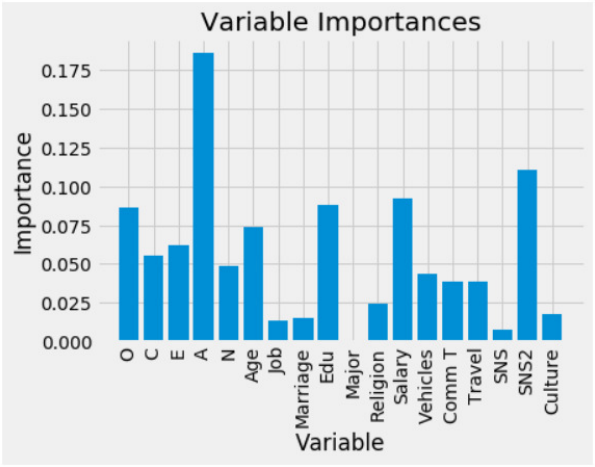
For these reasons, most participants rarely went to hospital. One participant frequently visited the hospital during the data collection period because of the need for continuous processing, and this was caused by accident but not by personality or other factors. For foreign institutions, we think significant results could be obtained if the number of participants increased and the age group varied.

However, in the case of hospital, we decided that the visit frequency was not affected by personality or the personal factors we collected. In the case of service industry, it is difficult to describe this category as specific place because it contains too many locations as previously discussed. For example, banks, beauty salons, massage parlors, bus terminals, hotels, guest houses and photo studios are included in service industry. These diversities of location category maybe attenuate the accuracy. Prediction accuracy is 59.79%, not very low, but it is also not that high. This would identify better predict accuracy and the affecting factors if the categories were more granular and grouped into units with one characteristic. For the category of museum, gallery, historical sites and tourist spots, the value importance is considered to have a significant result, although the predict accuracy 44.44% which was not high enough.

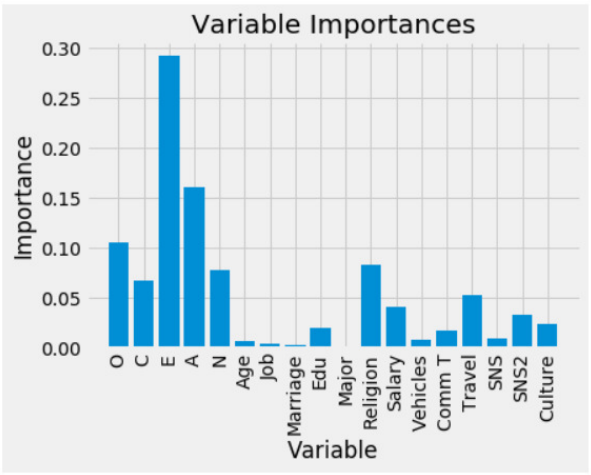
The results of experiments showed that openness and travel frequency affected the visit of museum, gallery, historical sites and tourist spots. Intuitively, open people like to travel because



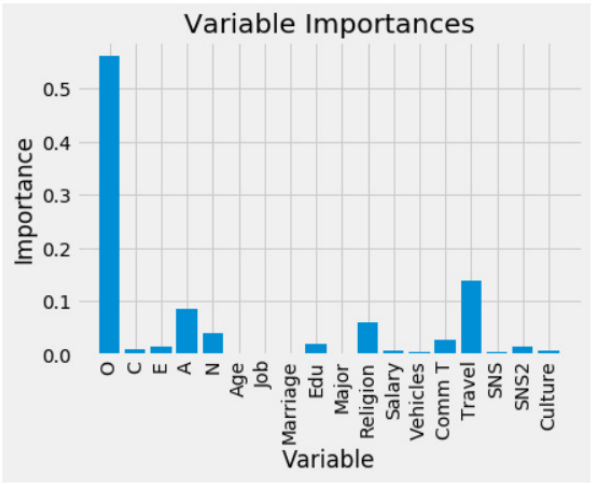
(a) label=Restaurant



(b) label=Pub



(c) label= Institutions of Education



(d) label=Museum, Gallery, historical sites and tourist spots

Fig. 3: Value Importance Graph

TABLE V: Summary of Results

	FI	**RB**	**SI**	**Restaurant**	**Pub**	**BS**	**TC**	**IE**	**Hospital**	**MG**
SMAPE(%)	97.43	50.92	40.21	21.74	34.41	29.97	35.54	23.02	87.4	55.56
Accuracy(%)	2.57	49.08	59.79	78.26	65.59	70.03	64.46	76.98	12.6	44.44
Feature 1	O	A	A	E	A	Religion	N	E	Age	O
	0.55	0.44	0.16	0.22	0.19	0.21	0.21	0.29	0.17	0.56
Feature 2	C	N	Age	C	SNS2	E	Religion	A	Edu	Travel
	0.11	0.12	0.15	0.21	0.11	0.11	0.21	0.16	0.17	0.14
Feature 3	A	O	O	Religion	O	SNS2	Salary	O	A	A
	0.1	0.09	0.12	0.13	0.09	0.09	0.13	0.11	0.13	0.09
Feature 4	E	E	C	O	Edu	N	E	N	E	Religion
	0.06	0.06	0.1	0.09	0.09	0.08	0.08	0.08	0.09	0.06
Feature 5	Job	C	Marriage	Culture	Salary	Salary	C	Religion	C	N
	0.03	0.05	0.09	0.09	0.09	0.08	0.07	0.08	0.08	0.04

TABLE VI: Statistics on Survey

Answers	Age	Job	Marriage	Edu	Major	Religion	Salary	Vehicles	Comm T	Travel	SNS1	SNS2	Culture
1	0	32	1	0	0	23	11	9	13	9	25	4	10
2	30	0	33	25	0	5	17	1	8	16	9	14	8
3	3	2		5	0	3	3	0	13	6		7	16
4	1			3	34	3	1	24	0	3			
5							2						

they love adventure. Frequent travel increases the chances of visiting museum, gallery, historical sites and tourist spots. We also expect to have a high degree of predict accuracy if it gets data from a wider range of ages and occupational groups.

B. Interpretation of Results

The experimental results show that the predict accuracy is usually high when the characteristics of the category are clear. For example, Restaurant, Pub, Beverage Store, Theater and Concert Hall, Institutions of Education are clear categories. While, Retail Business, Service industry, Museum, Gallery historical sites and tourist spots are not easy to clarify.

Therefore, it is hard to say that the category has one characteristic. For these reasons, it would have been difficult to predict by personality or personal factors. The highest predict accuracy was restaurant category, which was 78.26%. The most affected features are E (Extraversion) and C (Conscientiousness) among personality factors, and followed by Religion, O (Openness) and Cultural Life. For the institutions of education, the predict accuracy is 76.98%, followed by restaurant with a higher predict accuracy. Effective features include E(Extraversion), A(Agreeableness), O(Openness), N(Neuroticism), and Religion.

For the two categories of restaurant and school, we found distinguished results. At this time, it was determined that effective value importance value is greater than 0.1. Considering that most experimental participants of the study are students in their twenties, Extraversion, Agreeableness, and Openness leads to frequent visit to schools. Extraversion, Conscientiousness, and religion also affect the frequent visit to restaurant. To infer why these results came out, we expect that extroverted, enthusiastic and sincere students would have often eaten outside because they would often come to school and stay for a long. Otherwise extroverts are expected to engage in various activities. There would have been many visits to Restaurant in the process. In this context, visits to the beverage store will also have an impact on Extraversion.

Some of the questions in the experiment are that religion has a lot of impact on visiting the beverage store, and the theater and concert hall. In addition to religion, Neuroticism and salary affect theater and concert hall visits.

As mentioned earlier, people with high Neuroticism are stress sensitive and impulsive. Therefore, it is expected that stress will be solved through cultural life such as movie. Also, because cultural life costs, salary will also be effective. For the category pub, Agreeableness, SNS usage frequency, and Openness are effective. It can be inferred that people who get along well with many people are cooperative, have a communal personality, and often have drinking parties.

V. Conclusion and Future Work

In this research, we found that various factors including personality factors effects the selection of visiting place. Especially, factors such as salary, religion, SNS usage were newly distinguished as effective factors for favorite location selection. Several matters must be considered for more precise evaluation. First, most of participants were in their twenties. Table VI shows that several values are skewed. Therefore, these skewed values attenuate the relationship toward visiting places. Once we can get more personal factors including more various age, we guess that more general results with more credible results can be analyzed. Second, we need to adjust location category. For the current categories of location service, two categories contain too many location subcategories. For example, large general retailing and service business contain restaurant and bar but such categorization cannot characterize the locations. This phenomenon leads to inaccurate prediction result. Therefore, ramified categories must be applied in such case so to improve accuracy of analysis. Third, the more data must be collected, especially the location data. Most of participants are not eager to collect their visiting location using SWARM app or does not know the usage of SWARM app. This sort of collection is called as ‘check-in’. Collecting continuous geo-positioning data is passive, meaning that the geo-positioning data is automatically collected, while active check-in is required to use SWARM app. For the next research, we need to give more guidance of SWARM to participants. As well, some participants are too eager to collect check-in data so that even bus stops were checked in. This phenomenon may affect the analysis results. Several location categories are regarded as non-associated with personal factors we designed. For example, in case of hospital, accidents or disease may leads to the visit to hospital rather than personal factors. Therefore, personality, gender, hobby is regardless of such locations. Since most of the participants are students, educational locations are frequently visited. Maybe the job of students will affect the visit to educational locations. Therefore, we need to collect more various data to deduce meaningful result. Our analysis result could be applied to various area requiring visiting places prediction. For example, Location Based Service (LBS) and recommendation system

maybe best application area of our research. With the combination of personal factors and favorite visiting places, the usefulness of LBS and recommendation system can have more value added results and high quality of service.

References

[1] H. Y. Song and E. B. Lee, "An analysis of the relationship between human personality and favored location," AFIN 2015, p. 12, 2015.

[2] H. Y. Song and H. B. Kang, "Analysis of relationship between personality and favorite places with poisson regression analysis," ITM Web of Conferences, vol. 16, p. 02001, 2018. doi: 10.1051/itmconf/20181602001. [Online]. Available: https://doi.org/10.1051%2Fitmconf%2F20181602001

[3] S. Y. Kim and H. Y. Song, "Predicting human location based on human personality," in Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 70–81. [Online]. Available: https://doi.org/10.1007%2F978-3-319-10353-2_7

[4] S. Kim and H. Song, "Determination coefficient analysis between personality and location using regression," in International conference on sciences, engineering and technology innovations. Bali, ICSETI, 2015, pp. 265–274.

[5] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324

[6] P. T. Costa and R. R. McCrae, "Four ways five factors are basic," Personality and Individual Differences, vol. 13, no. 6, pp. 653–665, jun 1992. doi: 10.1016/0191-8869(92)90236-i. [Online]. Available: https://doi.org/10.1016%2F0191-8869%2892%2990236-i

[7] J. Hoseinifar, M. M. Siedkalan, S. R. Zirak, M. Nowrozi, A. Shaker, E. Meamar, and E. Ghaderi, "An investigation of the relation between creativity and five factors of personality in students," Procedia - Social and Behavioral Sciences, vol. 30, pp. 2037–2041, 2011. doi: 10.1016/j.sbspro.2011.10.394. [Online]. Available: https://doi.org/10.1016%2Fj.sbspro.2011.10.394

[8] D. Jani, J.-H. Jang, and Y.-H. Hwang, "Big five factors of personality and tourists' internet search behavior," Asia Pacific Journal of Tourism Research, vol. 19, no. 5, pp. 600–615, 2014. doi: 10.1080/10941665.2013.773922

[9] D. Jani and H. Han, "Personality, social comparison, consumption emotions, satisfaction, and behavioral intentions," International Journal of Contemporary Hospitality Management, vol. 25, no. 7, pp. 970–993, sep 2013. doi: 10.1108/ijchm-10-2012-0183. [Online]. Available: https://doi.org/10.1108%2Fijchm-10-2012-0183

[10] O. P. John, S. Srivastava et al., "The big five trait taxonomy: History, measurement, and theoretical perspectives," Handbook of personality: Theory and research, vol. 2, no. 1999, pp. 102–138, 1999.

[11] Y. Amichai-Hamburger and G. Vinitzky, "Social network use and personality," Computers in Human Behavior, vol. 26, no. 6, pp. 1289–1295, nov 2010. doi: 10.1016/j.chb.2010.03.018. [Online]. Available: https://doi.org/10.1016%2Fj.chb.2010.03.018

[12] M. J. Chorley, R. M. Whitaker, and S. M. Allen, "Personality and location-based social networks," Computers in Human Behavior, vol. 46, pp. 45–56, 2015. doi: 10.1016/j.chb.2014.12.038

[13] Foursquare Labs, Inc., "Swarm app," https://www.swarmapp.com/, 2019.

[14] G. Biau and E. Scornet, "A random forest guided tour," TEST, vol. 25, no. 2, pp. 197–227, apr 2016. doi: 10.1007/s11749-016-0481-7. [Online]. Available: https://doi.org/10.1007%2Fs11749-016-0481-7

[15] M. R. Segal, "Machine learning benchmarks and random forest regression," 2004. [Online]. Available: https://escholarship.org/uc/item/35x3v9t4

[16] International Standard Classification of Occupation, "ISCO," https://www.ilo.org/.