Gender Participation in Labor Force

Data Analytics on Labor Force Survey in Armenia

Final Project

BUS 288 - Business Analytics

Instructor: Hrant Davtyan

Student: Alisa Aleksanyan

May 15, 2020

American University of Armenia

Introduction

In the scope of the project, our group works with the Armenian Labor Force Survey Data for 2018.The Labor Force Survey Data includes all the employed and unemployed individuals residing in Armenia, who together comprise the product and service industry.
In this paper, we will bring particular focus to the issue regarding women participation in the Armenian labor force. For many years it has been a culturally accepted norm in Armenia for women unemployment to be higher than that of men, thus the gender unemployment gap remains a critical social and economic issue. To target the indicated problem, the paper will work to analyze which factors play a prominent role in labor participation amongst women in Armenia.

Particularly, the paper will address the following data in the survey; age, gender, education level, education status, marital status, job type, industry, income, hours worked in main and secondary jobs, unemployment rate, willingness to work, and reason for not working. Eventually, the paper will identify the variables that affect women participation in the labor force the most.

Consequently, the paper will suggest measures to increase the employment of women.

Data Description

The target variable of gender is categorical ("Male", "Female"), and is recorded as 1 for "Male" and 2 for "Female". Gender distribution in the data is balanced.

The initial data consists of 28297 observations of 199 variables. The data was converted into 10000 rows and 79 variables by filtering rows which contained answers for both the first and the second part of the questionnaire as well as selecting the variables which were concerning the problem the paper investigates. Further examination of the data led to dropping several variables and finally the data was left with 64 columns (not all of them were used but were supported).

The data does not contain duplicated, all rows represent different individuals.
All missing values for categorical variables were replaced to 0, indicating not answered for some variables and “No” for others where “1” refers to “Yes” and “NA” to “No” (such as unemployment records).
The missing values for numeric variables were replaced with 0, considering that if not answered then either informal work or not an influencing factor in the economy. Filling with mean or median would not be reasonable considering unemployed people do not have income.
Additionally, the data was subset to include only individuals in the age group between 15 to 75 to represent the labor force. The final data consists of 7193 rows and 63 columns.

Descriptive analysis

All the data involves both the presentation as of 1000 persons and as share of total (%).

Labor Force

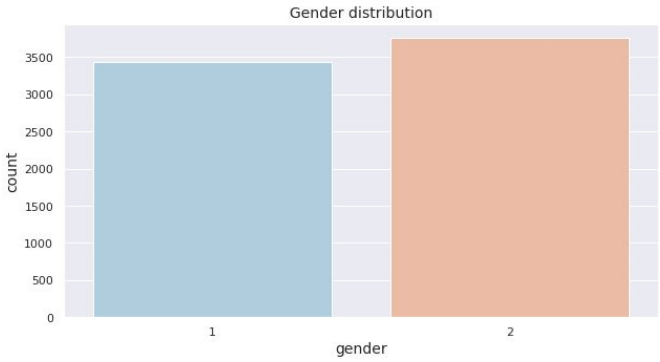
The data that we have here, represents the part of the population, which are from the age group 15-75 and are actively involved in economic activity in the timeframe of the observations.
The data does not include those individuals who create products for their own usage. Those individuals are not considered to be employed, thus, they are out of the above mentioned labor force. This data covers all the types of employment, more specifically, the permanet, temporary, occasional and non-regular types that exist in RA.

Hours of Work

Usual hours of work are characterized by the amount of hours people do the specific job in a week. Hours worked refers to the actual time spent on producing the good or the service in a week. The number can be both higher or lower than that of the usual hours of work. Hours of work describes two types of work: the **main** work of the household, and the **secondary** work. In the main and secondary workplaces the households get their salaries according to the hours that they worked.

Gender Distribution

As already mentioned, the data consists of 7193 households. Out of the 7193, the 3759 households are female (52.3%), and 3434 of them are male (47.7%) .

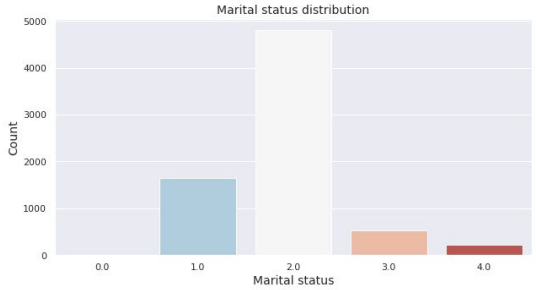


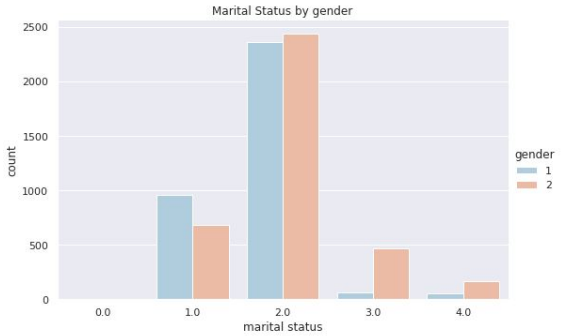
Demographic Information of Respondents

Here we see that male respondents are on average 45 years old, whereas Female households are on average 48 years old. As for the hours worked, the number of working hours of males is bigger than the number of hours worked of females (23 vs 14).

Marital status

Responses from the households about their marital status are provided with the following numbers:



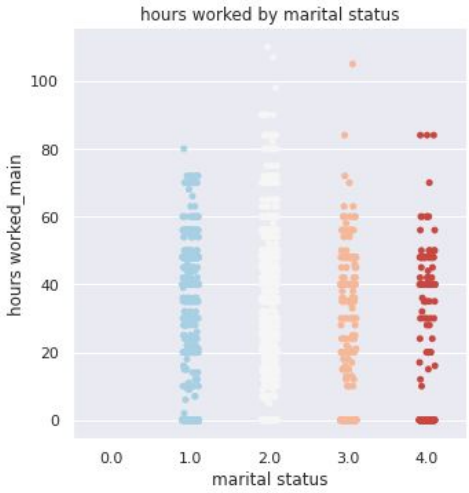


0.0-No response, 1.0-Never married, 2.0-Married, 3.0-Widowed, 4.0-Divorced/Separated
There were no male and female households who did not answer the question regarding the marital status.

- 27.8% of the males are not married yet (957 out of 3434).
- 68.78% of the males from the data are married (2362 out of 3434).
- 1.83% of males are widowed (63 out of 3434).
- 1.51% of males are divorced or separated (52 out of 3434).

- 18.27% of females are not married yet (687 out of 3759).
- 64.91% of females are married (2440 out of 3759).
- 12.42% of females are widowed (467 out of 3759).
- 4.39% of females are divorced or separated (165 out of 3759).

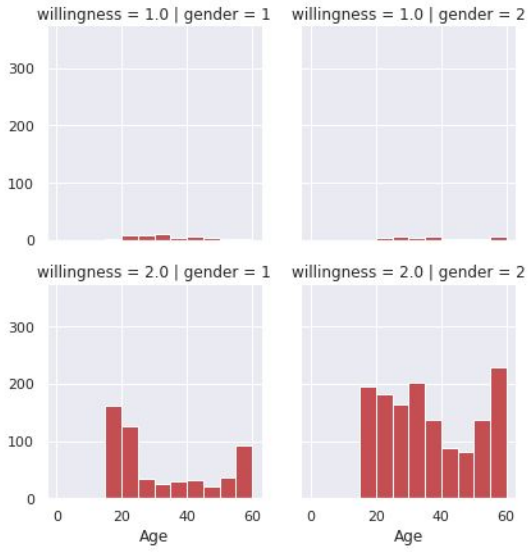
Hours Worked by Marital Status



As can be seen from the graphs, married households are the ones to work more than the others. Next, the hours worked by gender and marital status graph shows that married and divorced men are more likely to work more hours in their main workplace. Married men are the ones who work the most in their secondary workplace. About the Females, the separated / divorced households work more hours than the rest, and the widowed Female households work more hours in their secondary workplace.

Willingness to work

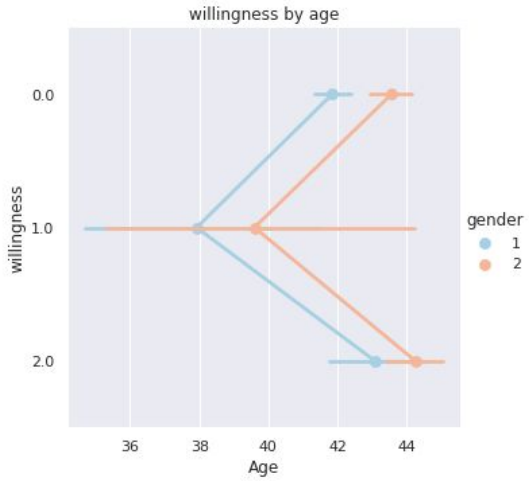
Willingness shows whether households are willing to work. Our data shows the willingness of both Female and Male households. The following figure shows the distribution of willingness by gender.



In this visualization, the important variable for us is the willingness = 2.0, which shows those respondents, who do not look for a job. For males(gender=1) we see that the age groups where they do not look for a job is 20 and more than 60. For females (gender=2) it is shown that most of the female respondents who belong to the age group 20-40 and <60, do not look for a job. The conclusion that we can make from this is the emphasized **effect of willingness to gender classification, specifically, we see that most of the females do not look for a job compared with**

males.

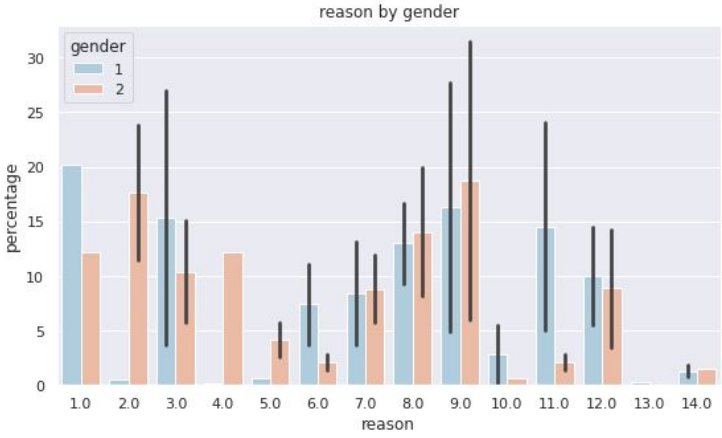
Willingness by Age



The below figure visualizes the willingness to work of the genders by their ages. Thus, **Males are most willing to work in their 38-43 years, and Females are willing to work when they are 40-44 years old.**

Reason by Gender

For the reason variable we have the data that indicates why respondents did not look for a job in the last month. While analyzing the responses it is shown that most of the women (almost 20%) mentioned the reason to be "No hope to find a suitable job" (=9.0) and "Family circumstances"(=2.0) and "Child care"(=4.0). For males the largest percentage(20%) is the reason "Studying / going to continue the education"(=1.0) , "No hope to find a suitable job" (=9.0) and "Illness / injury / incident" (=3.0). Thus, we can conclude that for females the reason for not looking for a job is family circumstances, child care or no hope for finding a suitable job, whereas for males the largest reason is studying, no hope to find a suitable job or incident.



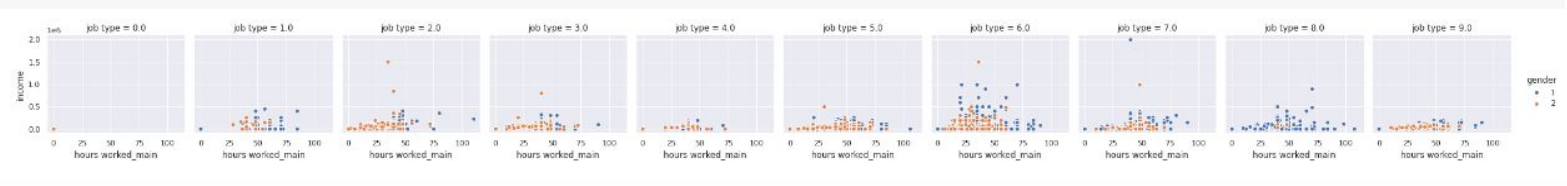
Income Distribution

This chart visually shows the relationship between 2 variables: gender (1=Male, 2=Female) and income class. For the income class, we have the following responses:

- 1.0 = Up to 55 000 AMD
- 2.0 = 55 000 AMD
- 3.0 = 55 001 - 110 000 AMD
- 4.0 = 110 001 - 220 000 AMD
- 5.0 = 220 001 - 440 000 AMD
- 6.0 = 440 000 - 600 000 AMD
- 7.0 = 600 001 - 700 000 AMD
- 8.0 = Refused to answer
- 9.0 = Do not know / difficult to answer

Thus, it is concluded that the majority of male respondents have the salary range **55 001 - 110 000 AMD** or answered do not know / find it difficult to answer. The third salary range that the male respondent answered to have is 110 001 - 220 000 AMD. For females, the biggest percentage answered to have the salary range 55 001 - 110 000 AMD. Interestingly, **more male respondents refused to answer the question than female respondents.**

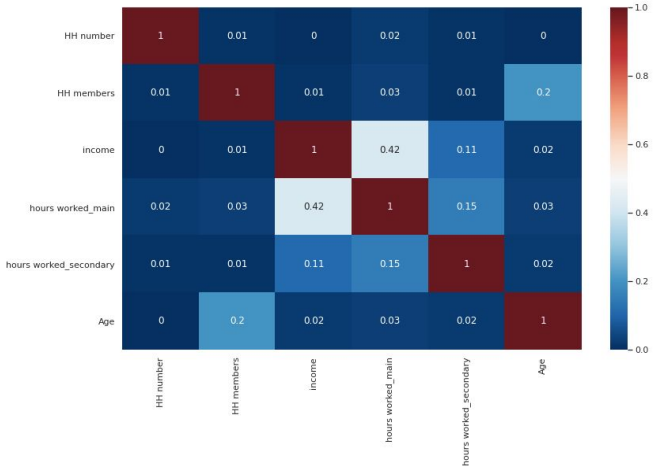
Hours Worked Distribution



This scatter plot helped us to visually see in which job type most of the male and female respondents are engaged. We have 10 categories, that correspond to the job types and are as follows: *0.0 = No response; 1.0 = Legislators, senior officials, managers; 2.0 = Professionals; 3.0 = Technicians professionals; 4.0 = Clerks; 5.0 = Service & sales workers; 6.0 = Skilled agricultural workers; 7.0 = Craft workers; 8.0 = Operators & assemblers; 9.0 = Elementary occupations.*

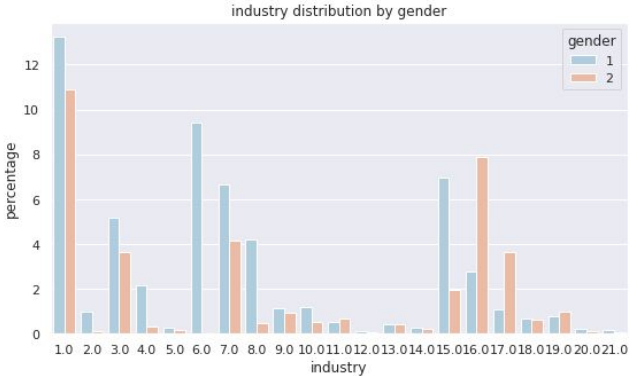
Thus, **mainly the respondents are skilled agricultural workers. For female respondents, the big part is professionals.**

Relationship Between the Numeric Variables



The heatmap shows the relationships between the variables. Thus, the **amount of income is connected with the hours worked at the main workplace(**=0.42). Next, **the age describes the household members variable**(0.2). Lastly, **hours worked in the main workplace are connected with the hours worked in the secondary workplace**(=0.15).

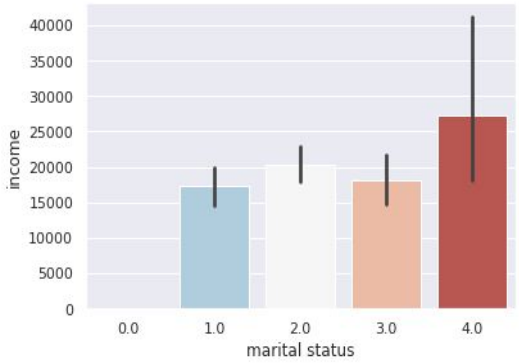
Industry and Job Type Distribution



Here we have the visualization of the variables gender and industry. The biggest number of respondents are working in **agriculture**, hunting and forestry(=1.0) type of economy and the part of males is bigger than females. The following type that has a big amount of responses is **construction** (=6.0) where all the workers are males based on the responses. For the **females**,

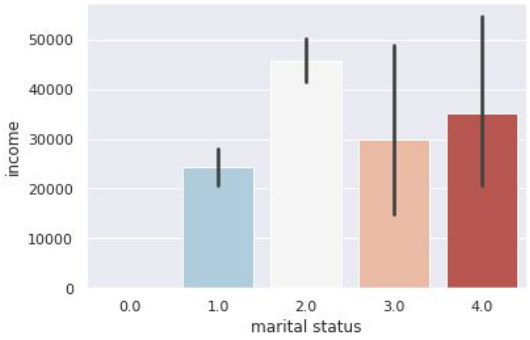
bigger percentage has th**e private households with employed persons**(=16.0)

Marital Status and Income, Based on Gender

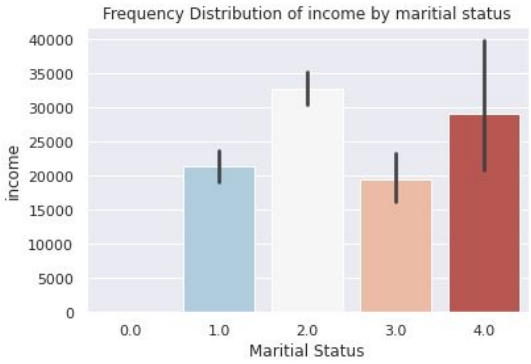


1.0-Never married ; 2.0-Married; 3.0-Widowed; 4.0-Divorced/Separated

The following diagram shows the frequency distribution of salaries for the employed **Female** households of the data. Thus, out of the data, the **divorced / separated Females are the ones to get comparably higher income.**



The diagram on the left shows the frequency distribution of salaries for the employed **Male** households of the data. Thus, the data **shows that the married men are the ones to get the highest salaries**.



The following diagram on the right shows the frequency distribution of salaries for all of the employed households of the data. From the data it can be seen that **never married, married and divorced / separated households have approximately alike numbers as their incomes, and only widowed households have the lowest ones.**

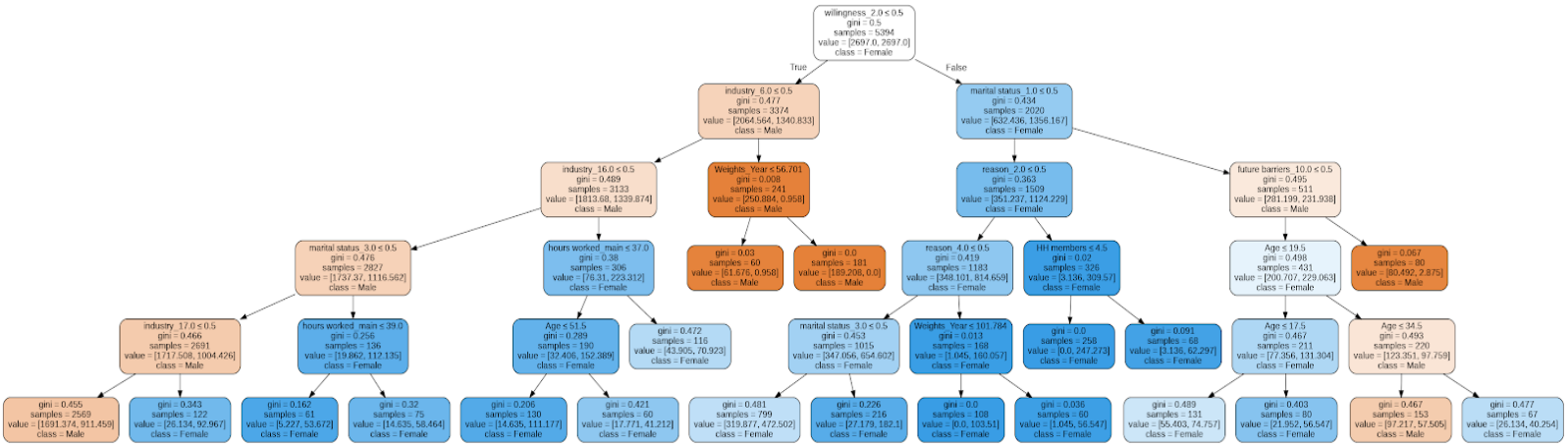
Methodology

Since we have a big data of 7193 observations we implemented a decision tree as well as a random forest classifier, first without manually making the dummy variables for the categorical variables. Further we run logistic regression on features that were identified to be important by DT and RF classifiers. We will compare the three models with mean cross-validated ROC AUC scores.
To implement classification and regression trees and logistic regression, first we split our data into train and test datasets with 75% and 25% proportions respectively using sklearn train_test_split. We use the whole data (labor) and separate the target variable of gender from other attributes and then split the resulting dataset. We also eliminate the variable for the number of households, since we find it useless.

Decision Tree on transformed data

First, we apply for a fully grown decision tree, without tuning the hyperparameters. From the results of the DT model we got 0.999 and 0.71 ROC AUC scores on train and test sets, respectively. So, we have an overfitting model which means we cannot generalize the trained dataset. To deal with the overfitting issue

we will tune the hyperparameters using GridSearch. With GridSearch cross validation we search for the most optimal parameters and fit our model with the latter. Consequently, we extract a mean 5-fold ROC AUC score of 0.75 for not transformed DT, which is a good indicator. Then, we visualize the resulting tree.




```
feature_importance
```

willingness_2.0	0.294166
marital status_3.0	0.116991
industry_6.0	0.116621
marital status_1.0	0.103156
industry_16.0	0.095877
industry_17.0	0.059500
reason_2.0	0.058188
future barriers_10.0	0.048936
reason_4.0	0.045245
Age	0.024356
hours worked_main	0.019726
income	0.013580
education level_10.0	0.002217
education level_5.0	0.000678
HH members	0.000533
marzes_8	0.000169
education level_7.0	0.000052
area_2	0.000008
reason_11.0	0.000000
job type_2.0	0.000000

dtype: float64

According to the DT classifier model, each internal node of the tree corresponds to a feature, and each leaf node represents a gender class that is either Male or Female. The most significant features are placed on the root of the tree. Accordingly, we can see that in the top of the decision tree we have the willingness 2.0, which indicates those who do not actively look for a job in a timeframe of the survey following with marital status 3.0 which is 'widowed'. Afterwards we have the features of industry 6.0 which corresponds to the Construction as the type of economic activity. The marital status 1.0 indicates those respondents who were never married.

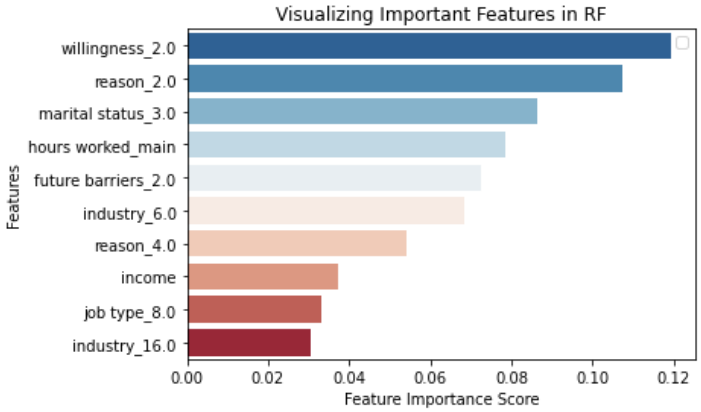
Given that the willingness 2.0 (respondent does not look for a job) is less than 0.5 (True), the respondent looks for a job and wants to work. Given that industry 6.0 (Construction) is less than 0.5(True), the respondent does not work in the construction field. Given that industry 16.0 (Private households with employed persons) is less than 0.5 (True), the respondent works in this field. Given that hours worked at the main job (how many hours they spend in their primary work) is less than 41.5 (True) respondents spend less than 41.5 hours in their primary work. Given that income < 111000 is True, the respondent has a salary less than 111000. All these lead us to conclude that the respondent is Female.

Random Forest on transformed data

For potentially better accuracy, we will implement a random forest classifier. Random forest uses several decision trees by random sampling of the training data, and by predictions voting or bagging, selects the

best option. Since the random sampling is done by replacement (bootstrapping), the forest reduces the overall variance.

We will use already defined parameters for Decision Tree Classifier in GridSearch to obtain best parameters for Random Forest. The mean 5-fold ROC AUC score for the RF classifier is 0.8, which outperforms DT. As per results of the classification we have the following feature importance order:



Logistic Regression based on Random Forest results

For logistic regression we have decided to choose only the features that RF classified as important since we have many columns with diverse categories. We perform the regression using each important variable indicated in the above chart and then evaluate the results.

```
                         Results: Logit
================================================================
Model:                Logit              Pseudo R-squared:   inf
Dependent Variable:   gender_2           AIC:                inf
Date:                 2020-05-15 17:39   BIC:                inf
No. Observations:     5394               Log-Likelihood:     -inf
Df Model:             10                 LL-Null:            0.0000
Df Residuals:         5383               LLR p-value:        1.0000
Converged:            1.0000             Scale:              1.0000
No. Iterations:       10.0000
----------------------------------------------------------------
                     Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
----------------------------------------------------------------
const               -0.2275   0.0646 -3.5214 0.0004 -0.3542 -0.1009
willingness_2.0      0.4406   0.0845  5.2125 0.0000  0.2749  0.6062
hours worked_main   -0.0024   0.0019 -1.2852 0.1987 -0.0062  0.0013
reason_2.0           2.9253   0.7426  3.9391 0.0001  1.4698  4.3809
reason_4.0           4.8914   1.0045  4.8694 0.0000  2.9226  6.8603
marital status_3.0   2.0333   0.1659 12.2589 0.0000  1.7082  2.3584
income              -0.0000   0.0000 -2.9591 0.0031 -0.0000 -0.0000
future barriers_2.0  2.1214   0.9118  2.3265 0.0200  0.3342  3.9085
industry_6.0        -5.1918   1.0051 -5.1655 0.0000 -7.1618 -3.2218
job type_8.0        -2.8927   0.4237 -6.8271 0.0000 -3.7232 -2.0623
industry_16.0        1.5036   0.1426 10.5443 0.0000  1.2241  1.7831
================================================================
```

In the main output above, we can see that all the predictor variables are significant except “hours worked”, since the p-values are smaller than 0.05 at the 95% confidence interval.

$$log(P(Female = yes)/1 - P(Female = yes)) = -0.2275 + 0.44willingness_2 + 2.93reason_2 + ... + 1.5industry_{16}$$

Given the features mentioned above, we can calculate the probability of the gender participation in the labor force.

Main output summary suggests that willingness to look for a job or try to launch own business increases the log odds of women employment by 0.44 times, while family circumstances and child care affect the women's participation in the labor force by tripleing and increasing the log odds five times respectively. Being widowed increases the chances twice, while the odds are decreasing for women in the construction industry and job type of operators and assemblers. Surprisingly, income does not change the odds for women, however this might be a pitfall of the model.

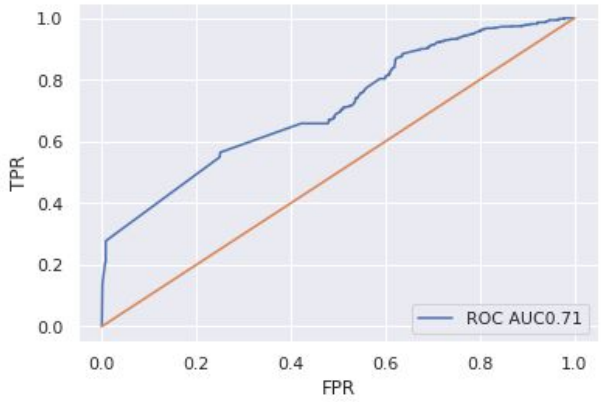
Logit Marginal Effects

Dep. Variable:	gender_2					
Method:	dydx					
At:	overall					
	dy/dx	std err	z	P>|z|	[0.025	0.975]
willingness_2.0	0.0845	0.016	5.276	0.000	0.053	0.116
hours worked_main	-0.0005	0.000	-1.286	0.198	-0.001	0.000
reason_2.0	0.5607	0.142	3.951	0.000	0.283	0.839
reason_4.0	0.9376	0.192	4.886	0.000	0.562	1.314
marital status_3.0	0.3898	0.030	12.921	0.000	0.331	0.449
income	-2.755e-07	9.28e-08	-2.969	0.003	-4.57e-07	-9.36e-08
future barriers_2.0	0.4066	0.175	2.328	0.020	0.064	0.749
industry_6.0	-0.9952	0.192	-5.185	0.000	-1.371	-0.619
job type_8.0	-0.5545	0.080	-6.901	0.000	-0.712	-0.397
industry_16.0	0.2882	0.026	11.058	0.000	0.237	0.339

According to Marginal Effects summary, if the person has no willingness to work there is on average 8.4 % more probability that they are women. Family circumstances and child care affect women participation on average by 56% and 93.8% respectively. Being widowed improves women participation around 40%. Future assumptions about family decreases the chances of an employee to be women by 41%.

There is on average 99.5 % lower probability of women being in construction industry and 55.5% for working as operators & assemblers

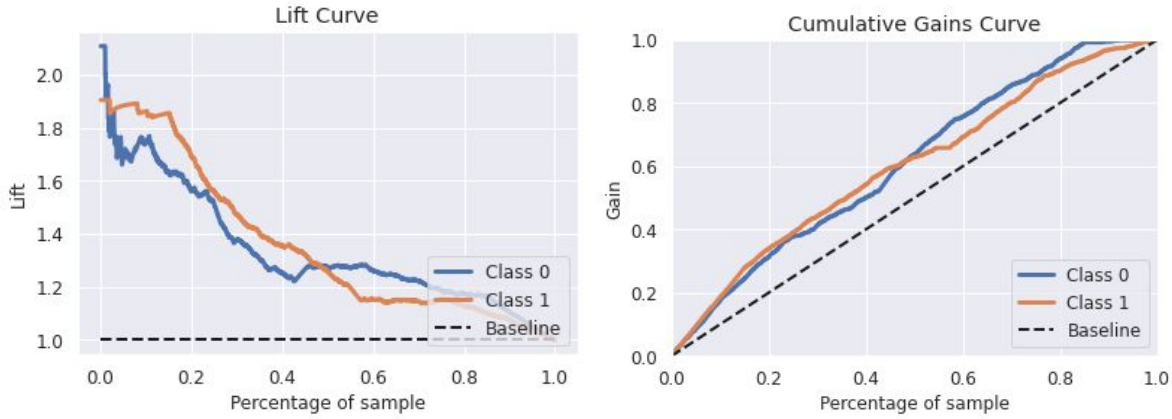
Model Evaluation



	Accuracy	ROC_AUC	Recall
Training set	0.566926	0.678472	0.611230
Testing set	0.595331	0.711194	0.658201

Logistic regression shows a pretty high mean 5-fold ROC AUC score of 0.68. Accuracy is shown in the table above as per result of test and train sets.

Recall shows if there are women in the test set, the model will be able to identify them 65.8% of the time.



Lift Curve and Cumulative Gain curves both indicate how good the model is relative to a completely random one. The Gain curve shows that if our target is to have 0.8 cumulative gain, more than 70% of the sample should be considered.

Conclusion

Overall, female labour force participation is highly affected by marital status, family circumstance and childcare. Women are abstained from participating in the labor force given their expectations of future family and there is on average 8% lower willingness for women to participate in the labor force. There is low involvement of women in male dominated sectors such as the construction industry, as well as there is a low probability of female participation in jobs such as operators or assemblers. Our analysis can be helpful for further consideration of gender gap resolvement in Armenia. Taking into account the results, the RA government may consider incentivizing married women with children to enter the labour market and participate in the economic activities.

References

1. **Link to code for the project:** https://colab.research.google.com/drive/1OG8n3LL5EsZvCPKLmisivrjKzWAaR6Zk?usp=sharing
2. National Statistical Service Republic of Armenia (n.d.) Labour Force One-Off Survey Questionnaire. Retrieved from https://www.armstat.am/file/article/rep_ashx_09e_8.pdf
3. State Council of Statistics of RA (n.d.) Instructions for filling in the labour force survey (LFS) questionnaire. Retrieved from https://www.armstat.am/file/doc/99506713.pdf
4. Statistical Committee of the Republic of Armenia. (n.d.). Labour Force Survey anonymised microdata database and questionnaire (by household's members). Retrieved from https://www.armstat.am/en/?nid=212
5. Statistical Committee of the Republic of Armenia. (n.d.). Report on labour force and informal employment in Armenia (on the results of one-off sample survey). Retrieved from https://www.armstat.am/en/?nid=80&id=1008&fbclid=IwAR3sGI0Q-yH9kLDqgYajk9IX2vpt4MEpUYKgTJaaoqMoSdwBlqy-HHJIdmo