Grad Kaggle Project Report 841 (Team MJ) Name: Murad Ahmed ID: 20839186
Name: Jaskirat Singh Bhatia ID: 20817626

ABSTRACT

Life satisfaction is a phenomenon difficult to accurately define or capture. In our study we have been given 270 possible determinants of life satisfaction. Several classification and learning techniques have been employed to determine the best possible combination of such determinants can best explain a person's satisfaction in Europe. Domain knowledge was first used to eliminate some variables that seem to reasonably irrelevant to the study at hand. Variables such as the timing of the interview, the gender of the third to thirteenth person at home, and more. Also, variables with over 50% missing were removed to avoid any bias to the results. This left us with about 184 variables. LASSO was used to further decrease the number of variables to a total of 136 variables. A combination of backward stepwise regression and the importance function in random forests further reduced the number of predictors to a subset of 19 variables. The analysis was done on both the 136 variables and the 19 variables to guarantee that there was no potential loss. Techniques such as boosting, logistic regression, trees, random forests, and neural networks were used. Almost all yielded similar results. Only using an ensemble of trees, random forests and boosting did we get our highest result which was 0.88741 on Kaggle and 0.896 on our test data.

LANGUAGES USED

- Python
- R

EXPLORATORY DATA ANALYSIS AND CLEANING

Domain Knowledge

The main source of exploring data was the long code codebook which contained all the necessary information regarding each variable. We used domain knowledge to eliminate variables that seemed irrelevant such as time of interview, the gender of third to thirteenth person at home, third to thirteenth person in the household in their relation to the respondent, and other variables. Our domain knowledge is mainly from the World Bank's, OECD's definitions of Satisfaction as well as some research on the predictors and determinants of Citizen Satisfaction. Predictors of life satisfaction are generally divided into five categories: economic, trust, safety, demographic, and happiness variables. Economic variables are variables such as income level, satisfaction with country's economy, etc. Trust variables are ones that have to do with trust in the institutions such as parliament, police, political parties, etc. Demographic variables are ones that have to do with gender, age, country, etc. Safety variables have to do with feeling of safety when walking in the streets, able to have a role in politics, freedom of speech, etc. Finally, happiness variables have to do with asking the respondent directly about their state of mind and aspects of satisfaction within their daily life. All such definitions have variables that represent them in our dataset. Accordingly, the following variables were deemed irrelevant to our definitions of satisfaction and thus were removed.

- Variables Dropped: v26, v27, v28, v29, v30, v31, v32, v33, v34, v39, v45, v46, v50, v90, v124, v125, v126, v127, v128, v129, v130, v1131, v133, v192, v193, v194, v195, v196, v197, v198, v199, v200, v201, v202, v203, v204, v205, v206, v207, v209, v210, v211, v212, v213, v214, v215, v258, v260, v261, v262, v263, v264, v265, v266, v267, v268, v269, v270, id (58 variables)

Exploring Missing Data

- 24.35% variables had some missing data
- 3.3% variables had 50% missing data
- 1.8% variables had 80% missing data

The golden rule is to eliminate variables with 80% or more missing data but eliminating variables with more than 50% missing data gave better results.

- Variables Dropped: v5, v22, v23, v69, v71, v86, v87, v88, v89, v91, v92, v93, v94, v95, v96, v97, v123, v151, v153, v158, v164, v168, v174, v182, v191, v241, v243, v259 (28 variables)

Some variables were read incorrectly by the software as categorical; example – age, years of full-time education, number of people living in the household, number of hours worked.

The data had two types of missing values

- Values which were actually NA – Not Recorded
- Values which corresponded to "not applicable" or "refused to answer" which were represented by ".a", ".b", ".c", ".d".

All the missing values in the second type were replaced by NA hence were treated the same as type 1 variables were treated.

Treating Missing Data

- All categorical values were imputed by the mode
- All Numerical values were imputed by the median

It was also made sure that categories were equally spaced (that is the difference between categories is one unit)

- Variables with inconsistent numbered categories have had those numbers changed such as in the case of v56, v58, v60, v62, v65, v66, v67, v68 which all reflect the level of education for respondent as well as respondent's parents and had a category numbered as "55" while other categories were 1 to 5 or 1 to 7. The categories were recoded to be in the same direction as well such that as we go higher in categories, the more positive the variable is (i.e. in v1 "Able to take active role in political life" the higher the response the better or more positive the variable).

- Ordinal variables were also converted to ordered factors to retain the significant and meaning of a category's order.

Using Least Absolute Selection and Shrinkage Operator (LASSO)

Lasso regression is called the Penalized regression method, it is used to select the subset of variables. The LASSO uses the sum of the absolute values of the model parameters and imposes a constraint on them, where the sum has a specified constant as an upper bound. Because of this constraint, the regression coefficients of some variables become zero. This allowed us to identify which variables were most strongly associated with predicting life satisfaction. To optimize LASSO, we need to find the suitable tuning parameter λ which as increases, increases bias and decreases variance. We start out with a default value for $\lambda = 100$ then we plot the misclassification error for different values of λ to get the ideal tuning parameter. The tuning parameter is important because it controls the strength of the penalty, which implies a greater threshold for importance. LASSO excluded about 48 variables and left us with 136 variables.

- Variables Dropped due to LASSO: v3, v4, v6, v7, v8, v11, v12, v14, v16, v17, v20, v21, v24, v25, v37, v40, v44, v47, v49, v52, v53, v54, v55, v57, v89, v61, v63, v72, v78, v100, v107, v132, v134, v150, v154, v155, v157, v161, v165, v166, v167, v171, v175, v176, v185, v221, v230, v257 (48 variables)

Using Random Forest

Calculating the importance of each variate using the permutation which shows how a score decreases when a variable is not included and Gini impurity criteria in Random Forest, we were left with 6 variables as the most importance (Gini >= 13 and Permutation >=8)

Variable ID	Variable Name	Gini Impurity score	Permutation score
V98	How happy are you	100	100
V74	Enjoyed life, how often past week	18.2	11.56
V224	How satisfied with present state of economy in country	21.6	15
V79	Felt depressed, how often past week	13.27	8.075
V101	Feeling about household's income nowadays	17.55	11.44
V253	Were happy, how often past week	19.64	11

Note that the variables with the most importance are one's that literally ask about happiness, enjoyment, depression, etc. which all reflect the secondary satisfaction explained in our domain knowledge. The other variables ask about the economy which is also reflected in the economy part of our domain knowledge. Also note that v98 which is simply a question of how happy is a person correlates directly to our question of satisfaction in both a common sense perspective and an analytical perspective since it has a score of 100 in both criteria and has much greater influence than the other 5 variables combined according to our results.

Using Backward Stepwise Regression

It is a popular method to eliminate irrelevant variables. It starts with taking into consideration all the variables then eliminating the least significant variate one by one. This elimination process continues till all the insignificant variables are eliminated. Using Backward Stepwise Regression at its default values, we were left with 19 variables.

- In addition to the variables retained by Random Forest, Backward Stepwise Regression also retained the following 13 variables which also verify the soundness of our domain knowledge

Variable ID	Variable Name
V81	Felt lonely, how often past week
V1	Able to take active role in political group
V225	State of education in country nowadays
V226	How satisfied with the national government
V99	Subjective general health
V217	Doing last 7 days: retired
V220	How often socially meet with friends, relatives or colleagues
V246	Doing last 7 days: unemployed, not actively looking for job
V82	Felt sad, how often past week
V223	How satisfied with the way democracy works in country
V236	Trust in country's parliament
V180	Most people can be trusted, or you can't be too careful
V179	Most of the time people helpful or mostly looking out for themselves
V233	Trust in the legal system

- The 6 variables calculated by Random Forest were a subset of the 19 variables calculated by Backward Stepwise Regression.
- We decided to work with 2 datasets:
  - Dataset with 19 variables from Backward Stepwise Regression
  - Dataset with 136 variables from LASSO

Training and Testing split

- The train data initially was split into both training and test data to calculate the area under the curve.
  - The split was 70 - 30
- Later for ensembling the train data was split into training 1 and training 2 in a 50-50 split.

ANALYSIS

Initial Analysis

Initially, we went with using various classifiers; after tuning and testing each classifier we got (based on our train/test split): -

Classifier	Accuracy – 19 variables	Accuracy – 136 variables	Accuracy - Kaggle
Logistic Regression	0.8864	0.8841	0.88290 for 19 variables
KNN	0.822	0.7941	0.8007 for 19 variables
Neural Networks	0.877	0.881	0.87700 for 136 variables
Random Forest	0.8883	0.8878	0.88349 for 19 variables
Boosting	0.8897	0.891	0.88463 for 136 variables
Decision Tree	0.8078	0.8078	0.80062 for 136 variables

Note that we conducted the analysis for both 19 and 136 variables on our test/train split then the combination with the higher accuracy was the one tested on Kaggle that is why some classifiers have 19 variables and others have 136 variables.

Tuning Parameters

- **Decision Tree –** It was pruned till we got minimum value for test error.
- **Random Forest –** Performed grid search on 1 – 30 trees at 10 or 20 nodes over a threefold cross validation to get 19 trees and node size of 20 for 136 variables and 8 trees for 19 variables.
- **Boosting –** The number of trees were set to 1000, shrinking factor was 0.1.
- **KNN –** For K = 5, we got maximum accuracy.

Final Analysis

Since Random Forests and Boosting had the highest accuracies and trees is of the same family of classifiers and had a much lower accuracy, we did an ensemble of random forests, boosting, and trees.

- Used 50% data to train a random forest, boosting and decision trees.
- Implemented an ensemble classifier on the above three trained models with logistic regression to get their weights.

Classifier	Weight – 19 variables	Weight – 136 variables
Random Forest	0.49	0.5
Decision Trees	0.03	0.02
Boosting	0.48	0.48

- Accuracy with 19 Variables – 88.3%
- Accuracy with 136 Variables – 89.6%

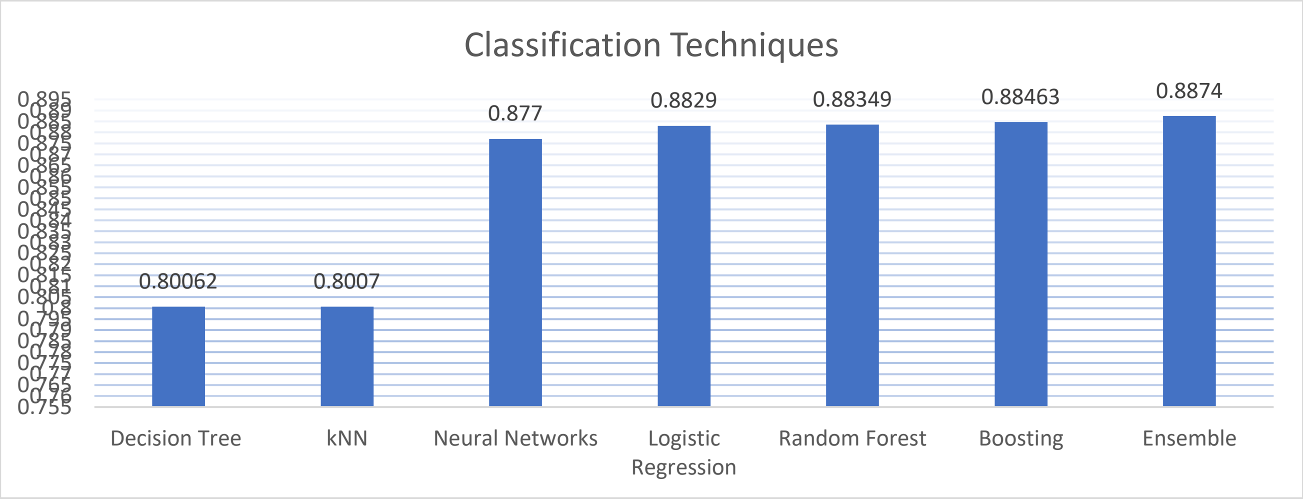
Note that we noticed that Logistic Regression, Random Forests, and Boosting had the best results, so we tried to stack them with one another, but they gave the same result as ensembling.

RESULTS AND CONCLUSIONS

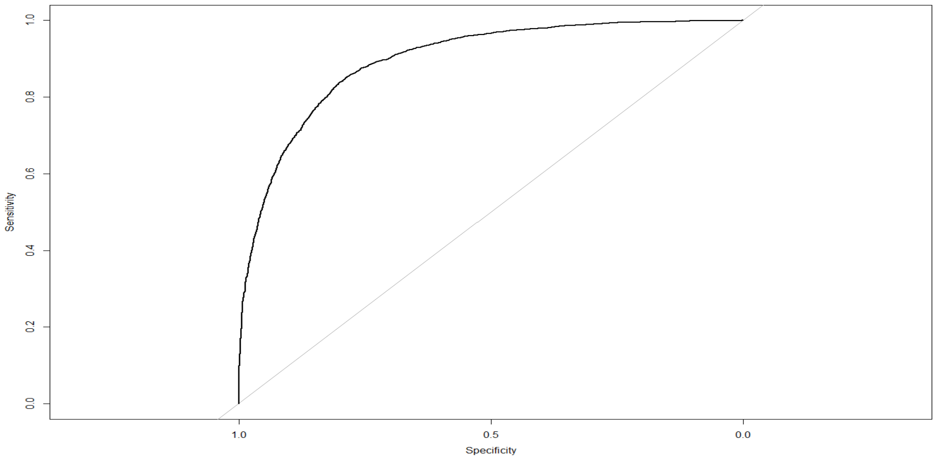
- Accuracy on our test data – 89.6% (136 variables)
- Accuracy on Kaggle – 88.74% (136 variables) and 88.484% (19 variables)
- Position on leaderboard – 30
- V98 "How happy are you?" is the most important and strongest predictor of satisfaction.

Graphical Summary of Analysis

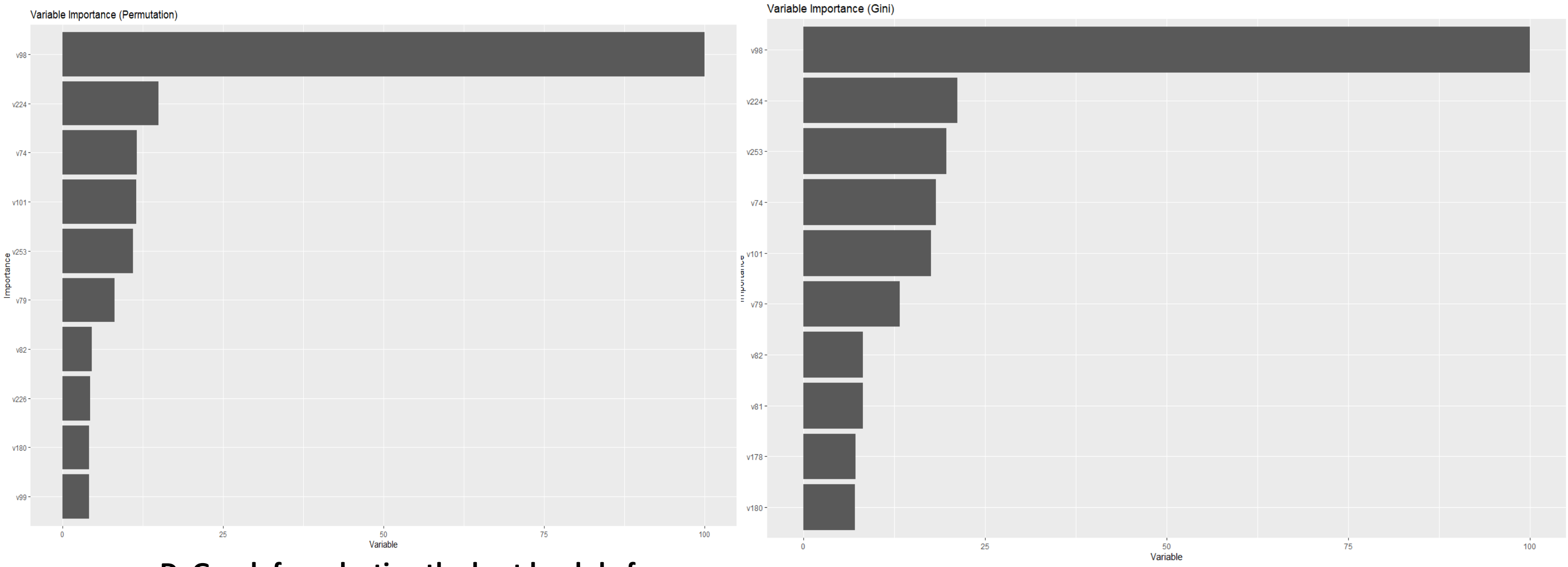
a. Bar Graph of All Analyses



b. ROC curve of Final Analysis (0.896)

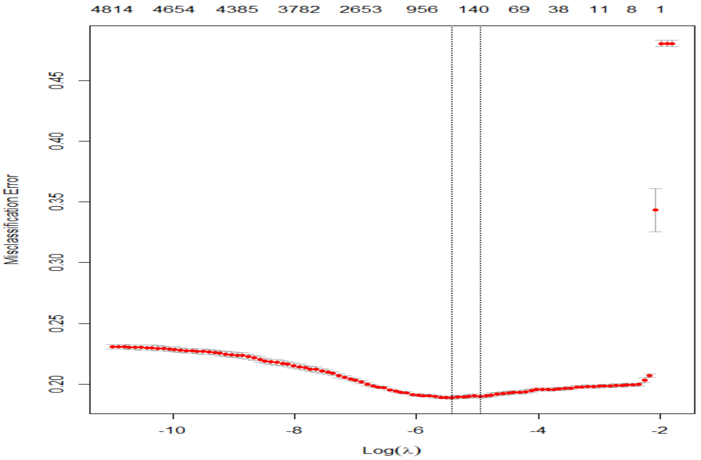


C. Variable Importance using Random Forests (Gini and Permutation criteria top 10)



D. Graph for selecting the best lambda for LASSO

Selecting the best value for lambda for LASSO, we can observe from the plot that the best $\log(\lambda) = -5.426$ which means our $\lambda = 0.00443$



REFERENCES

Domain Knowledge and Ideas for Analysis

1) **CARINUGAN T. JOAN, REARIO A. RITA MA., LIM E. MELODY, ANOG MAYBELLE (2015) Citizen Satisfaction Index System (CSIS) Implementation in Santiago City, Isabela. Asian Review of Public Administration, Vol. 26, Nos. 1 & 2, pp. 32-38**
2) **Kelly, JM & Swindell, D (2002) Service quality variation across urban space: First steps toward a model of citizen satisfaction. Journal of Urban Affairs, Vol 24, No. 3, pp. 271-288.**
3) **The Organisation for Economic Co-operation and Development (2011) https://www.oecd.org/gov/48250728.pdf[accessed 27/2/2020]**
4) **Şandor D. S. (2007) Determinants and Outcomes of Citizen's Satisfaction With Public Services In Cluj-Napoca.Transylvanian Review of Administrative Sciences, 21 E, pp. 103-112**
5) **Song, Miyeon & J Meier, Kenneth (2018) Citizen Satisfaction and the Kaleidoscope of Government Performance: How Multiple Stakeholders See Government Performance. Journal of Public Administration Research and Theory, 10.1093/jopart/muy013.**
6) **https://saeguide.worldbank.org/sites/worldbank.org.saeguide/files/documents/4_Citizen%20Satisfaction%20Surveys.pdf[accessed 1/3/2020]**
7) **Zenker, S, Petersen, S & Aholt, A (2013) The Citizen Satisfaction Index (CSI): Evidence for a Four Basic Factor Model in a German Sample. Cities, vol 31, pp. 156-164. DOI: 10.1016/j.cities.2012.02.006**

Documentation for Packages used for analysis

1) **https://cran.r-project.org/web/packages/caret/caret.pdf for trees and random forest**
2) **https://cran.r-project.org/web/packages/FNN/FNN.pdf for fast kNN**
3) **https://cran.r-project.org/web/packages/gbm/gbm.pdf for boosting**
4) **https://cran.r-project.org/web/packages/glmnet/glmnet.pdf for LASSO**
5) **https://cran.r-project.org/web/packages/MASS/MASS.pdf for stepwise regression**
6) **https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf for Neural Networks**
7) **https://cran.r-project.org/web/packages/ranger/ranger.pdf for fast random forest**
8) **https://cran.r-project.org/web/packages/Rfast/Rfast.pdf for fast logistic regression**
9) **https://cran.r-project.org/web/packages/rpart/rpart.pdf for trees**