Evaluating Conventional Classification Algorithms predicting Energy Level by Mobile Phone Usage

Marlijn Y. Moonen
m.y.moonen@tilburguniversity.edu
ANR 2017374
U254929

THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN DATA SCIENCE BUSINESS AND GOVERNANCE

DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES

TILBURG UNIVERSITY

Thesis committee:
Drew Hendrickson
Çiçek Güven

Tilburg University
School of Humanities and Digital Sciences
Tilburg, The Netherlands
December 2019

Table of contents

Preface

This thesis, ‘Evaluating Conventional Classification Algorithms predicting Energy Level by Mobile Phone Usage’, would not have been possible without the support and guidance of my thesis supervisors Drew Hendrickson and Giovanni Cassani. The supervision meetings were really helpful due to the possibility to collaborate with fellow students. These meetings encouraged me to keep progressing on my work week by week. Additionally, I would like to sincerely thank my thesis supervisors for the insightful feedback on my work. Moreover, I would like to thank everyone who kept supporting me throughout this period.

Marlijn Moonen,
December, 2019

Evaluating Conventional Classification Algorithms predicting Energy Level by Mobile Phone Usage

Marlijn Moonen

Abstract

The main goal of this study was to evaluate conventional classification algorithms at predicting a mobile phone users' energy level by their phone activities. Before evaluating the classifiers on the data, this study aimed to examine whether the predefined target variable 'energy level', which consisted of six classes, was constructed in a reliable way. Subsequently we found that there were no significant differences between some classes, hence the original six classes were merged into four significantly different classes. As the reordered target variable has four classes, this study is concerned with a multi-class classification problem. Additionally, the target variable is merged into two classes, thereby making it a binary target variable. As such we were able to evaluate the classifiers on both multi-class and binary classification problems and thereafter compare the results. Four classifiers are used on these classification problems, namely k-Nearest Neighbor, Support Vector Machine, Random Forest and Logistic regression. The Random Forest classifier is the best performing classifier (0.521) of the multi-class classification task and the k Nearest Neighbor classifier (0.702) for the binary classification task. Furthermore, this study examined to what extent feature importance affects the models. Most of the models were rarely affected by the removal of features and still managed to perform well.

1. Introduction

Classification problems occur in certainly all diversified business related fields and handling these problems is an important area in machine learning. In his book ‘Machine Learning with R’, Brett Lantz (2013, p.33) expresses the classification principle in a clear way: “things that are alike are likely to have properties that are alike”. The classification algorithm tries to find a relationship between a set of feature variables and a target variable. Classifiers can roughly be distinguished into two categories: multi- or two-class classification problems. Multiclass classification problems have target variables with more than two classes. When a target variable has two classes only, it is a binary classification problem.

A variety of studies have performed classification tasks which have been applied to decision-making scenario’s (Buttenberg, 2015; Jovanovic & Perovic, 2010; Liu & Gegov, 2015; Vardasca, Vaz, & Mendes, 2017). Classification problems are common in fields of customer target marketing, business failure prediction, dept risk management, medical disease diagnosis or social network analysis (Aggarwal, 2014). Previous studies clarified that there is a wide variance in the performance of classification algorithms under contrasting scenarios. The strengths and limitations of different classification algorithms among contrasting scenarios have been studied by Kiang (2003). Herein, the researcher denotes that there is no single method which performs best for all learning tasks. Therefore, the best way of finding a method which fits a specific learning task is to utilize multiple classification algorithms and scrutinize which one works best.

This study examined four conventional classification algorithms which used data on how energetic people were feeling and the way they used their mobile phone during the same period of time. Nowadays, we cannot imagine not having mobile phones anymore. Mobile phone usage has changed enormously during the past few years due to technological opportunities. A few years ago, when you would ask people why they used their mobile phone they would presumably answer that they used their phone for texting and calling only. Today, answering this question is more difficult as we tend to use our mobile phone for almost everything. As technology empowers companies to build applications in all formats, people nowadays have a great variety of applications to choose from and can download their most preferred choice. Hence, research to examine relationships between the phone user’s mental as well as physical state in relation to their mobile phone behavior is of utmost importance in today’s society.

The objective of this study is to apply conventional classification algorithms to the data and examine the relationship between mobile phone usage and its influence on the user’s energy level. Lazy learning is said to be the most intuitive type of learning (Sun, Bo, Luo, & Yang, 2019). Lazy learners are instance based classifiers and are called ‘lazy’ as they store all the training data and only start building a classifier when a new unseen instance needs to be classified (Jadhav & Channe, 204). The lazy learner k Nearest Neighbor (kNN) is investigated in this study.

Probabilistic models are said to be the most fundamental among all data classifiers (Aggarwal, 2014). Both Naive Bayes and logistic regression are probabilistic models, but as Naive Bayes is particularly good for binary classification problems, this study examines the Logistic regression classifier. Probabilistic models are said to achieve comparable performance with more sophisticated classification methods such as Decision Trees and Random Forests (Deng, Sun, Chang, & Han, 2014). Therefore, the Random Forest (RF) classifier is examined in this research as well.

When a classification task should make discriminations based on a weighted sum of all the attributes, a Support Vector Machine (SVM) is one of the most preferable classifiers. Some machine learning algorithms have a tendency to overfit and thus result in high testing errors. This is a consequence of generalization to unseen test instances. SVMs take maximum margins into account which makes overfitting less likely (Wang & Lin, 2014). Hence, also SVMs are examined in this research. More explanation on the classifiers used in this study is given in section 3.2.

1.1 The overarching problem statement and research question

As the conventional classification algorithms are evaluated in this thesis, the following research question acts as a guideline throughout the research:

"To what extent do conventional classification algorithms differ in performance when predicting energy level based on mobile phone usage?"

In order to answer this research question, three sub-questions were formulated. First, it is important to compare different features in the mobile phone data with descriptive statistics. Features of the mobile phone data are analyzed. Additional to analyzing these different features and finding correlations, an important aspect of this study is analyzing how people perceived the 6-point Likert scale in relation to their energy level. Did people indeed observe a difference between 'not energetic at all' and 'very slightly' or is there no significant difference? Results might show that energy level should actually be divided into two classes 'not energetic' and 'energetic' which would make this a binary classification problem. On the other hand it could be a multiclass classification problem if there are significant differences in the six energy levels. This led to the formation of the first sub-question:

Sub-question 1: "*To what extent is energy level a reliable target variable with regard to how it is constructed for this dataset and perceived by the participants?"*

As described earlier, four classification algorithms are examined to see how each model performs in predicting energy level based on mobile phone usage and which features are important in this prediction. This led to the formation of the second sub-question:

Sub-question 2: *"To what extent is there a difference in feature selection between the classification algorithms and how is that affecting the target variable?"*

In addition to building models and analyzing how feature selection is affecting the target variable, it is important to interpret the classification models and evaluate their performance. Hence, feature importance across the models is evaluated. This led to the third sub-question:

Sub-question 3: *"How do feature performances affect the model's behavior?"*

1.2 Structure of the thesis

This research paper is divided into seven parts. The second section provides background information about classification algorithms, theories regarding mobile phone usage and prior studies which are relevant to this research. The third section provides the methodology of this study, wherein more details are given on the data and algorithms which are used. Additionally, this section provides a clear view on the experiments which are carried out for this research and the evaluation metrics used. Section four provides the results obtained by the experiments and then in section five the results are discussed and recommendations for future research are given. Finally, section six is a concluding section with the conclusions of this study.

2. Related work

Comparing several conventional classification algorithms among multiple subjects and datasets has been actively investigated in the last decade. Nevertheless, research on how conventional classification algorithms differ in their performance when applying it to one single dataset is still incipient in literature. This section surveys the literature that addresses mobile phone usage in combination with mood (2.1) as well as studies evaluating different conventional classification algorithms (2.2).

2.1 Mobile phone usage and mood

The literature in the field of mobile phone usage in combination with emotions and mood has a long and rich history. Since mobile phones became an indispensable part of our lives, many researchers have investigated the impact of it (Elwood, 2015; Gunter, 2019; Sugiyama, 2010; Topalli, 2016; Zheng, 2015). A study examined by Stragier, Hendrickson, Vanden Abeele, and De Marez (2019) describes the use of mobile phones as a complex, fragmented and patterned behavior and therefore their study gained insight into these patterns by investigating the structure of smartphone sessions. Their findings show that the key functionality of smartphones is mobile communication. Furthermore, the 'Encyclopedia of Mobile Phone Behavior' by Zheng (2015) is a source for scholarly research on mobile phone usage and contains approximately 120 articles. These articles include investigations on mobile phone behavior and on how mobile phones are revolutionizing the way people learn, work, and interact with one another. A book written by Gunter (2019) exists of various researches on the role of mobile phones in the lives of children and young people. This book provides insights into the effect that mobile phones have on young people and how it is still a challenge to regulate and control this technology.

Additionally, there are studies which in particular focus on emotions and well-being of people in combination with mobile phone usage (Mehrotra, Tsapeli, Hendley, & Musolesi, 2017; Turner, Love, & Howell, 2008). For example, Mehrotra et al. (2017) investigated correlation and causation between user's emotional states and mobile phone interaction. Their study concluded that user's emotions have a causal impact on different aspects of mobile phone interaction and that specific applications showed a causal impact on activeness, happiness and stress level. One of their conclusions stated that people tend to use their phone more when they are active and thus also increases the likelihood of using more apps and number of clicks. Another study, examined by Becker et al. (2016) analyzed how mood levels of healthy clients are affected by behavioral and environmental information through smartphones and how this data contributes to the prediction performance of various statistical models. This study eventually assumed that variables which were more meaningful regarding the target variable could potentially increase the prediction performance.

The interaction between people and mobile phone usage is widely associated with the cognitive context. In recent years, multiple studies tried to discover relationships between users' cognitive context and the way they used their mobile phone (Mehrotra et al., 2017). LiKamWa, Liu, Lane & Zhong (2013) found that the communication history and application usage patterns of participants are good indicators of a user's daily mood average. These indicators predicted a user's daily mood average with an accuracy of 66%. With improved results, they eventually built a service named MoodScope which analyzes mobile phone usage history to predict user's mood. Furthermore, a study by Alvarez et al. (2014) investigated changes in mobile phone usage patterns (focused on application usage) of bipolar patients and how these changes correlated with the patients' self-reported state. Their study showed that there is a strong correlation of patterns of app usage and self-reported mood state. Smartphone data has not only been investigated as the indictors of mood but is additionally used to predict app usage; Srinivasan et al. (2014) introduced a system MobileMiner that discovers frequent co-occurrence patterns and uses these patterns to predict the future applications. Additionally, the prediction of application usage has also been studied by Cao & Lin (2017), who reviewed and discussed literature containing predictive modeling methodologies used for app usage prediction. They found that some common statistics which were used in multiple studies regarding app usage were the duration of the usage sessions and the interaction time. Here, interaction time refers to the time spent on sessions during a fixed period, for example a day. After surveying different studies, the daily interaction time was concluded to be ranging between a few minutes to more than 500 minutes, wherein 90% of the users had an interaction time range between 20 to 100 minutes. Furthermore, they mention that categorical labels of applications such as finance, communication and new can also be correlated to mobile phone statistics like the duration of a session. As mentioned earlier, Stragier et al. (2019) concluded that communication is a key factor of mobile phone usage, which is also confirmed by Cao & Lin as they state that of all the usage sessions, 49.6% were started by the category of communication and this category is therefore the trigger of using a mobile phone.

2.2 Classification algorithms

In addition to the literature focused on mobile phone usage, classification algorithms have been studied and compared to one another across the years as well (Ahn & Kim, 2011; Nieciecka, 2018; Nasa & Suman, 2012; Yusa & Utami, 2017). Nieckiecka (2018) investigated the performance of the classifiers neural networks and decision trees to predict happiness. Only one of the two experiments showed significant difference between the classifiers and therefore it is suggested that further research should investigate if other classifiers to a multiclass target classification problem perform better when using survey data for predictions. As mentioned in the introduction, one of the studies which investigates multiple classification algorithms is a study by Kiang (2003). The strengths and limitations of different classification algorithms among contrasting scenarios are analyzed in his study. The five classification algorithms used in his paper are neural networks, decision trees, discriminant analysis, logistic regression and k-Nearest Neighbor. Kiang investigates the strengths and limitations of these classification algorithms by systematically adjusting the data characteristics. In this way, imperfections such as nonlinearity, multicollinearity and unequal covariance are introduced. The altered data characteristics proof to have a considerably impact on the performance of the classification algorithms. For example, the degree of multicollinearity has an effect on the logistic classifier (clarification of multicollinearity for this study is given in section 3.1.4). Another conclusion of the study of Kiang is that there is no single classification method that outperforms all methods in all problem situations.

Two types of classification tasks exist: binary and multi-class (Tharwat, 2018). Therefore, the literature focusing on classification algorithms varies widely. Fernández-Delgado et al. (2014) compared 179 different implementations of seventeen classification algorithm families on 121 public datasets. The classification problems investigated throughout this study were binary as wel as multi-class problems. The classifier which is most likely to be the best classifier is proofed to be the random forest. The random forest classifiers achieved the best results on average in comparison to the other classifiers. This study has been replicated by Wainer (2016) wherein a few ‘imperfections’ of Fernández-Delgado his study were addressed and adjusted. The replicated study transformed all the datasets in such a way, that all the classification tasks were binary. Furthermore, the number of classifiers were downsized to see how different families of algorithms perform in comparison to one another, as this was not clear from the original study. In addition to the standard null-hypothesis significance test (NHST) to examine if the algorithms were significantly different from one another, this study also used a Bayesian analysis. Wainer concluded that the random forest, gradient boosting machines, and RBF SVM are the classifiers which are most likely to have the highest accuracy.

In addition to these binary classification studies there are numerous studies which have investigated multi-class classification problems as well (Aly, 2005; Lango, 2019; Mosley, 2013; Student, Pieter, & Fujarewicz, 2016). One of those studies (Aly, 2005) solved multiclass classification problems as they were actually binary. A multiclass classification problem was unfold by extending the binary classification technique for various classification algorithms. The algorithms investigated in this study are neural networks, decision trees, k-Nearest Neighbor, Naive Bayes, and support vector machines. Furthermore, Bi and Zhang (2018) compared classification algorithms for multi-class imbalanced data and thereafter introduced a new multi-class classification algorithm, named the Diversified Error correcting Output Code (DECOC). They applied their new classification algorithm on nineteen public datasets and compared the performance of the DECOC with seventeen classification algorithms on predefined accuracy measures, with as result that the accuracy of the DECOC was significantly higher than the other classifiers.

Classification algorithms should be evaluated on their performance and there are distinctive approaches in doing so. A study conducted by Tharwat (2018) applied several assessment methods to both binary and multi-class classification algorithms. The extent to which a dataset is either balanced or imbalanced influences the assessment measures, and Tharwat presents the dissimilarity for each metric.

Furthermore, this study gives illustrative examples of how to calculate those measures on both binary and multi-class classification problems. Additionally, a description is given on how accuracy, error rate, sensitivity, specificity, likelihood ratio and the Yourden's index metrics can be utilized to evaluate a classifiers performance. Moreover, a study conducted by Demsar (2006) investigated statistical comparisons of classification algorithms over multiple datasets. They examined the Wilcoxon signed ranks test to compare the performance of two different classifiers and the Friedman test with corresponding post-hoc tests for comparisons of multiple classifiers over multiple datasets. The previous mentioned study by Wainer (2016) followed the full Demsar procedure of comparing classification algorithms completely in addition to using the Bayesian ANOVA to verify statistically significant differences between classifiers. Even though this study was replicating the original study of Fernández-Delgado et al. (2014), they differ in the methods they use for comparing the performance of the classifiers as the original study developed a paired t-test to compare the accuracies of the different classifiers and did not follow the Demsar procedure. The evaluation metrics used for this thesis will be described and explained briefly in section 3.4.

3. Methodology

This section first provides a description of the data (**3.1**), where the data is explained as well as the preprocessing steps and a description of the final dataset used for this study. The second part specifies the conventional classification algorithms (**3.2**) and the third part outlines the experimental procedure and implementation (**3.3**) to train the classifiers. The fourth part provides the evaluation criteria (**3.4**) for the models and lastly the baseline is described (**3.5**).

3.1 Dataset Description and Pre-Processing Phase

Two datasets were used in this study to investigate the relationship between mood and mobile phone usage. Tilburg University is the provider of the datasets. The datasets contain data obtained from students between seventeen and twenty years old. The mobile phone data is gathered during a time period of one month, wherein mobile phone behavior was recorded throughout an application installed on the mobile phones of the participants prior to the start of the research. Furthermore, data was gained throughout a questionnaire wherein the participants were asked to fill in a survey regarding their mood four times a day.

3.1.2 Mood Dataset

The mood dataset consists of data regarding the conditions the participant was in. In total 146 participants were recorded in this dataset. The participants were asked to fill in a questionnaire four times a day at different time slots. The questionnaire included questions with reference to how participants were feeling on a 6-point Likert scale from ‘not *m* at all’ to ‘very *m*’ where *m* is one of the twelve diverse mood-types. For this study, the only mood-type taken into account is the degree to which people felt energetic. The six levels of feeling energetic are; (1) ‘Not at all’, (2) ‘Very slightly’, (3) ‘A little’, (4) ‘Moderately’, (5) ‘Quite a bit’ and (6) ‘Extremely’. In section 3.3.1 more clarification will be given on how these levels are used throughout this study.

Participants were asked to fill in a questionnaire four times a day, nevertheless some participants did not manage to fill in the survey at all four timeslots every day. From the first to the fourth timeslot the observations have a descending pattern. Therefore, this study contrived a feature which resembles the average energy level of participants per day. Due to missing values in the data, as a result of participants not consequently filling in their survey four times each day, the average is a suitable metric.

The original mood dataset contains 16016 observations with 35 features, nevertheless fourteen observations turned out to be duplicate and are deleted as this is observed to be a technical mistake and only unique observations are essential for this study.

3.1.3 Phone Use Dataset

The Phone Use Dataset consists of data regarding how the participants used their mobile phone. There are 124 people who participated in the experiment. This dataset contains 586,792 observations of 9 different features. The dataset contains +/- 20,000 duplicate rows and these rows are deleted as these are not relevant for this study.

As the target variable 'energy level' is constructed as a daily average (as mentioned in section 3.1.2) the features in the phone use dataset are all constructed as daily measurements as well. An explanation of the features derived from the phone use dataset is given in table 1.

Feature	Description	Type
Apps_per_day	Amount of applications opened per user per day	Numeric
Mean_app_time_a_day	The average time spent on a single application per user per day	Numeric
Sessions_per_day	The amount of sessions per user per day	Numeric
Mean_time_session_per_day	The average time spent on a single session per user per day	Numeric
Mean_apps_per_session	The average amount of apps used in one session per user per day	Numeric
Most_used_categ_day	The most used application per user per day	Factor
Unique_apps_in_session_per_day	The amount of unique applications in one session per user per day	Numeric

Table 1: An overview of the features derived from the phone use dataset.

3.1.4 Description of the final dataset

In the interest of this study, the two above mentioned datasets were merged into one final dataset. The datasets were merged by 'date' so that every participant has one single row for each day they participated in the experiment. Due to some technical issues there were some major outliers. For some participants the duration of a single session lasted multiple days which led to unreliable statistics for the constructed features. As it is essentially impossible to have single sessions which lasts multiple days and these outliers led to unreliable statistics for different constructed features, these participants were deleted from the dataset. Eventually, after the outliers were removed and the data was merged in such a way that all users had a single row for each day they participated in the experiment, the final dataset contained 2,202 observations of 116 participants. The final dataset consists of the target variable ‘energy level’ as well as the seven predictors which are mentioned in table 1 in section 3.1.3.

High correlations between the predictor variables, also called multicollinearity, might affect the model. Therefore, a correlation plot was utilized to test the multicollinearity. The correlation plot shows that there is some intercorrelation between the features in the final dataset. The two features which have a positive correlation are ‘apps used per day’ and ‘unique apps in a session per day’. These features were expected to correlate as it is reasonable to state that when people use a lot of applications per day it is comprehensible that the amount of unique applications opened consequently increases. Furthermore, instant messaging and social networking are negatively correlated. As these are the only correlations found and they do not correlate with any other predictors the features are not adjusted.

As the classifiers kNN and SVM are not that adaptive to categorical predictor variables, the predictor variable ‘most used category’ is transformed into dummy variables. This feature originally had 70 different categories of applications which were used during the day. As mentioned before, only the most used categories per day were taken into account and the final dataset contained 22 unique categories which belonged to the ‘most used category per day’. As eight of the 22 categories were significantly more popular than the others, the other fourteen categories are all re-categorized into one combined category named ‘else’. This was done to reduce the number of dimensions in the data.

Furthermore, kNN and SVM classifiers are sensitive to the measurement scale of the input variables (Lantz, 2015), and as the input variables in this study have huge differences in ranges of the scales the variables will be normalized and rescaled. The min-max normalization method is applied to normalize the data as this provides linear transformation and the relationschip among orginal values is retained (Patro & Sahu, 2015).

3.2 Algorithms

This section provides a description of the algorithms used in this study. First, classification will be explained according to the literature (3.2.1). The rest of the section provides information about the four classification methods applied in this study; K-Nearest Neighbor (3.2.2), Support Vector Machine (3.2.3), Random Forest (3.3.4) and Logistic Regression (3.3.5).

3.2.1 Classification algorithms

Classification algorithms differ in performances as real-world problems do most of the time not fulfill all assumptions of all methods. Therefore, one approach of handling classification problems is not just selecting one method but instead implement various appropriate methods and selecting the method that provides the best solution (Kiang, 2003). Classification algorithms are built to find a model which best fits the relationship between the predictor variables and target variable, which is also referred to as composing a learning algorithm. Moreover, the model should predict the unseen class labels from the target variable as correctly as possible and therefore the key objective of a classification algorithm is to have a good generalization capability (Tan, Steinbach, Karpatne & Kumar, 2014).

3.2.2 K-Nearest Neighbor

k-Nearest Neighbors (kNN) is a non-parametric supervised classification algorithm. The idea of kNN is finding a group of *k* samples that are closest to unknown samples, based on certain distance measures (Akbulut, Sengur, Guo & Smarandache, 2017). The class to which the unknown samples belong is determined by calculating the average of the response variables, also referred to as the class attributes of the *k* nearest neighbor. The value of *k* is important in this classification algorithm as this is the key tuning parameter of kNN (Noi & Kappas, 2018).

3.2.3 Support Vector Machine

The support vector machine (SVM) is said to be one of the most robust and successful classification algorithm amidst all others. The idea behind the SVM is to encounter a hyperplane which separates the features into different domains. The goal of the hyperplane is to divide the data into as homogenous domains as feasible (Lantz, 2013). A SVM uses this hyperplane to find the maximum distance between the two nearest datapoints (Nazir, Qi & Silvestrov, 2016). Figure 1(a) represents the multiple options the hyperplane has in separating the data. The algorithm searches for the Maximum Margin Hyperplane (MMH) which indicates the maximum distance of the nearest support vectors as shown in figure 1(b). Each class is represented by at least one support vector.

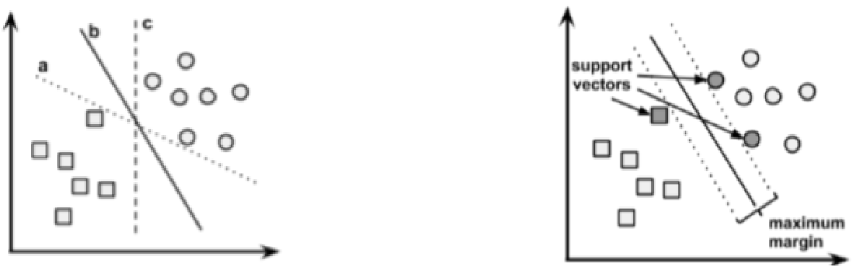


Figure 1 a) hyperplane options of separating the data | b) the Maximum Margin Hyperplane (Lantz, 2013).

This is an example of linear separable data. When the data is not linear separable, the support vector machine algorithm uses ‘slack variables’. Slack variables are data points which fall into the incorrect side of the margin hyperplane. As it is important to have as minimum slack variables as possible, a cost value is applied to the slack variables. The cost value is a parameter, denoted as *C*. The higher the cost value, the harder it is to achieve a 100 percent perfect seperation between datapoints (Lantz, 2013).

3.2.4 Random forest

The random forest (RF) algorithm is a non-parametric ensemble learning classifier. As this classifier is non-parametric, the linearity of the data does not need to be taken into account (Gupte et al. 2014). It aggregates a large number of decision trees and reduces the variance compared to a single decision tree. The decision tree algorithm divides the data into smaller portions to identify patterns and features which are used for predictions. The data is partitioned into smaller subsets based on feature values. The classifier starts by choosing the feature value which has most predictive power towards the target class. Subsequently, this method continues until there are no remaining feature values to divide among classes. However, it is also possible that the algorithm stops when the tree has reached a predefined size or when all the nodes have the same class and cannot be divided anymore. The random forest model combines all results it obtained from the trees to get the prediction accuracy. The three parameters of the RF classifier are the node size, the number of trees in the forest and the number of predictors used to decide om which nodes to split (Lakshmanaprabu, 2019). The influence of a variable on the classification in the random forest classifier can be indicated by either the Mean Decrease Impurity (MDI) or the Mean Decrease Accuracy(MDA).

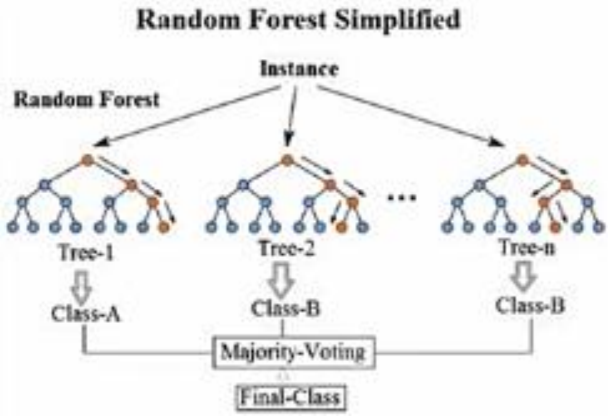


Figure 2 Random Forest classifier (Edureka, 2017).

3.2.5 Logistic Regression

The logistic regression algorithm analyzes the relationship between the predictor variables and the target variable. It is a supervised learning algorithm. As this study has both two-class as well as multi-class classification tasks which are explained more specifically in section **3.3.1** of the experimental setup, both binary logistic regression and ordinal logistic regression are applied to the data. The classifier is mostly used for binary target variables, but nevertheless also suitable for multi-class classification tasks. The logistic regression model is classifying inputs to be in one class or the other by the use of logarithmic odds (Kranse, Roobol, & Schröder, 2008).

3.3 Experimental Procedure

The main goal of this study is to answer whether predictions can be made about how energetic people are feeling using data regarding mobile phone behavior. This section describes the procedures which are taken to answer the sub-questions as feasible as possible.

3.3.1 Experiment 1

This study first investigated to what extent the six levels of the mood type 'energetic' are reliable in the way they are constructed in this dataset. The first experiment in this study aims to answer the following sub-question:

"*To what extent is energy level a reliable target variable with regard to how it is constructed for this dataset and perceived by the participants?"*

There are multiple studies discussing the best way to analyze and treat Likert scale responses (Knapp, 1990; Chang, 1994; Dawes, 2008; Clarifo & Perla, 2008; Harpe, 2015; Xu & Leung, 2018). Likert scale responses are very popular item scoring schemes to quantify people's opinions and there are multiple options to conduct such an item scale (Bishop & Herron, 2015). Some of the Likert scale response categories have a 'neutral' option in the middle. When having this 'neutral' option, some would argue that the response options are balanced as there as many response options to the left side of 'neutral' as there are options on the right of 'neutral' (Dawes, 2008). Nevertheless, if a response set does not contain a 'neutral' option such as in this study, there possibility exists that there is less distance between 'very slightly' and 'a little' than between 'a little' and 'moderately'. Knapp (1990) even suggest that in this case some would argue to reorder the two middle terms. Furthermore, the study conducted by Chang (1994) found that 6-point scales led to greater reduction of reliability than 4-point scales. The goal of the first experiment is to find out how different classes of 'energy level' were perceived by the participants and if some of these classes can be combined.

The dataset is split in 5 different datasets, where each dataset only contained two of the classes. The first dataset consisted of the classes 'not at all' and 'very slighty', the second dataset consisted of the classes 'very slightly' and 'a little', the third dataset consisted of the classes 'a little' and 'moderately', the fourth dataset consisted of the classes 'moderately and 'quite a bit' and the fifth dataset consisted of the classes 'quite a bit and 'extremely'. The logistic regression classifier was applied to each dataset to see if the predictor variables were significant for that specific dataset. If the predictor variables were not significant, the two classes were combined and if they were significant these classes were separated.

3.3.2 Experimental setup for investigating algorithms

Four classification algorithms are applied to the data, namely the kNN, SVM, RF and LGR. This section provides the experimental setup of each of the classifiers. All classifiers and both datasets are trained using 10-fold cross validation, which will be further explained in section 3.4. The implementation information of the classification algorithms are listed below.

- knn, k-Nearest neighbors classifier. Package: class (Venables and Ripley, 2002)
- glm. Logistic regression. Package: stats (McCullagh & Nelder, 1989)
- polr. Ordinal logistic regression. Package: MASS (Venables and Ripley, 2002)
- random forest. Package: randomForest (Breiman, 2001)
- svmRadial. A SVM with RBF kernel. Package: e1071 (Meyer et al., 2014)
- varImp. Feature importance. Package: caret
- gmulti. Feature importance logistic regression. Package: gmulti (Calcagno & Mazancourt, 2010).

This study uses nonlinear kernal functions for the SVM to transform the data into a high-dimensional feature space. These kernal functions make the data more seperable as the data is not perfectly linearly seperable. The svmRadial is used for this study in combination with the radial basis function kernal (RBF).

3.3.3 Experimental setup for feature importance

The feature importance is analyzed for each of the classifiers. The Logistic regression classifier has a package called ‘gmulti’ which is useful for finding the optimal model with the most important features. Gmulti is a package for automated model selection. The algorithm starts from a full model and step by step removes features which do not significantly reduce the fit of the model. The output of the gmulti analysis consists of the best performing models with their support (Calcagno & Mazancourt, 2010). For example a value of 80 means that in 80 of the 100 fitted models this feature is seen as an important predictor and is thus part of the final model.

To assess the importance of the predictors in the other classifiers, the varImp function of the caret package is used. These variable importance tables clarify the strength of the effect associated with each predictor in the model. The variable importance function will give each predictor a separate variable importance for each class. For classification problems, the function conducts a ROC curve analysis on each predictor in the model. Thereafter, the area under the curve (AUC) is computed and this area is used as the measure of the variable importance (Kuhn, 2007). The importance scores of the classifiers are rescaled to a value between 0 and 100 so it is comparable to one another.

The features that have no strong effect on the model are removed to see if the classifiers perform better. Additionally, the features that are most important are also removed to see what impact removing these features have on the model. The threshold which is set to tell if a feature is ‘important’ is set to a value of above 80 and the threshold for ‘less important’ is set to a value lower than 40.

3.4 Evaluation criteria

Tharwat (2018) explains that accuracy is one of the most commonly used evaluation metrics for classifiers. Therefore, the models are all evaluated based on their accuracy. Accuracy comes in combination with misclassification rate, also referred to as Error Rate (ERR). As we are dealing with a multi-class classification problem the weighted accuracy is used, this is because the weighted accuracy gives equal contribution to the predictive performance for all the classes in the target variable and is not dependent of the number of observations of each class (Döring, 2018).

To select the optimal parameters for the classifiers, each classifier is trained using 10-fold cross-validation. 10-fold cross validation involves splitting the data into ten folds, wherin nine folds are used to train the model and the remaining fold to validate the model. This is repeated ten times with the result that each fold has been used as validation fold once. The classifiers were thereafter evaluated on overall accuracy as well as weighted accuracy and F1 macro score. The F1 macro score was chosen as evaluation metric as a high F1 macro score indicates that the classifier performs well on each individual class and is therefore suitable if the classes are imbalanced (Döring, 2018).

3.5 Baseline

A baseline to compare the performance of classification models is used as reference point to compare the classification models with. The baseline model does not take any predictor variables into account. As the goal of this study is to predict the class of the target variable ‘energy level’, the baseline of this study is the most common class divided by the total of all classes.

4. Results

In order to answer the main reseach question of this study this section provides the results of the three sub-questions. The main research question is:

"To what extent do conventional classification algorithms differ in performance when predicting energy level based on mobile phone usage?"

4.1 Sub-question 1

The first sub-question "*To what extent is energy level a reliable target variable with regard to how it is constructed for this dataset and perceived by the participants?"* had as goal to investigate whether the target variable 'energy level' is constructed in a reliable way.

As discussed in section 3.3.1, the logistic regression classifier is applied to see if there are differences between the classes of the target variable 'energy level'. The results show that when modelling the logistic regression on the first dataset, where the target variable has the classes 'not at all' and 'very slightly', none of the predictors are significantly different for the two classes (significance is estimated with a value of $p < 0.5$).

In contrast to the first two classes, the classes 'very slightly' and 'a little' do have significant differences in the predictor variables. Here, the predictors 'sessions per day', 'applications per day' and 'average applications opened per session' are significantly different for 'very slightly' and 'a little'. Additionally, the classes 'a little' and 'moderately' of the third dataset also show significant differences. Of the 'most used categories' only the category 'dialer' is not significantly different between the classes. Furthermore, the features 'sessions per day' and 'applications per day' are also significantly different for 'a little' and 'moderately' .

The classes 'moderately' and 'quite a bit' are also significantly different from one another based on the data. Here, the features 'sessions per day', 'unique applications used in a session per day' and the most used category 'dialer' have significant differences between the two classes. The classes 'quite a bit' and 'extremely' do not have any significant differences between the features in the classes. Table 2 gives an overview of the classes which were merged as they did not differ from one another.

Original classes	*New class*
Not at all & very slightly	Low
A little	Low-med
Moderately	High-med
Quite a bit & extremely	High

Table 2. Merging the classes into four classes.

The classes which are not significant different from each other based on the predictor variables are combined and the final dataset which is used for the remaining questions of this study was therefore a dataset with a 4-class target variable, and will ahead be referred to as the ‘4-class dataset’. Additionally, the binary dataset is also used as it is interesting to see the differences between classifiers on multi-class classification problems as well as on binary classification problems. The binary dataset is constructed by separating the four classes of the 4-class dataset, wherein the classes ‘low’ and ‘low-medium’ are rescaled into ‘low’, and the classes ‘high-med’ and ‘high’ are rescaled as ‘high’. Figure 3 shows the distribution of the two datasets and table 3 shows the accuracy as well as the weighted accuracies and the F1 macro score of the 10-fold cross validation for each of the classifiers on the binary as well as the 4-class dataset.

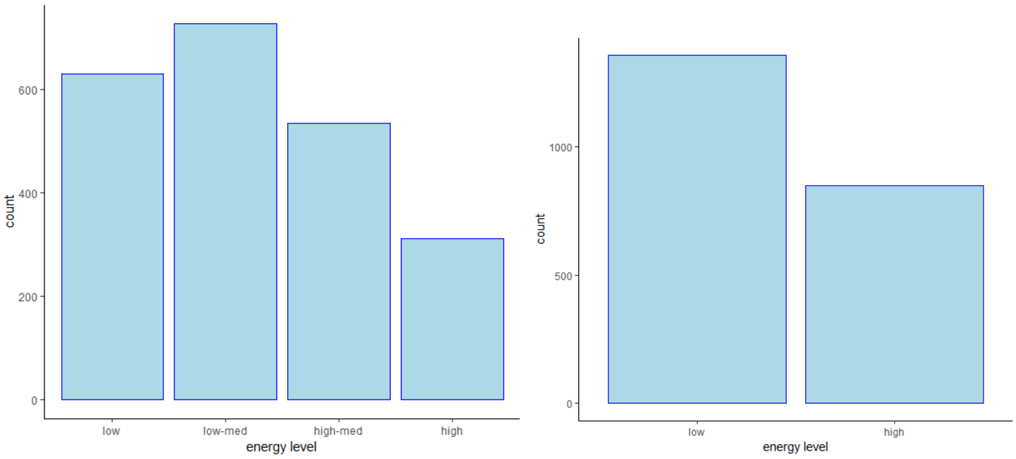


Figure 3 Distribution histograms of the 4-class dataset and the binary dataset.

Target variable	Classifier	Accuracy	Weighted accuracy	F1 macro score	Kappa	Parameters
Binary Baseline = 0.59	kNN	0.702	0.658	0.47	0.16	k = 7
	SVM	0.639	0.540	0.47	0.18	C = 0.5, kernel = ‘radial’
	RF	0.639	0.591	0.47	0.08	mtry = 2
	LGR	0.624	0.522	0.55	0.10	-
Four-class Baseline = 0.33	kNN	0.516	0.483	0.49	0.09	k = 7
	SVM	0.437	0.371	0.43	0.14	C = 1, kernel = ‘radial’,
	RF	0.521	0.446	0.55	0.15	mtry = 2
	LGR	0.351	0.277	0.32	0.12	-

Table 3. Accuracy, weighted accuracy, F1 macro score, Kappa and parameter selection of the classifiers using 10-fold cross validation on the binary and 4-class datasets.

For the 4-class dataset it is interesting to see how different classes have different influence on the predictors. For example figure 4 shows the distribution of the predictor ‘most used category per day’ for the four different classes. This distribution gives insight into which categories are more popular to specific classes. It is notable that dating applications are only part of the ‘most used category of the day’ when energy level is ‘low’. Furthermore, when looking at the category ‘dialer’ we can see that of all classes, dialer is mostly common as ‘most used category of the day’ when people their energy level is high. The category Instant Messaging is not varying widely. Nevertheless, when energy level is ‘low’ the users are using Instant Messaging less than when their energy level is ‘high’ and this is exactly opposite to Social Networking.

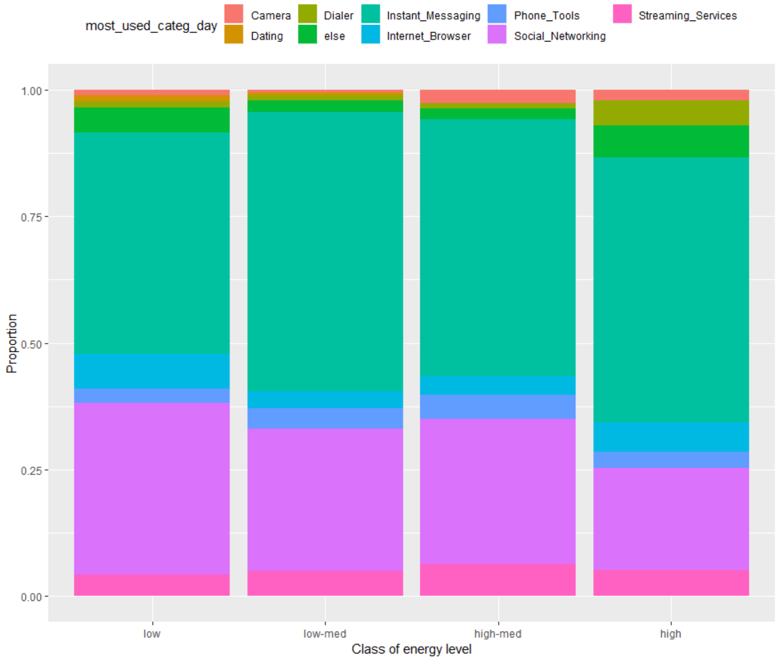


Figure 4. distribution of different categories in the 4-class dataset.

4.2 Sub-questions 2 and 3

The second research question: “*To what extent is there a difference in feature selection between the classification algorithms and how is that affecting the target variable?”*, had as goal to investigate the extent to which certain specific features are important and if there is a difference in which features are important to the classifiers. And the third sub-question *“How do feature performances affect the model’s behavior?”* had as goal to see how certain features affect the classifier’s performance. This section discusses the importance of the features for each classifiers and the effect of removing the important and less important features. The threshold value which is used to determine if a feature is seen as important or less important is explained in section 3.3.3. The results of these questions are divided into two sections. The first section provides the results of the binary dataset (4.2.1) and the second section provides the results for the 4-class dataset (4.2.2)

4.2.1 Binary dataset

The feature which is an important predictor for all four classifiers of the binary dataset is ‘applications used per day’ which indicates that the amount of applications opened per day is one of the predictors which has most effect in the models when predicting energy level. This section provides information about the importance of features in each model on the binary dataset and the effect of removing the important and less important features. At the end a table is provided to summarize the performance results of the classifiers.

The kNN classifier shows that there are six predictors which have more effect on the model than the others. These features are: ‘amount of applications used per day’, ‘average time spent on a session per day’, ‘average amount of applications per session’, ‘most used category Social Networking’, ‘amount of unique applications in a session’ and ‘average time spent on a single application’ First, these predictors were removed from the model to see how the model performed without those features. Results show that the overall accuracy as well as the weighted accuracy of the model decline slightly (1%). Secondly, the least important features were removed, and therefore the six most important features where the only predictors in the model. Results show that the overall and weighted accuracy after removing the least important predictors decrease even more than after removing the most important predictors (3%). The order of feature importance before and after removing the important and less important predictors is also shown in Appendix C.

The predictors which have most effect on the model of the SVM classifier are the features ‘applications per day’ and ‘most used category Social Networking’. After removing those two predictors, the overall accuracy as well as the weighted accuracy both increase by approximately 2%. When following the threshold values mentioned in section 3.3.3 ten of the predictors are removed as their effect on the model is below 40%. As a result, when the less important features are removed the accuracy decreases. The SVM classifier predicting energy level is performing better when more predictors are added to the model, even though these predictors do not have a huge effect on the model. The order of feature importance before and after removing the important and less important predictors is shown in Appendix C

The random forest classifier shows that, in contrast to the other classifiers, the feature ‘most used category’ is not of great importance in the model. Therefore this feature was removed to see if the performance of the model improved. Results show that when removing these features, the overall and weighted accuracy increase enormously. When the accuracy was originally 63.9%, the accuracy after removing the less important features is 99.4%. The order of feature importance after removing the ‘most used category’ feature is shown in Appendix C.

For the logistic regression classifier, the gmulti algorithm is used to get insights into the feature importance of the model. The gmulti algorithm wants to have a as low AIC as possible and finds the best performing model. More detailed explanation of this package is given in section 3.3.3. The algorithm shows that after 70 repetitions of throwing out features the best performing model is said to be predicting ‘energy level’ with (1) a selection of the most used categories per day, (2) amount of sessions per day, and (3) amount of applications per day. The overall and weigted accuracy for the logistic regression model is decreasing when removing important as well as unimportant predictors.

Nevertheless, removing the most important predictors does have a greater effect on the model than removing the less important predictors. The order of feature importance before and after removing the important and less important predictors is shown in Appendix C.

Classifier	Accuracy before removing features	Accuracy after removing important features	Accuracy after removing least important features	Weighted accuracy before removing features	Weighted accuracy after removing important features	Weighted accuracy after removing least important features
kNN	0.702	0.690	0.672	0.658	0.643	0.617
SVM	0.639	0.655	0.625	0.540	0.565	0.505
RF	0.639	0.642	0.994	0.591	0.538	0.992
LGR	0.624	0.615	0.623	0.522	0.500	0.519

Table 4. *Accuracy and weighted accuracy of binary dataset before and after removing important and less important features.*

4.2.2 Four-class dataset

This section provides information about the importance of features in each model on the 4-class dataset and the effect of removing the important and less important features. At the end a table is provided to summarize the performance results of the classifiers. Additionally, an overview of precision and recall measures per classifier before and after removing the important and less important features are given in Appendix C.

In addition to the feature importance of the kNN model on the binary dataset, the kNN model for the 4-class dataset has two additional features which are important in the model. These features are ‘sessions per day’ and ‘most used category Instant Messaging’. After removing the important features the accuracy of the model decreases slightly with 0.6% and after removing the less important features the accuracy increases slightly with 0.2%. The order of feature importance before and after removing the important and less important predictors in the kNN model is shown in Appendix D.

The feature importance of the SVM classifier on the 4-class dataset is slightly different from the feature importance of the SVM on the binary dataset. Here, the least important features of the classifier are all ‘most used categories’ except for the categories Instant Messaging and Social Networking. As in almost all models, the features regarding the amount of applications and sessions per day are important predictors again. The order of feature importance before and after removing the important and less important predictors is shown in Appendix D.

The model of the RF classifier shows the same feature importance results as with the binary dataset and the accuracy also increases enormously after removing the less important features (99.3%). The 'most used categories' are less important in the model. Furthermore, it is interesting to see that after removing the 'most used category' features, the feature 'amount of unique apps in a session per day' is now seen as the least important predictor and is not even used in the model. The order of feature importance before and after removing the important and less important predictors is shown in Appendix D.

The gmulti algorithm shows the most important features of the logistic regression on the four-class dataset. After 90 repetitions of finding the best model, the best model for the ordinal logistic regression is said to be the model with the features 'most used category a day', 'sessions per day', and 'apps per day'. The accuracy and weighted accuracy after removing the important and less important predictors decrease slightly. The predictor 'average time spent on a session per day' does not have a huge effect on the original model, nevertheless after removing the three most important features the effect of this predictor increases. This shows that when important features are removed from the logistic regression model, it finds new important features so that the accuracy of the model does not decrease enormously. The order of feature importance before and after removing the important and less important predictors is shown in Appendix D.

Classifier	Accuracy before removing features	Accuracy after removing important features	Accuracy after removing less important features	Weighted accuracy before removing features	Weighted accuracy after removing important features	Weighted accuracy after removing less important features
kNN	0.516	0.510	0.518	0.483	0.471	0.478
SVM	0.437	0.439	0.414	0.371	0.373	0.345
RF	0.521	0.439	0.993	0.446	0.364	0.991
LGR	0.351	0.337	0.351	0.277	0.260	0.278

Table 5. *Accuracy and weighted accuracy of 4-class dataset before and after removing important and less important features.*

5. Discussion

General discussion

This study investigated whether mobile phone behavior could make predictions about how energetic people were feeling. Conventional classification models were developed which used mobile phone data to predict the class of ‘energy level’. The results of the first experiment provided more insight in how people perceived the 6-point Likert scale. The way to analyze Likert scale responses has been a subject of discussion in the literature (Knapp, 1990; Chang, 1994; Dawes, 2008; Clarifo & Perla, 2008; Harpe, 2015; Xu & Leung, 2018). Knapp (1990) already suggested that reordering Likert scale responses might be necessary when the possibility exists that there is for example less distance between the first two classes than between the second two classes. The first experiment of this study showed that there were no significant differences between the classes ‘not at all’ and ‘very slightly’ and between the classes ‘quite a bit’ and ‘extremely’. With these results, the original six classes of the target variable were reduced into four classes as the classes which did not have any significant differences between one another were merged together. This finding is in accordance with the study of Chang (1994) who found that 6-point scales led to greater reduction of reliability than 4-point scales. Classification problems can either be multi-class or binary and as it is interesting to see how classification models perform differently for binary and multi-class classification problems with respect to feature importance, the target variable was reordered into a binary variable with the values ‘low energy level’ and ‘high energy level’.

Previous studies demonstrated that mobile phone usage was a powerful prediction tool for user’s emotions (Mehrotra, 2017; Likamwa, Lia, Lane & Zhong, 2013). Therefore, this study investigated whether mobile phone usage data could predict the extent to which someone felt energetic. Conventional classification models were built which aimed at using mobile phone usage data in order to predict energy level. The best performing model for the reordered 4-class dataset was the Random Forest, with a performance that was 19.1% better compared to the baseline. This finding is in accordance with the studies of Fernández-Delgado et al. (2014) and Wainer (2016), where Random Forest was concluded to most likely be the best classifier among the others based on the highest accuracy. The Random Forest is followed by the k-Nearest Neighbor, the Support Vector Machine and lastly the Logistic regression. All models performed better than the baseline. The F1 macro score is not that high for the models in the 4-class dataset. Again, the Random Forest has the highest score with a F1 macro score of 0.55. The higher the F1 macro score, the better a classifier performs on each individual class of the target variable.

The results for the binary dataset were slightly different. Even though all models performed better than the baseline, the k-Nearest Neighbor classifier is performing better than the Random Forest. The classifiers which perform slightly worse in comparison with the k-Nearest Neighbor are the Random Forest and Support Vector Machine, which have the same overall accuracy but the weighted accuracy is higher for the Random Forest model. Again, the logistic regression has the lowest performance. The F1 macro scores for the models in the binary dataset did not differ, they were all below 0.50 so they did not outperform the F1 macro score of the Random Forest classifier in the 4-class dataset.

Furthermore, feature importance was investigated for the different classifiers. As Mehrotra et al. (2017) mentioned, specific applications have impact on activeness level and furthermore the likelihood of using more apps increases when people feel active. As we could see in figure 4 the most used categories were Social Networking and Instant Messaging. The fact that these two categories were the most used categories is in accordance with the study of Cao & Lin (2017) which stated that most mobile phone sessions start by the category of communication. The feature importance analysis additionally showed that these features were one of the most important predictors in the model together with 'applications used per day' and 'amount of sessions per day'. After removing the most important features from the model, overall the accuracy for the classifiers decreased. The feature importance analysis also showed which features were less important to the model. When removing these features the accuracy of the Support Vector Machine decreased with 2.3%, whereas the accuracy of all other classifiers increased. Especially the accuracy of the Random Forest classifier increased enormously. When looking at the results of the 4-class dataset the accuracy of the Random Forest before removing any features was 52.1% and after removing the less important features the accuracy increased to 99.3%.

Limitations and further research

There were a few limitations in this study. First, the target variable 'energy level' as well as the predictors are measured per day. Nevertheless, it could be possible that the change in energy level and mobile phone usage during these days lead to better predictions. Therefore, a better understanding in different measurement periods of those variables would provide more certainty in interpreting the results. For example, the change in energy level in the morning and phone use in the morning against energy level in the evening and phone use in the evening might lead to other performances.

In this study the dependency structure of the data is not taken into account. For each participant, there are repeated measurements over time. Due to missing values in the data, as a result of participants not consequently filling in their survey four times each day, the average per day is taken as metric. Nevertheless, multi-level models for example are able to take the difference of repeated measurements between participants into account and would be suggested for future research. These models use fixed effects like gender, which cannot be divided into more categories than male and female, but also use random effects like time and location. With these random effects one is able to correct the dependency structure. One of the models which can be used is the generalized linear mixed models (GLMMs).

Four classification algorithms have been evaluated on how they performed when predicting energy level by mobile phone usage. The limited time which was available for this research made it impossible to investigate whether changes in the cross validation method or changes in the parameters of the classifiers led to other performances. Therefore, for future research it would be interesting to examine the effects of changing the parameters of the classifiers. Additionally, different cross validation methods could be investigated such as the leave-one-out cross-validation (LOOCV).

6. Conclusion

The goal of this study was to answer the following research question: *"To what extent do conventional classification algorithms differ in performance when predicting energy level based on mobile phone usage?"*. Three sub-questions were formulated in order to answer this main research question:

Sub-question 1: *"To what extent is energy level a reliable target variable with regard to how it is constructed for this dataset and perceived by the participants?"*

Sub-question 2: *"To what extent is there a difference in feature selection between the classification algorithms and how is that affecting the target variable?"*

Sub-question 3: *"How do feature performances affect the model's behavior?"*

Firstly, this study investigated the extent to which energy level was a reliable target variable with regard to how it is constructed. It was found that the classes of energy level were not constructed in a reliable way for this research. When using mobile phone usage data in order to predict the class of energy level, the classes 'not at all' and 'very slightly' and the classes 'quite a bit' and 'extremely' were not significantly different from one another and it was therefore better to merge them.

The classification algorithms all use different features to get the best model. The Support Vector Machine and k-Nearest Neighbor classifiers have most occurrences in features which are of importance in the model. The Logistic regression and Random Forest classifiers differ the most. The feature 'most used category' in the Random Forest model was not important to the model in case the other features were present. After removing the 'most used category' feature, the performance of the classifier increased enormously. Nevertheless, when removing all other features and only keeping the predictor 'most used category', the Random Forest model was still able to perform better than the baseline.

The Random Forest is the only classifier which significantly improves by removing some of the predefined less important features. The accuracy of the other models stay approximately the same or decrease when removing less important features from the model. When removing features which are proven to have more effect in the model, the only model which seems to improve slightly is the Support Vector Machine. Concluding, feature performances do affect the model's behavior, but even though some of the important features were not present, the models still performed better than the baseline by converting the less important features into more important features for the model.

References

Aggarwal, C. (2014). An introduction to Data Classification. *Data Classification: Algorithms And Applications*, 6–7. doi: 10.1201/b17320

Ahn, H., & Kim, K.-J. (2011). Corporate Credit Rating using Multiclass Classification Models with order Information . *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, *5*(12).

Akbulut, Y., Sengur, A., Guo, Y., & Smarandache, F. (2017). NS-k-NN: Neutrosophic Set-Based k-Nearest Neighbors Classifier. *Symmetry*, *9*(9), 179. doi: 10.3390/sym9090179

Alvarez-Lozano, J., Osmani, V., Mayora, O., Frost, M., Bardram, J., Faurholt-Jepsen, M., & Kessing, L. V. (2014). Tell me your apps and I will tell you your mood. *Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments - PETRA 14*. doi: 10.1145/2674396.2674408

Aly, M. (2005). Survey on Multiclass Classification Methods. *Neural Networks Technical Report, Caltech., 19, 1-9., doi:10.1.1.175.107*

Becker, D., Funk, B., Riper, H., Bremer, V., Asselbergs, J., & Ruwaard, J. (2016). How to Predict Mood? Delving into Features of Smartphone-Based Data. Twenty-second Americas Conference on Information Systems, San Diego.

Buttenberg, K. (2015). The classification of customer- and brand-oriented marketing capabilities. *Transnational Marketing Journal*, *3*(1), 26–44. doi: 10.33182/tmj.v3i1.407

Bi, J., & Zhang, C. (2018). An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. *Knowledge-Based Systems*, *158*, 81–93. doi: 10.1016/j.knosys.2018.05.037

Bishop, P., & Herron, R. (2015). Use and Misuse of the Likert Item Responses and Other Ordinal Measures. *Internal Journal of Exercise Science, 8*(3), 297-302.

Breiman, L. (2001). Random Forests. *Machine Learning* 45(1), 5-32.

Cao, H., & Lin, M. (2017). Mining smartphone data for app usage prediction and recommendations: A survey. *Pervasive and Mobile Computing*, *37*, 1–22. doi: 10.1016/j.pmcj.2017.01.007

Calcagno, V., & Mazancourt, C. D. (2010). glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models. *Journal of Statistical Software*, *34*(12). doi: 10.18637/jss.v034.i12

Chang, L. (1994). A Psychometric Evaluation of 4-Point and 6-Point Likert-Type Scales in Relation to Reliability and Validity. *Applied Psychological Measurement*, *18*(3), 205–215. doi: 10.1177/014662169401800302

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, *42*(12), 1150–1152. doi: 10.1111/j.1365-2923.2008.03172.x

Dawes, J. (2008). Do Data Characteristics Change According to the Number of Scale Points Used? An Experiment Using 5-Point, 7-Point and 10-Point Scales. *International Journal of Market Research*, *50*(1), 61–104. doi: 10.1177/147078530805000106

Deng, H., Sun, Y., Chang, Y., & Han, J. (2014). Probabilistic models for classification. In *Data Classification: Algorithms and Applications*, 65–86. CRC Press. doi: 10.1201/b17320

Döring, M. (2018, December 4). Performance Measures for Multi-Class Problems. Retrieved November 27, 2019, from https://www.datascienceblog.net/post/machine-learning/performance-measures-multi-class-problems/.

Elwood, J. M. (2014). Mobile phones, brain tumors, and the limits of science. *Bioelectromagnetics*, *35*(5), 379–383. doi: 10.1002/bem.21853

Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research 15. doi: 3133-3181

Gunter, B. (2019). *Children and mobile phones: adoption, use, impact, and control*. Bingley, UK: Emerald Publishing.

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, *7*(6), 836–850. doi: 10.1016/j.cptl.2015.08.001

Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR)*, *5*(1), 1842-1845.

Jovanovic, A., & Perovic, A. (2010). On classification and decision making. *IEEE 8th International Symposium on Intelligent Systems and Informatics*. doi: 10.1109/sisy.2010.5647130

Kiang, M. Y. (2003). A comparative assessment of classification methods. *Decision Support Systems*, *35*(4), 441–454. doi: 10.1016/s0167-9236(02)00110-0

Knapp, T. R. (1990). Treating Ordinal Scales as Interval Scales. *Nursing Research*, *39*(2). doi: 10.1097/00006199-199003000-00019

Kranse, R., Roobol, M., & Schröder, F. H. (2008). A graphical device to represent the outcomes of a logistic regression analysis. *The Prostate*, *68*(15), 1674-1680.

Lakshmanaprabu, S. K., Shankar, K., Ilayaraja, M., Nasir, A. W., Vijayakumar, V., & Chilamkurti, N. (2019). Random forest for big data classification in the internet of things using optimal features. *International Journal of Machine Learning and Cybernetics*, *10*(10), 2609–2618. doi: 10.1007/s13042-018-00916-z

Lango, M. (2019). Tackling the Problem of Class Imbalance in Multi-class Sentiment Classification: An Experimental Study. *Foundations of Computing and Decision Sciences, 44*(2), 151–178. doi: 10.2478/fcds-2019-0009

Lantz, B. (2015). *Machine learning with R: learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications*. Birmingham: Packt Publ.

Likamwa, R., Liu, Y., Lane, N. D., & Zhong, L. (2013). MoodScope. Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys 13. doi: 10.1145/2462456.2483967

Liu, H., & Gegov, A. (2015). Collaborative Decision Making by Ensemble Rule Based Classification Systems. *Studies in Big Data Granular Computing and Decision-Making*, 245–264. doi: 10.1007/978-3-319-16829-6_10

Mehrotra, A., Tsapeli, F., Hendley, R., & Musolesi, M. (2017). MyTraces: Investigating Correlation and Causation between Users' Emotional States and Mobile Phone Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1*(3). doi: 10.1145/3130948

Mosley, L. (2013). A balanced approach to the multi-class imbalance problem. *Iowa State University Capstones, Thesis and Disstertations.* doi: 10.31274/etd-180810-3375

Nasa, C., & Suman, S. (2012). Evaluation of Different Classification Techniques for WEB Data. *International Journal of Computer Applications*, *52*(9), 34–40. doi: 10.5120/8233-1389

Nazir, T., Qi, X., & Silvestrov, S. (2016). Linear and Nonlinear Classifiers of Data with Support Vector Machines and Generalized Support Vector Machines. *Springer Proceedings in Mathematics & Statistics Engineering Mathematics II*, 377–396. doi: 10.1007/978-3-319-42105-6_18

Nieciecka, D. (2018). Predicting Happiness - Comparison of Supervised Machine Learning Techniques Performance on a Multiclass Classification Problem. *Technological University Dublin. School of Computing.*

Noi, P. T., & Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors, 18*(2), 18. doi: 10.3390/s18010018

Patro, S. K., & Sahu, K. K. (2015). Normalization: A Preprocessing Stage. *Iarjset*, 20–22. doi: 10.17148/iarjset.2015.2305

Srinivasan, V., Moghaddam, S., Mukherji, A., Rachuri, K. K., Xu, C., & Tapia, E. M. (2014). MobileMiner: Mining Your Frequent Patterns on Your Phone . *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp 14 Adjunct*. doi: 10.1145/2632048.2632052

Stragier, J., Hendrickson, D., Vanden Abeele, M., & De Marez, L. (2019). *Unlock, Chat, Lock. A Markov Chain Analysis to Unveil How Smartphone Use Unfolds in Everyday Life*. Paper presented at Annual Conference of the International Communication Association 2019, Washington, United States.

Student, S., Pieter, J., & Fujarewicz, K. (2016). Multiclass Classification Problem of Large-Scale Biomedical Meta-Data. *Procedia Technology*, *22*, 938–945. doi: 10.1016/j.protcy.2016.01.093

Sugiyama, S. (2010). Fashion and the mobile phone: a study of symbolic meanings of mobile phone for college-age young people across cultures. *Mobile Media and the Change of Everyday Life*. doi: 10.3726/978-3-653-01460-0/18

Sun, J., Bo, Y., Luo, J., & Yang, J. (2019). Application of the K Nearest Neighbor Algorithm Based on Scaling Weight in Intelligent Attendance System. *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. doi: 10.1109/icmtma.2019.00142

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2014). *Introduction to data mining* (2nd ed.). Boston: Pearson Addison Wesley.

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*. doi: 10.1016/j.aci.2018.08.003

Topalli, B. (2016). The Mobile Phone & Its Impact In Teenagers`Daily Life. *European Scientific Journal, ESJ*, *12*(8), 161. doi: 10.19044/esj.2016.v12n8p161

Turner, M., Love, S., & Howell, M. (2008). Understanding emotions experienced when using a mobile phone in public: The social usability of mobile (cellular) telephones. *Telematics and Informatics*, *25*(3), 201–215. doi: 10.1016/j.tele.2007.03.001

Vardasca, R., Vaz, L., & Mendes, J. (2017). Classification and Decision Making of Medical Infrared Thermal Images. *Lecture Notes in Computational Vision and Biomechanics Classification in BioApps*, 79–104. doi: 10.1007/978-3-319-65981-7_4

Venables, W. N. & Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth edition. Springer

Wang, P. & Lin, C. (2014). Support Vector Machines. In Data Classification: Algorithms and Applications, 187–204. CRS Press. doi: 10.1201/b17320-8

Wainer, J. (2016). Comparison of 14 different families of classification algorithms on 115 binary datasets. *Computing Insititute. University of Campinas.* ArXiv: 1606.00930

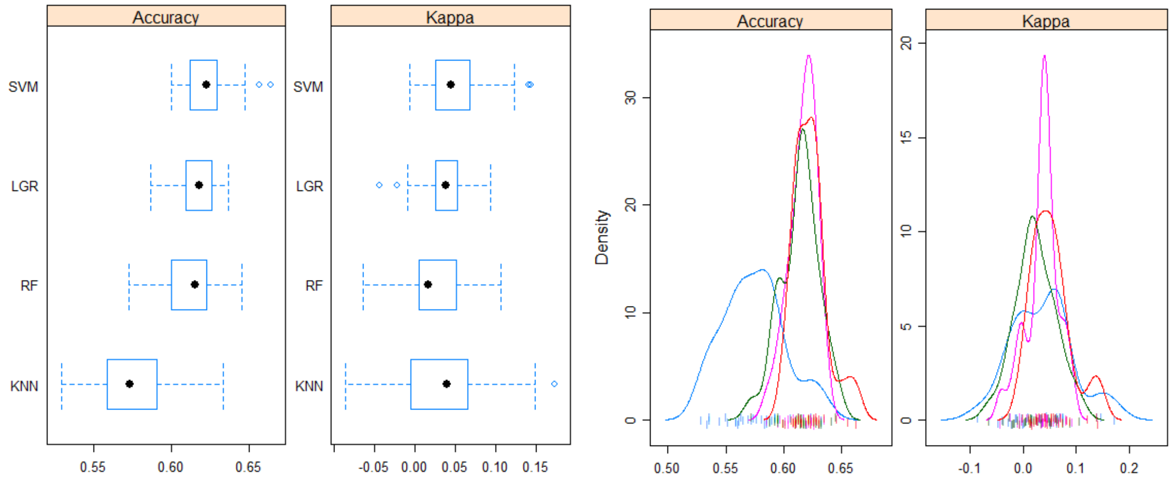
Xu, M. L., & Leung, S. O. (2018). Effects of varying numbers of Likert scale points on factor structure of the Rosenberg Self-Esteem Scale. *Asian Journal of Social Psychology, 21*(3), 119–128. doi: 10.1111/ajsp.12214

Yusa, M., & Utami, E. (2017). Classifiers evaluation: Comparison of performance classifiers based on tuples amount. *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. doi: 10.1109/eecsi.2017.8239204

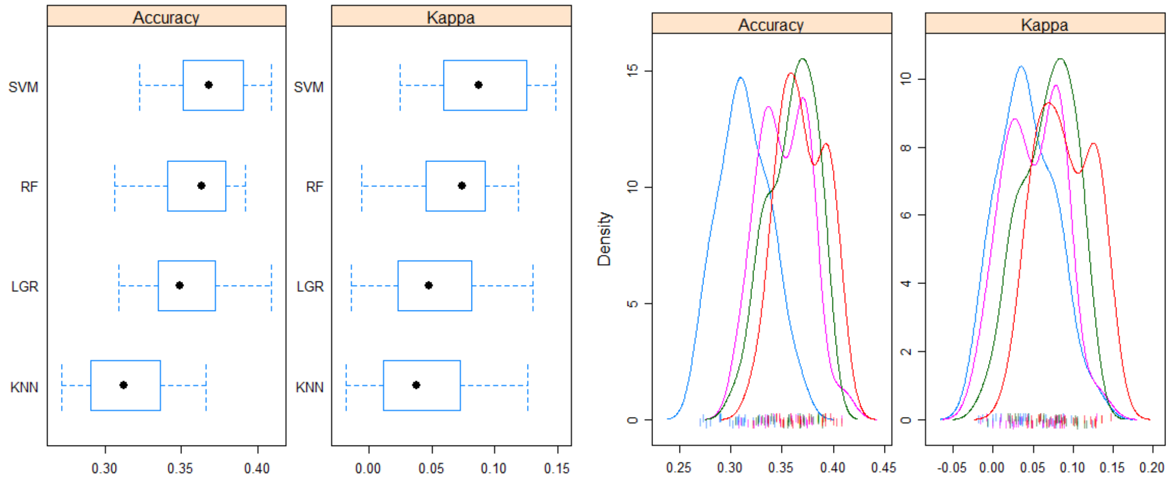
Zheng, Y. (2015). *Encyclopedia of mobile phone behavior*. Hershey, PA: Information Science Reference.

Appendix A: accuracy and kappa visualizations

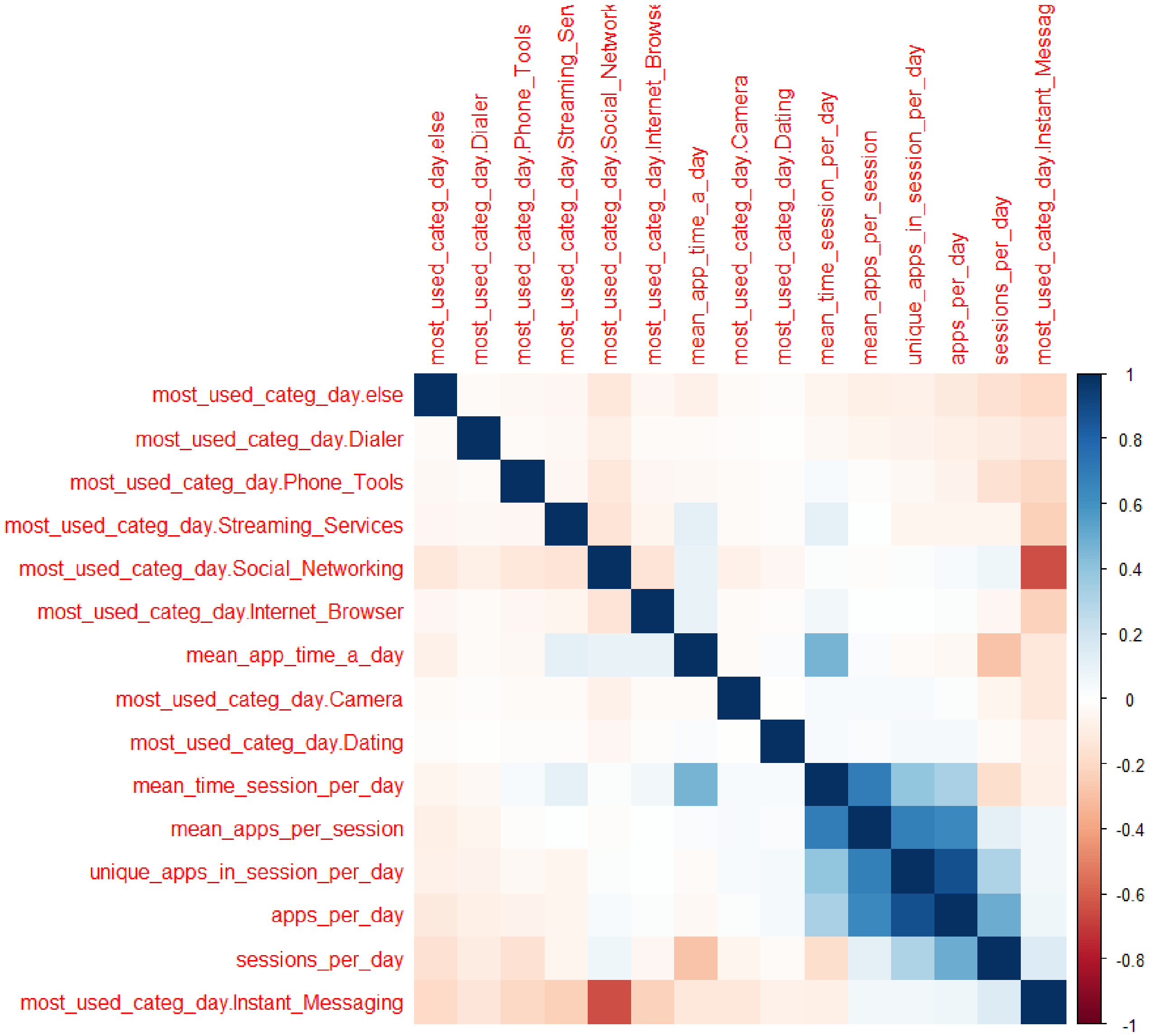
Binary dataset



4-class dataset

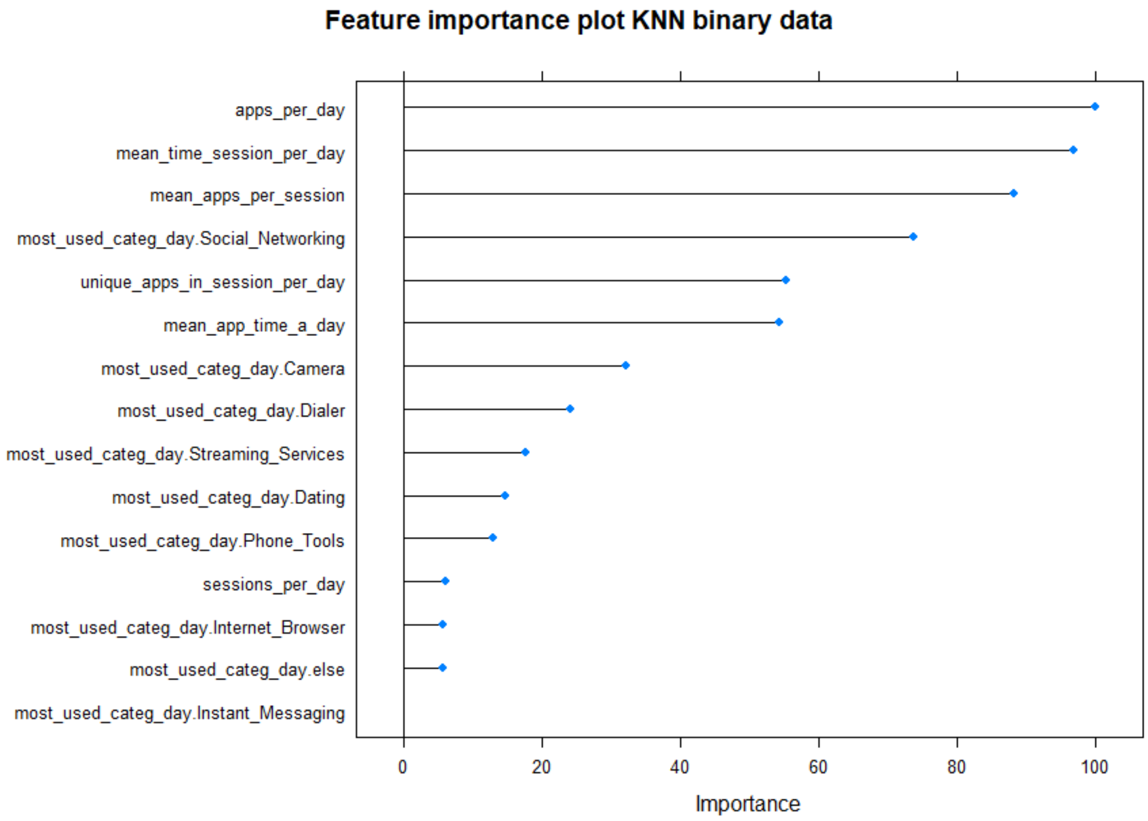


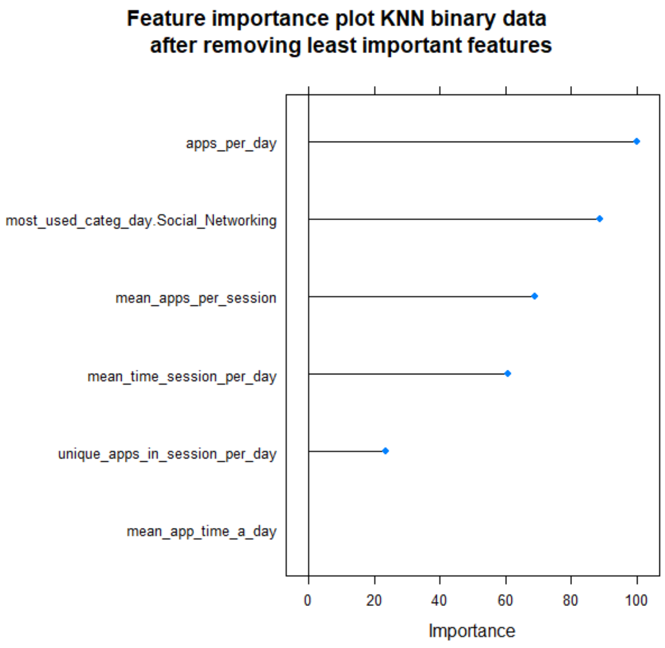
Appendix B : correlation plot

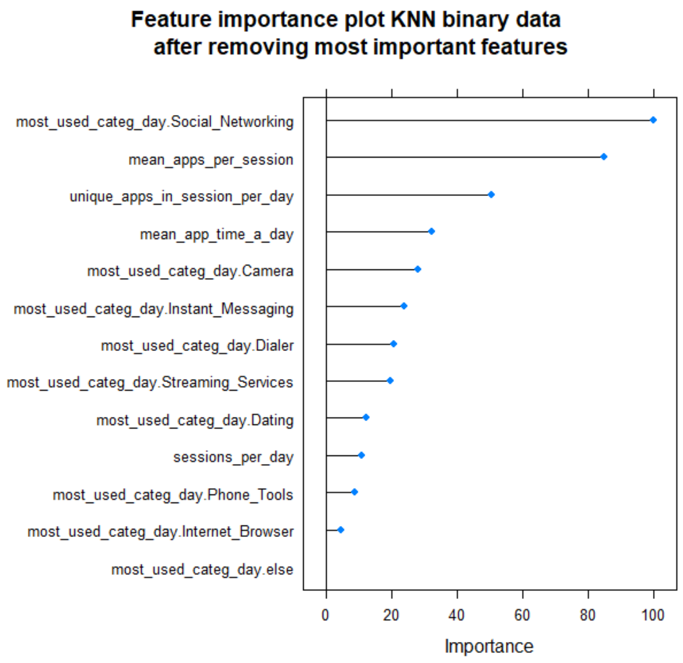


Appendix C: Feature importance plots binary data

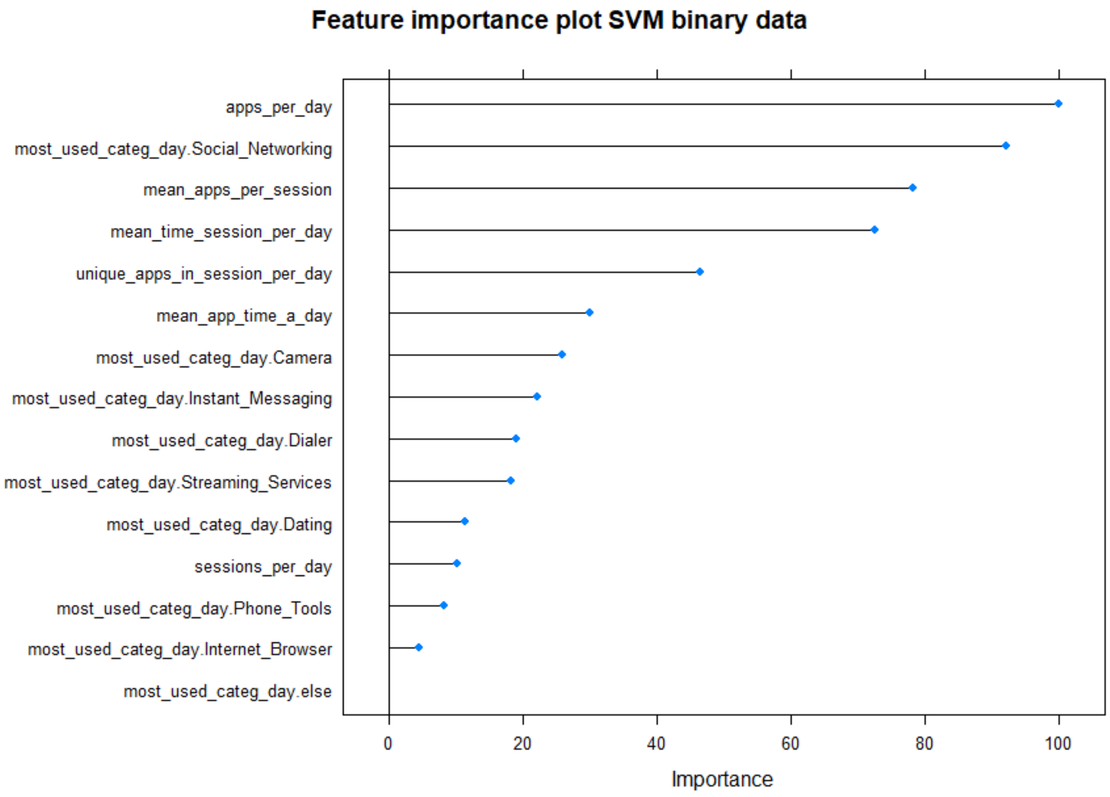
kNN feature importance

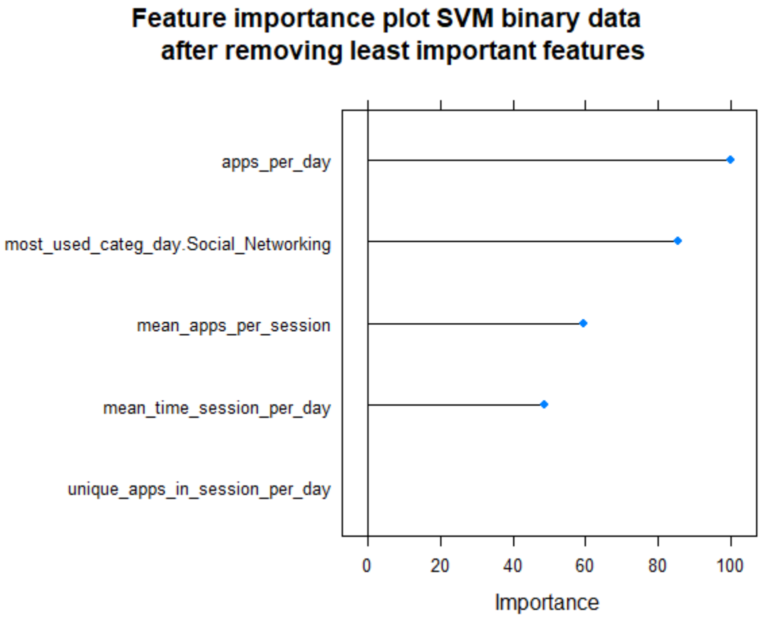


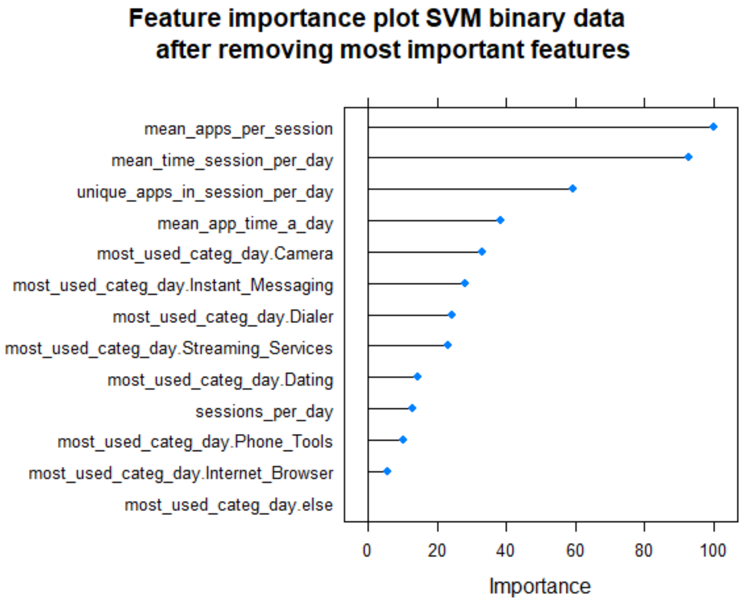




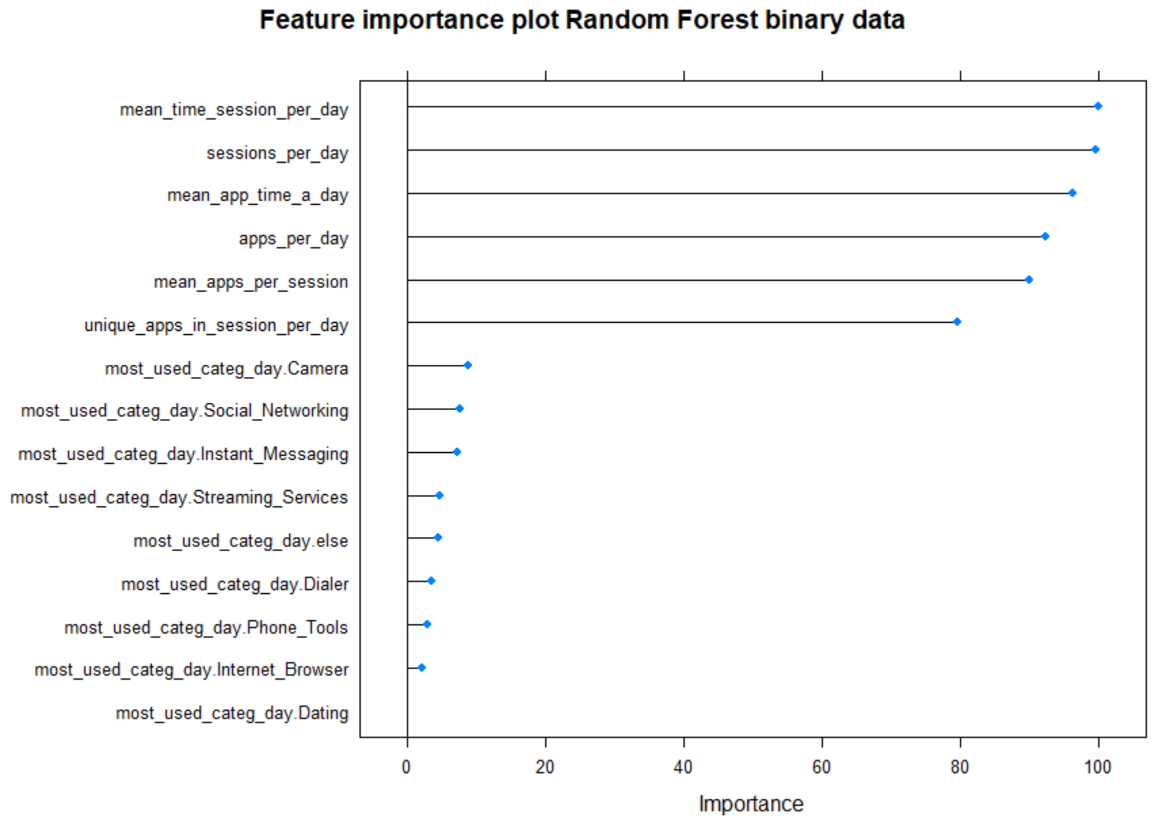
SVM feature importance

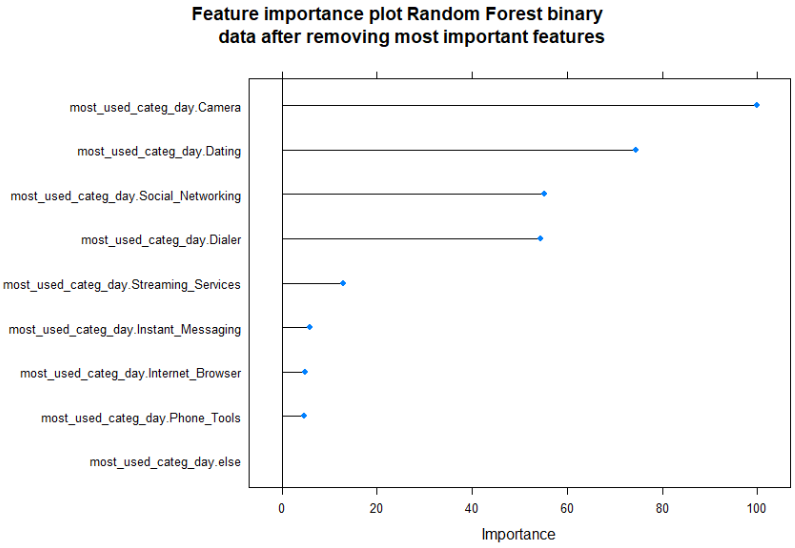


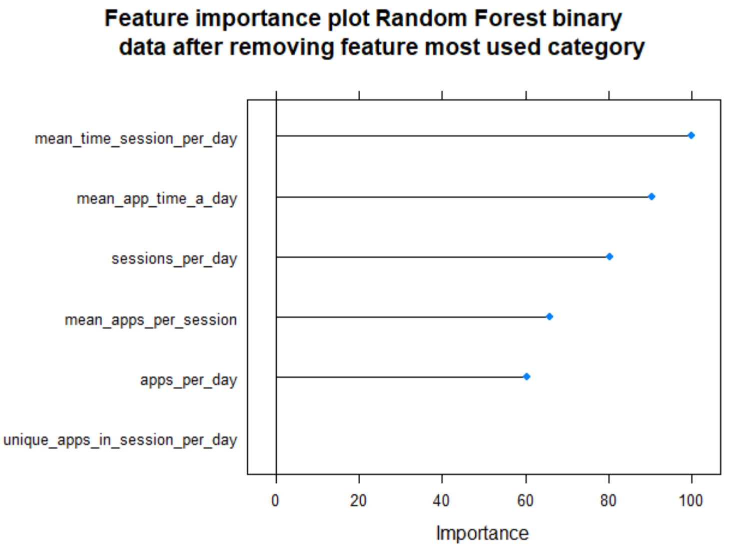




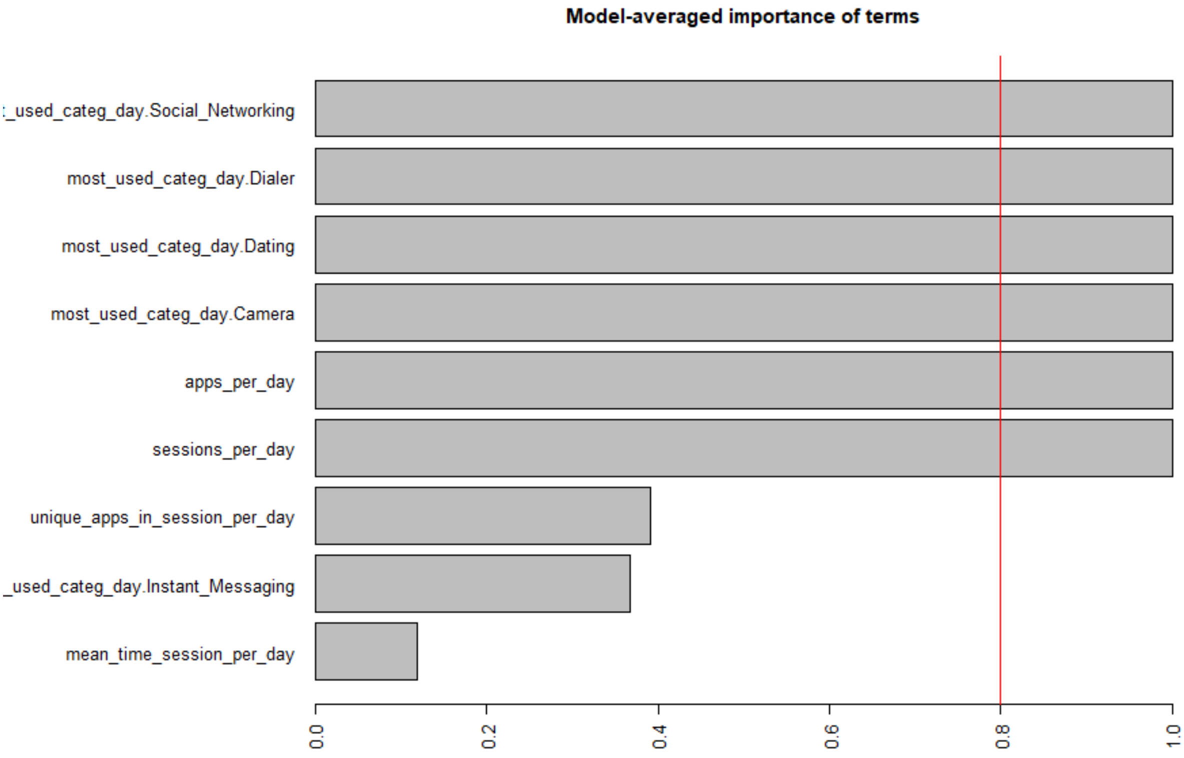
Random forest feature importance


Feature importance plot Random Forest binary data

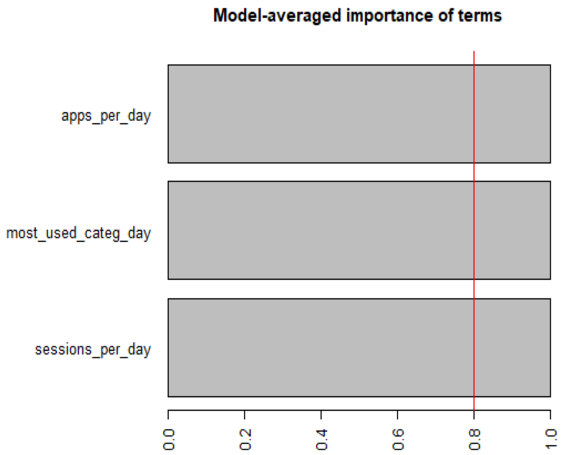

Feature importance plot Random Forest binary data after removing most important features

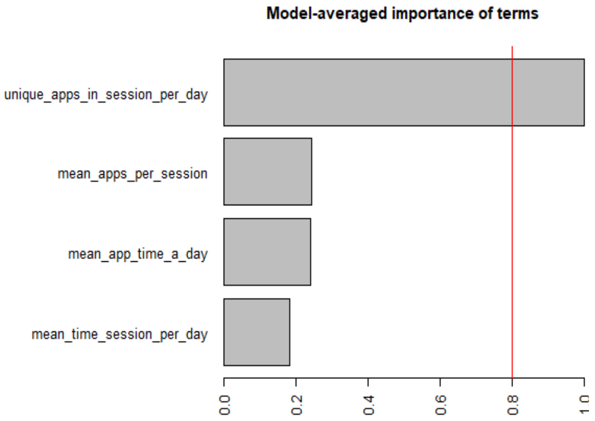

Feature importance plot Random Forest binary data after removing feature most used category

Logistic regression feature importance



Gmulti feature importance on the binary dataset with all features present

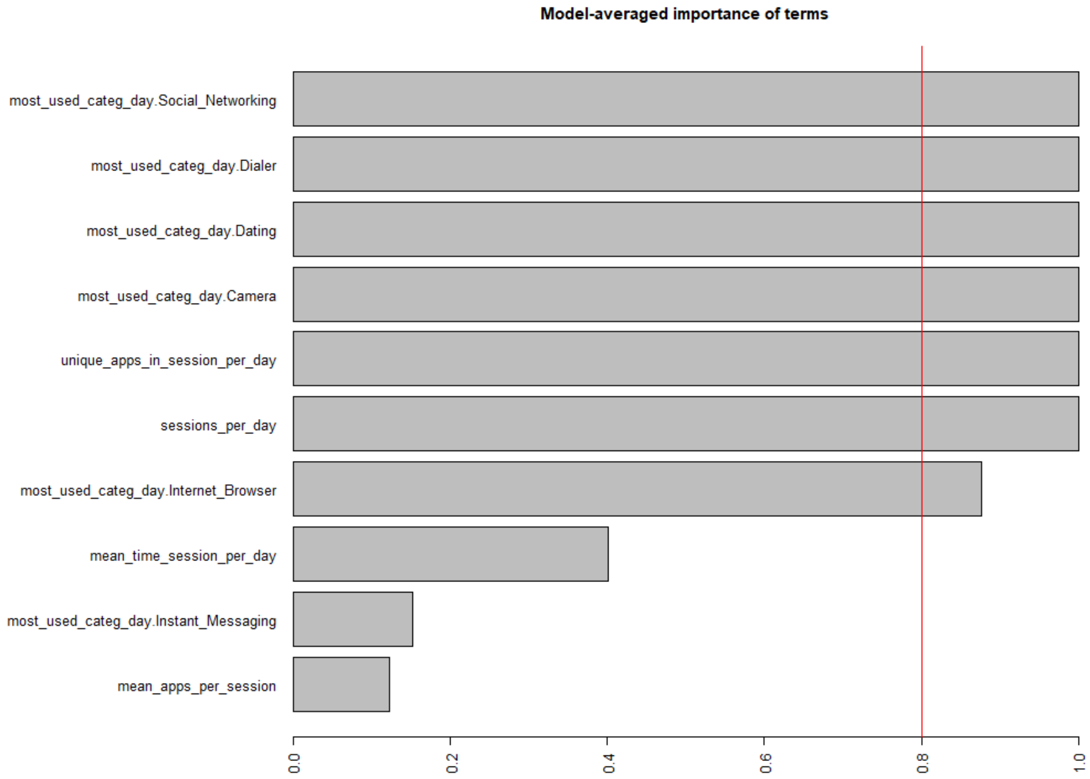




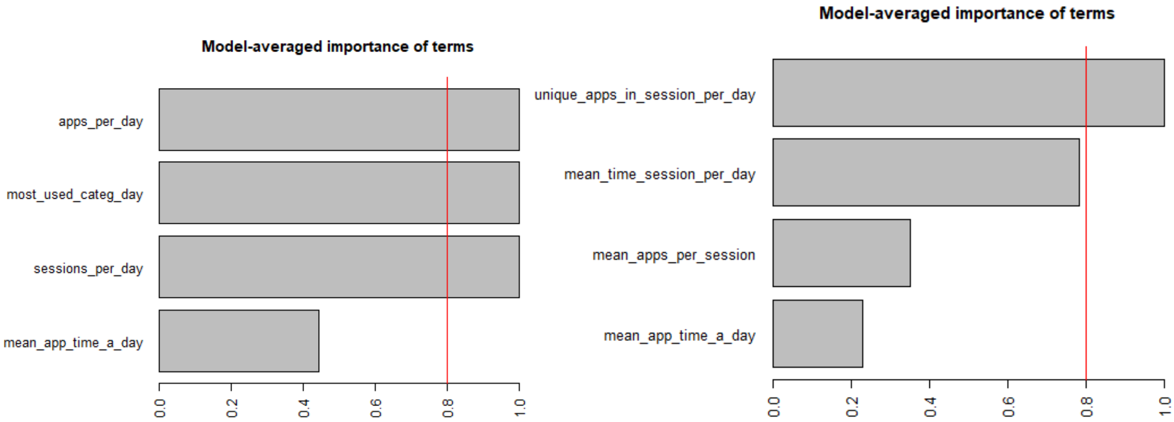
Gmulti feature importance on the binary dataset with only important features on the left side and without the most important features on the right side.

Appendix D: Feature importance plots 4-class data

Logistic Regression feature importance

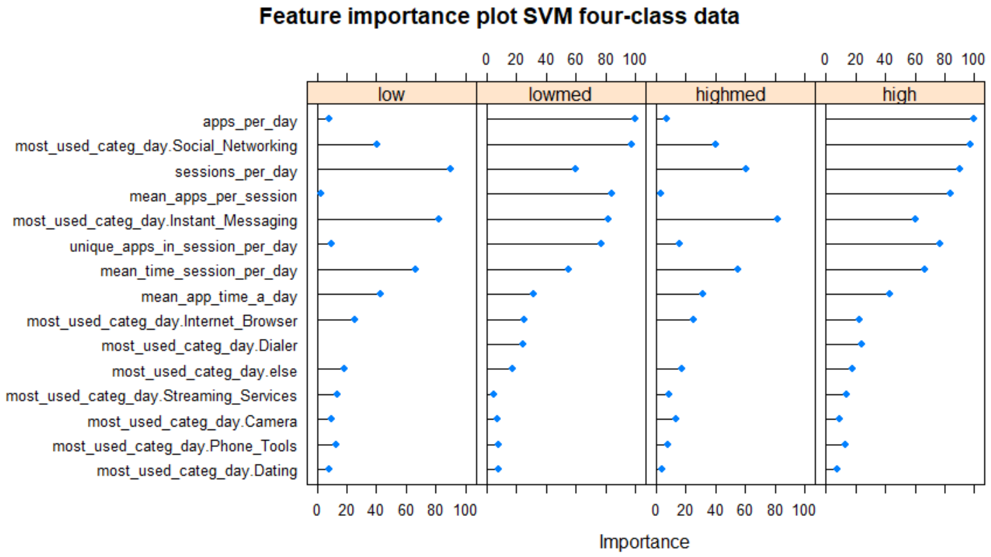


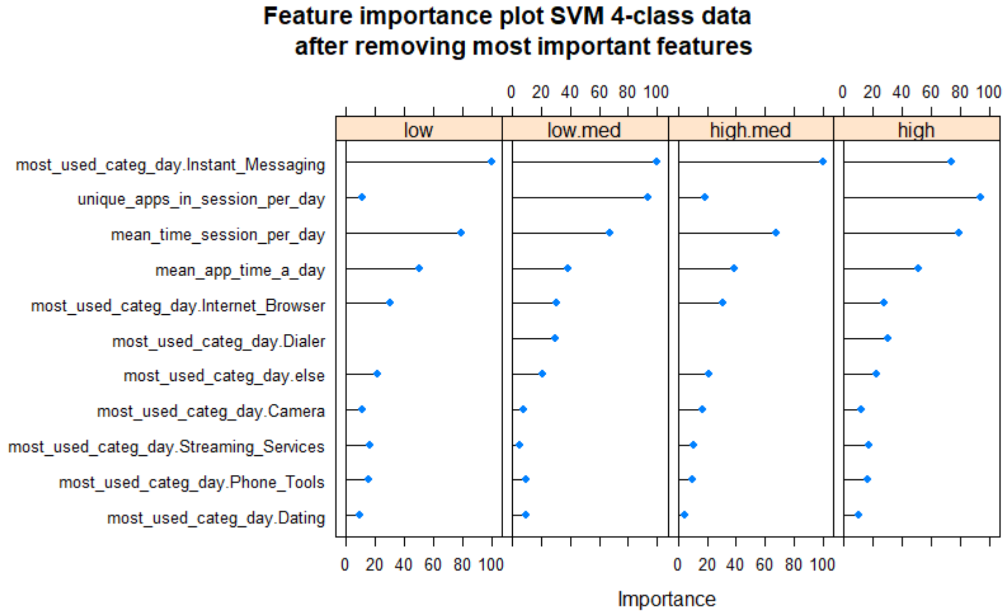
Gmuti feature importance on the 4-class dataset with all features.

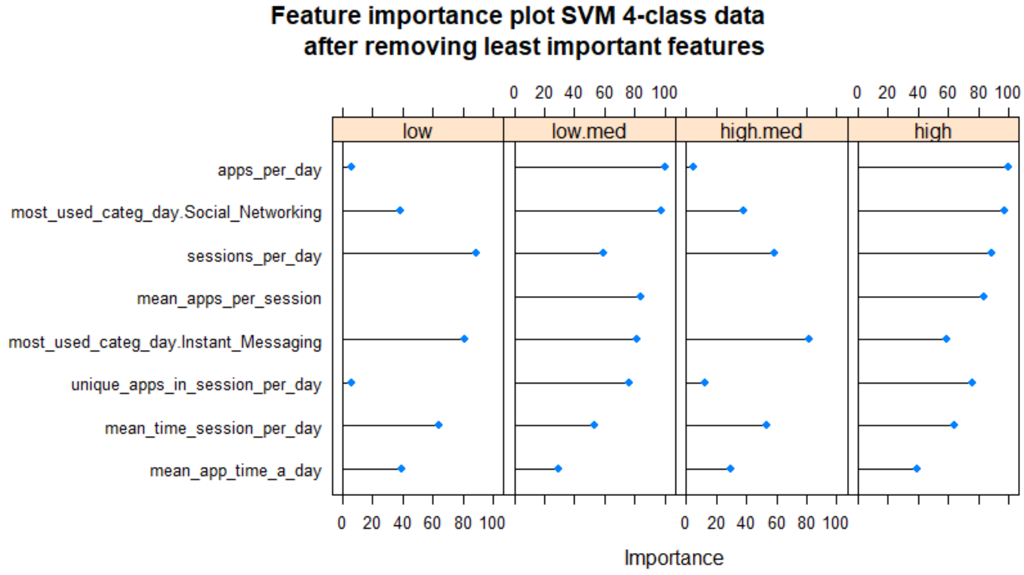


Gmulti feature importance on the 4-class dataset with only important features on the left side and without the most important features on the right side.

Support Vector machine feature importance

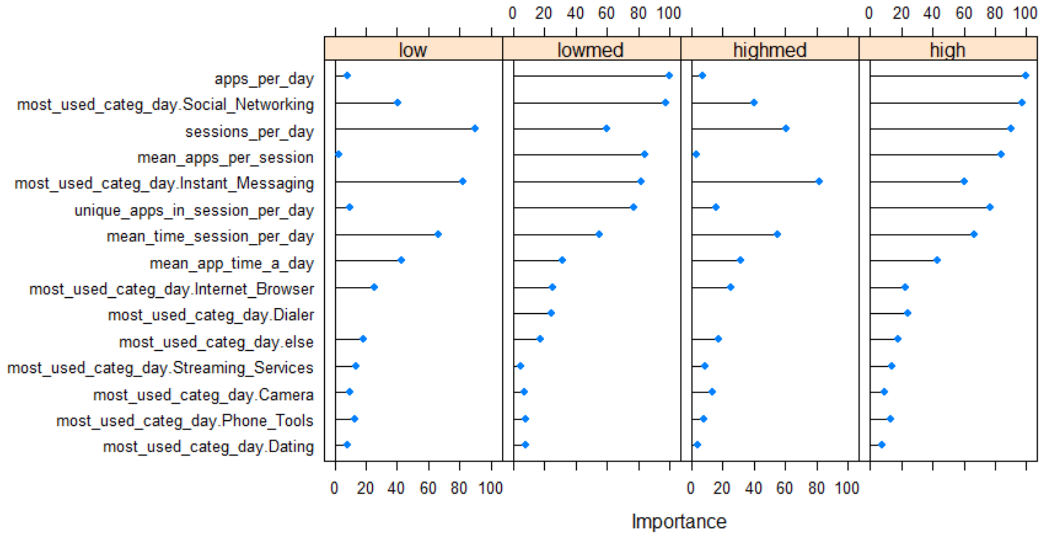




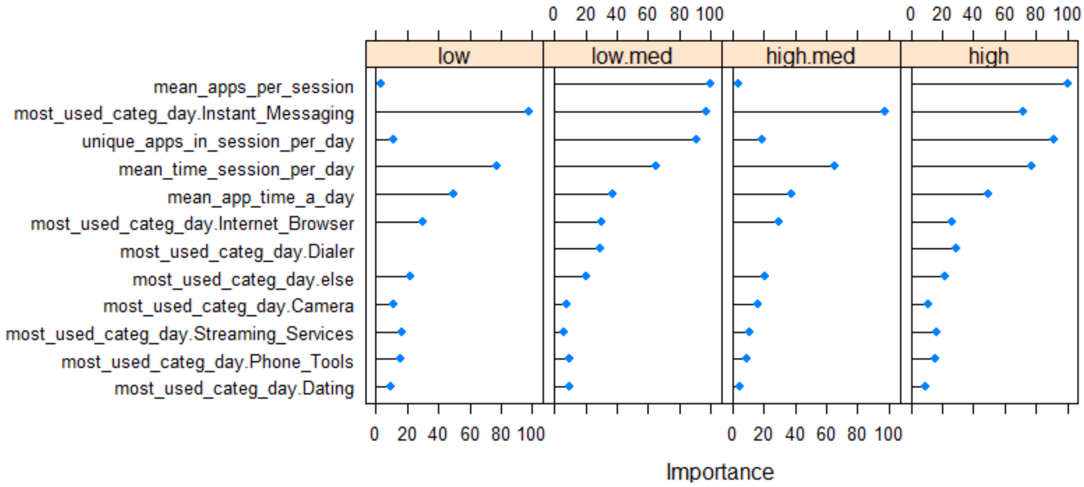


k-Nearest Neighbor feature importance

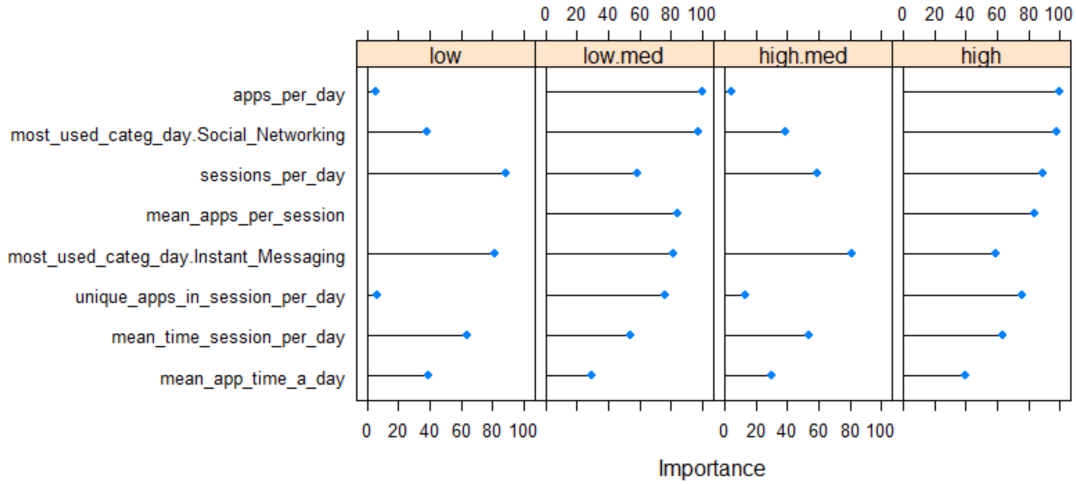
Feature importance plot kNN four-class data



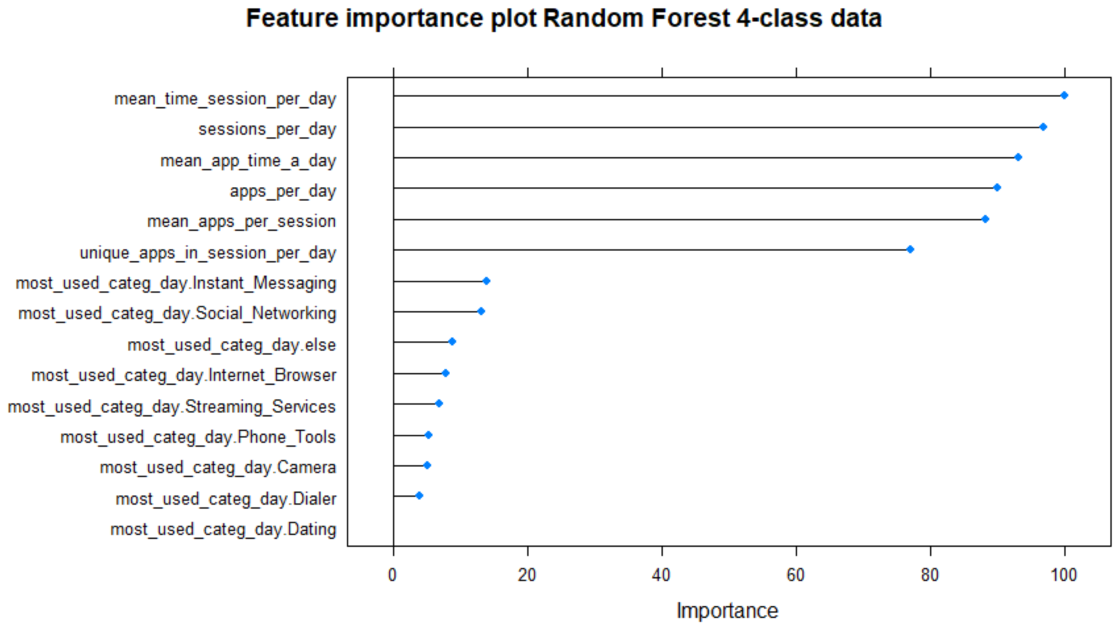
Feature importance plot kNN 4-class data after removing most important features

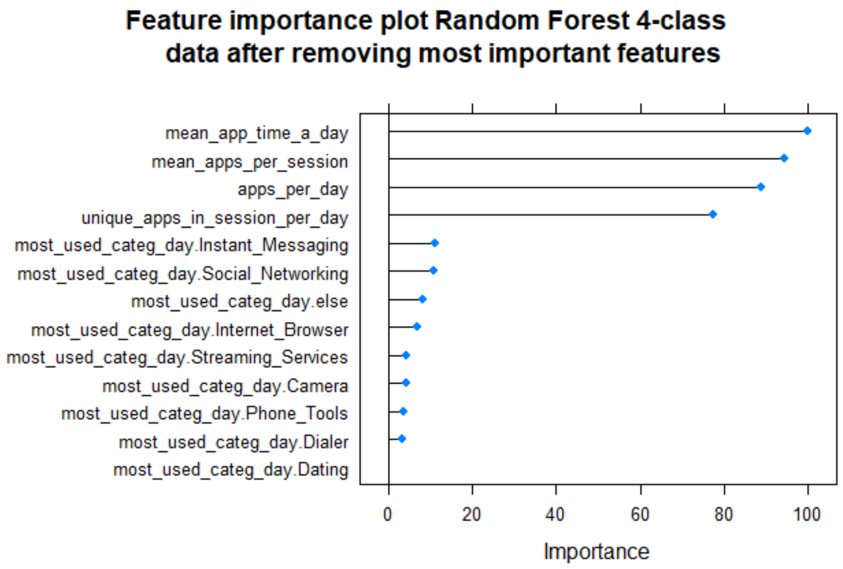


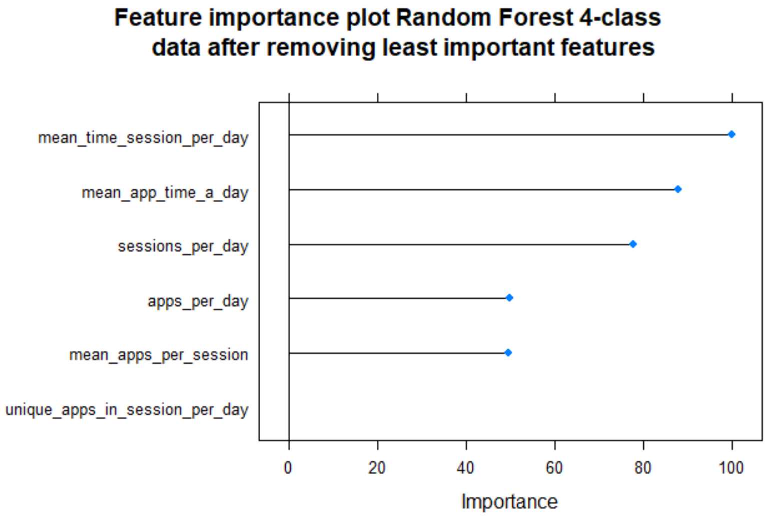
Feature importance plot kNN 4-class data after removing least important features



Random Forest feature importance


Feature importance plot Random Forest 4-class data


Feature importance plot Random Forest 4-class data after removing most important features


Feature importance plot Random Forest 4-class data after removing least important features

Appendix C: precision and recall for classes of 4-class dataset

Class	*Precision*	*Recall*	*Precision – less imp*	*Recall - less imp*	*Precision – imp*	*Recall - imp*
Low	0.508	0.409	0.444	0.445	0.501	0.413
Low-med	0.394	0.800	0.395	0.723	0.397	0.788
High-med	0.560	0.174	0.441	0.183	0.520	0.198
High	0.582	0.103	0.450	0.029	0.725	0.093

Table 1. precision and recall of the SVM classifier on the 4-class data

Class	*Precision*	*Recall*	*Precision – less imp.*	*Recall – less imp.*	*Precision – imp.*	*Recall - imp.*
Low	0.522	0.574	0.534	0.596	0.554	0.566
Low-med	0.521	0.610	0.526	0.612	0.512	0.628
High-med	0.506	0.447	0.528	0.454	0.462	0.432
High	0.460	0.276	0.398	0.250	0.482	0.260

Table 2. precision and recall of the kNN classifier on the 4-class data

Class	*Precision*	*Recall*	*Precision – less imp.*	*Recall – less imp.*	*Precision – imp.*	*Recall - imp.*
Low	0.639	0.545	0.994	0.997	0.570	0.354
Low-med	0.443	0.920	0.989	0.999	0.391	0.920
High-med	0.902	0.172	0.994	0.989	0.759	0.082
High	0.852	0.147	1	0.981	0.689	0.099

Table 3. precision and recall of the RF classifier on the 4-class data

Class	*Precision*	*Recall*	*Precision – less imp.*	*Recall – less imp.*	*Precision – imp.*	*Recall - imp.*
Low	0.427	0.208	0.429	0.216	0.431	0.115
Low-med	0.337	0.844	0.337	0.838	0.330	0.924
High-med	0.360	0.051	0.365	0.501	0.000	0.000
High	0.400	0.006	0.286	0.001	NA	0.000

Table 4. precision and recall of the LGR classifier on the 4-class data