2024

Economic Fluctuations and Job Satisfaction in Spain

DATA SCIENCE PROJECT

UJVARA FETOSHI

Table of Contents

INTRODUCTION **2**

EXPLANATORY DATA ANALYSIS AND DATA PREPARATION **3**

EXPLANATORY DATA ANALYSIS 3

Data File Understandability *3*

Identifying and Visualising Variables *4*

DATA CLEANING 7

DATA CATEGORIZATION 8

MODEL GENERATION **9**

MODEL 1 9

MODEL 2 12

MODEL 3 15

MODEL SELECTION **18**

MODEL INTERPRETATION **21**

COEFFICIENTS SIGNS 21

ODD RATIO 23

MARGINAL EFFECT 26

SCORECARD 28

CONCLUSION **29**

Introduction

The focus of this project is to analyse **How economic fluctuations influence job satisfaction among employees in Spain, considering the impact of job security and work-life balance.**

As a step to reach this goal, data from the European Social Survey was collected, from round 10, with respect to the year 2020. The main categories from each of the data to build the explanatory variables include: Media use and trust; Politics; Subjective well-being and social exclusion; Gender, age and household composition; Socio-demographic profile, Family, work and well-being and Personal and social well-being. The dependent variable under analysis belongs to the category of Personal and social well-being, and can be described as 'How satisfied are you in your main job'.

The methodology followed throughout the project respected the eight stages for building predictive econometric models, namely, Exploratory Data Analysis and Data Preparation, Creation of a Richer Set of Covariates, Variable Selection, Multicollinearity Issue, Model Estimation, Model Validation, Model Selection and Interpretation and Prediction. The current report will carefully detail each of the previously mentioned steps in the context of the Spanish reality regarding job satisfaction.

Explanatory Data Analysis And Data Preparation

Explanatory Data Analysis

As a starting point, all the data regarding Spain from the year 2020 was extracted from the European Social Survey portal. To initiate our preliminary analysis, we refined column names using the information from an HTML documentation. After that, we created a histogram of the percentage of missing values, which helped in the identification of columns with significant missing values.

The original dataset was then filtered to retain only the chosen independent variables and the dependent variable, 'How satisfied are you with your main job?' The codes representing non-applicable, refusal, don't know, and no answer responses in each variable were replaced with NaN for consistency. To ensure data quality, rows with null values in the dependent variable were removed, and a histogram was plotted to visualize the distribution of missing values across variables. A threshold of 10% was set, leading to the identification of columns with less than 10% missing values as final non-null columns. By following this variable selection process, we focused the analysis on key variables and made sure that we maintained data integrity.

Data File Understandability

In preparation for explanatory analysis and improving data interpretability, the JSON file containing variable category mappings was loaded. Numeric values in certain columns were mapped to corresponding categories, creating a more intuitive representation of the data. Furthermore, the column names were transformed for better clarity. Unnecessary stopwords were removed, and names were converted to lowercase. Special characters were replaced, and the resulting labels were standardized, which contributed to a more understandable dataset.

Identifying and Visualising Variables

Furthermore, a comprehensive data pre-processing procedure was executed. For numerical covariates, we identified and replaced outliers with 'nan' values. Descriptive statistics were initially visualized, demonstrating the distribution of each covariate before the outlier removal process. The outliers, identified using the Interquartile Range (IQR) method, were then replaced with 'nan' values. The resulting dataset was visualized again, highlighting the impact of outlier removal on the covariate distributions.

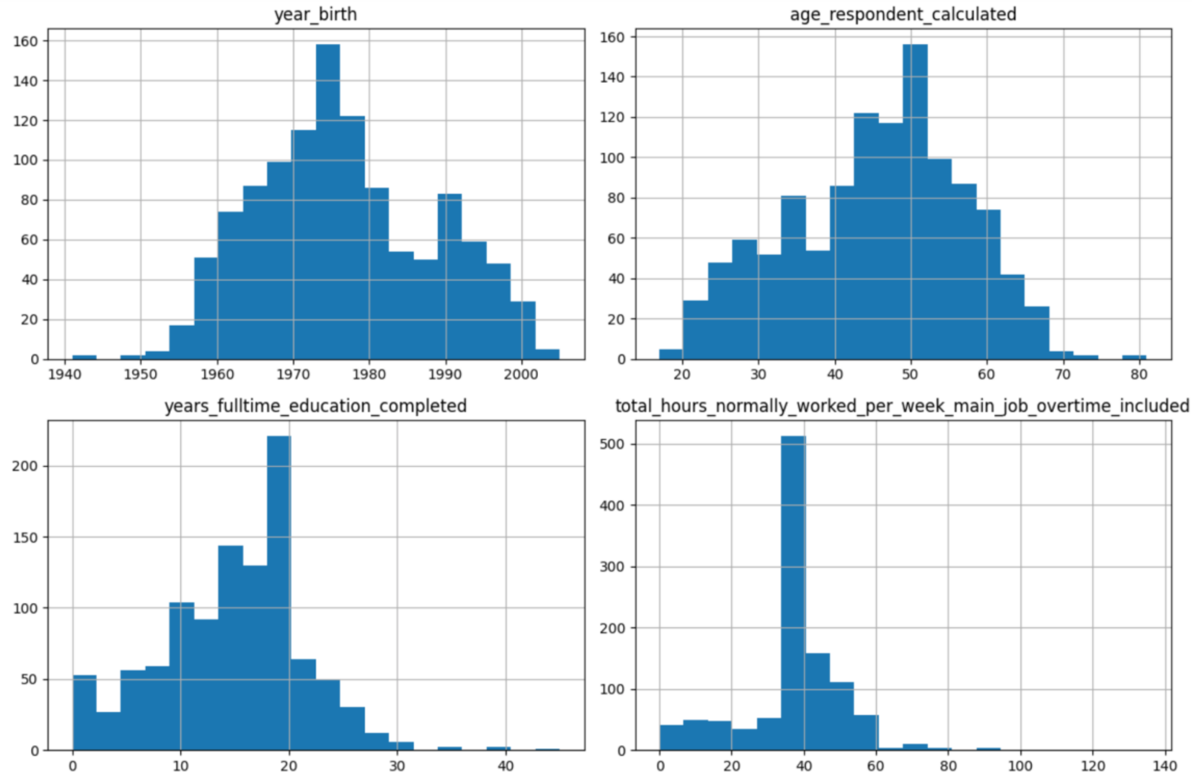


Figure 1 - Histogram of numerical features before removal of outliers

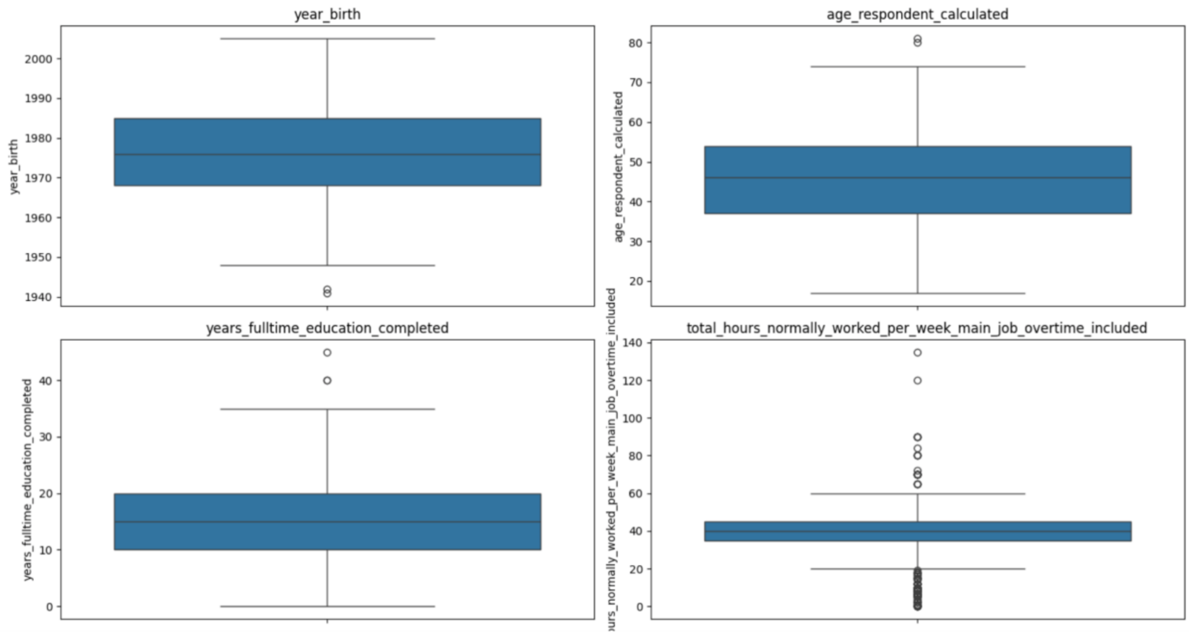


Figure 2 - Boxplot of numerical features

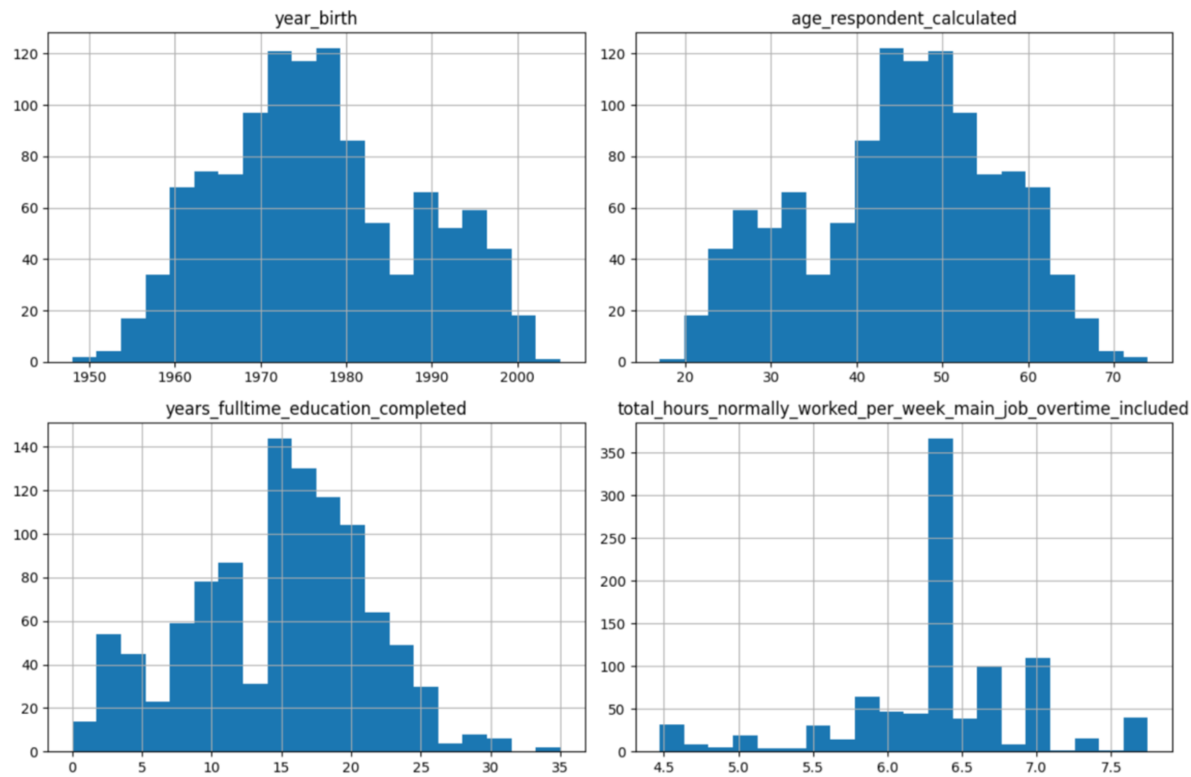


Figure 3 - Histogram of numerical features after removal of outliers

Furthermore, we explored the missing values in numeric covariates, providing a detailed summary of the percentage of missing values for each variable after the outlier removal process.

For nominal and ordinal covariates, frequency tables were generated to display the distribution of each category along with the percentage of missing values. We continued with exploratory data visualizations to gain insights into the distribution of key variables. The initial visualizations displayed count plots for each categorical variable in the dataset, providing an overview of the distribution of values. Furthermore, a detailed exploration of the distribution of job satisfaction scores was presented, showcasing both the count distribution and the calculated mean and median values, which pointed to the value 7 in both cases.

We concluded that a binary classification of job satisfaction should be introduced, considering scores above 7 as "satisfied" and scores equal to or below 7 as "not satisfied." We visualised the binary job satisfaction distribution which offered a better perspective on the proportion of satisfied and unsatisfied individuals.

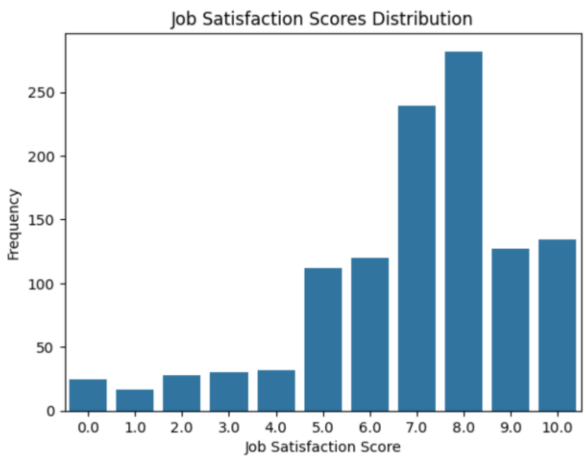


Figure 5 - Distribution of Job Satisfaction Score

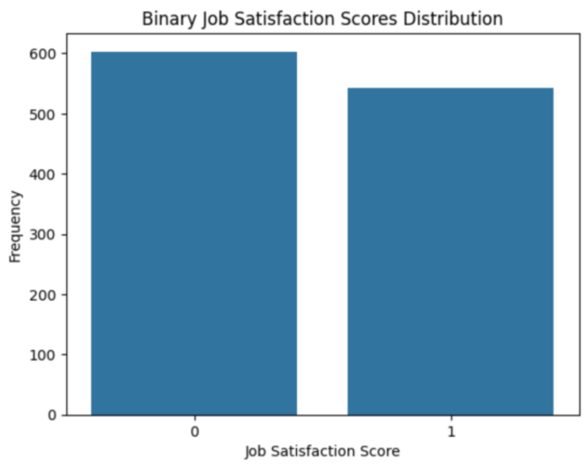


Figure 4 - Distribution of Binary Job Satisfaction Score

Additionally, kernel density plots were generated for selected variables, comparing the distribution between satisfied and unsatisfied groups. These density plots provided a better view of the variables' impact on job satisfaction, offering insights into potential patterns and differences.

Data Cleaning

Furthermore, we continued with the data cleaning process to handle missing values in the dataset and identify dummy variables that are significantly correlated with the dependent variable.

For numeric variables, a chi-square test of independence was conducted between these dummy variables and the binary job satisfaction variable. Dummies with a p-value less than 0.05 were retained, indicating a significant relationship with job satisfaction. The numeric variables associated with these dummies were filled with zeros for missing values, while others were replaced with the mean. We have followed the same procedure for ordinal and nominal variables.

Data Categorization

In the data categorization process for nominal variables, we calculated a chi-square test of independence for a set of covariates with the binary job satisfaction variable. For each of these retained nominal variables, a natural breaks clustering approach was employed based on the odds ratio of their class levels. Specifically, the Jenks Natural Breaks method was utilized to categorize these class levels into three clusters or groups, ensuring similar odds within each cluster.

The resulting cluster mappings for each nominal variable were stored in a dictionary (**jenksnaturalbreaks_mapping**). These mappings provide a linkage between the original class levels and their corresponding clusters/groups. The original nominal variables were then updated to reflect their cluster assignments based on this categorization. Additionally, we mapped clusters for specific nominal variables such as "main_source_household_income," "occupation_isco," "fathers_occupation_respondent_," and others. Each original class level in these variables was replaced with its corresponding descriptive category within the assigned cluster. Finally, we converted the categorical variables into binary indicators to provide cluster information represented as dummy variables. Finally, a summary was provided to confirm the inclusion of all relevant variables, covering numeric, ordinal, and nominal categories in the dataset.

Model Generation

After carefully proceeding with the exploratory data analysis and data preparation, we have proceeded with the model generation part. At this stage, 3 models were created:

- Model 1 - binned model with only categorical variables
- Model 2 - model with interaction variables
- Model 3 - model with original variables with transformations

For each model, a richer set of covariates was created, comprising non-linear effects and interaction effects. Followed by the variable selection and the handling of the multicollinearity issue. Afterwards, each model is estimated and validated. All of these stages can be described below.

Model 1

The objective for creating a richer set of covariates was to augment the feature space allowing the binned model to better capture patterns and relationships within the data. We started this process by binning numeric features such. We extended the binning process to nominal variables, where dummy variables were generated for each interval of the binned numeric features. Ordinal columns underwent a similar transformation, being categorized into five bins using the `pd.cut` method. Dummy variables were subsequently created for each bin. To ensure model stability, the least frequent bin in each ordinal column is dropped, mitigating the risk of multicollinearity. Finally, the true response variable, denoted as 'binary_job_satisfaction,' was extracted from the preprocessed dataset.

For the variable selection part, we employed the Chi-Square test of independence to evaluate the statistical significance of correlations between the binary job satisfaction variable (y) and each explanatory variable in the dataset. We further retained the variables with p-values below 0.05 in the 'reduced_dum_binned' dataset. This step ensured the inclusion of variables exhibiting meaningful relationships with job satisfaction, enhancing the model's predictive power and interpretability.

We continued our work by examining multicollinearity among variables in the reduced dataset. By calculating the Spearman correlation coefficient for all variable pairs, we identified and removed variables with correlation coefficients exceeding a threshold of 0.7. This process ensured that highly correlated variables, which can affect model interpretability and stability, are excluded. Additionally, we verified multicollinearity using the Variance Inflation Factor (VIF), providing insights into potential issues related to high collinearity. The VIF values for each variable were displayed in the 'vif_data', aiding in the assessment of multicollinearity within the dataset.

For the model estimation, the initial model was fitted and evaluated for convergence using the 'convergence' function, providing insights into the success of the optimization process. We computer and displayed the information criteria through the 'informationcriteria' function. Furthermore, we printed summary statistics of the initial model using 'binned_model_res.summary()'. Variables with significant impact on job satisfaction (with p-value < 0.05) were identified. Subsequently, a new logistic regression model was estimated, exclusively considering these significant variables. The process was repeated, and the final model was fitted, providing a comprehensive summary and highlighting variables significantly associated with job satisfaction.

Finally, the validation code for the logistic regression model assessed convergence, information criteria, and the global null hypothesis, which can be seen below.

Information Criteria

Criterion	Intercept Only	Intercept and Covariates
AIC	1587.551	1235.861
BIC	1592.595	1336.742
-2LogL	1585.551	1195.861

Model Convergence Status

Optimizer	newton
Starting parameters [0.0, 0.0, 0.0, 0.0, 0.0, 0.0,	0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
Max. iterations	35
Tolerance rate	1e-08
Req. iterations	6
Convergence Status	True

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr>ChiSq
Likelihood Ratio Test	389.6897	19	0.0
Wald	236.962	19	0.0

This way, we checked if the model has converged successfully, computed information criteria for model evaluation, and evaluated the overall significance of predictors. From the results, we can see that according to the values of AIC, BIC and Likelihood Ratio Test, the model with coefficients has significantly lower values for each information criterion, which suggests that adding coefficients improves the model's fit. These steps collectively ensured the model's reliability and effectiveness in predicting job satisfaction.

All of the steps above can be reflected on the model specification below:

Job satisfaction among a sample of 916 individuals from Spain.

$$y_i = \begin{cases} 1 \text{ if individual i is satisfied with their job} \\ 0 \text{ otherwise.} \end{cases}$$

The model reads as,

π_i = Λ(β0 + β1 * how_emotionally_attached_country_Low$_i$ +
β_2 * how_emotionally_attached_country_Medium$_i$ +
β_3 * how_happy_Low$_i$ +
β_4 * how_happy_Medium$_i$ +
β_5 * how_happy_Very_High$_i$ +
β_6 * line_manager_gives_workrelated_help_likely_Very_Low$_i$ +
β_7 * line_manager_gives_workrelated_help_likely_High$_i$ +
β_8 * line_manager_gives_workrelated_help_likely_Very_High$_i$ +
β_9 * line_manager_supports_employees_balancing_work_much_Medium$_i$ +
β_{10} * number_people_living_regularly_member_household_Very_Low$_i$ +
β_{11} * partner_family_fed_pressure_job_often_Very_Low$_i$ +
β_{12} * partner_family_fed_pressure_job_often_Low$_i$ +
β_{13} * too_tired_work_enjoy_things_like_home_often_Very_High$_i$ +
β_{14} * doing_last__days_unemployed_actively_looking_job_0_Not_marked$_i$ +
β_{15} * responsible_supervising_employees_1_Yes$_i$ +
β_{16} * occupation_isco_Drivers_Software_Developers_and_Miscellaneous$_i$ +
β_{17} * what_type_organisation_work_worked_Miscellaneous$_i$ +
β_{18} * industry_nace_rev_Transportation_and_Miscellaneous$_i$ +
β_{19} * region_Northern_and_Central_Regions$_i$)

Where:

π_i: Probability of individual i being satisfied with their job.

Λ: The logistic function applied to binned/nominal covariates.

β_0: Intercept.

β_1 to β_{19}: Coefficients for each of the predictors.

Predictors include various aspects like emotional attachment to the country, happiness levels, interaction with line managers, household dynamics, work-life balance pressures, tiredness levels due to work, employment status, supervisory responsibilities, occupation categories, type of organisation, industry sectors, and regional factors.

This model was generated from multiple iterations on all binned and categorical variables after thorough variable selection, based on multicollinearity and variable significance.

Model 2

We continued our work by working on the second model which was a model with the interaction variables. We started from enriching the set of covariates by introducing non-linear effects such as square and cube transformations for numeric variables. Additionally, two-level interaction effects were created for both numeric and ordinal covariates which included combinations of original features, providing a more comprehensive set of predictors. Similarly, interaction effects were constructed for numeric and nominal covariates, revealing complex relationships that contribute to the dependent variable.

Moreover, we calculated the total number of variable candidates resulting from these enrichments. It assesses the K/N ratio, where K represented the number of candidate variables, and N was the sample size. Since the ratio was above 15%, we concluded that there is a necessity to perform variable selection and check for multicollinearity among the predictors.

We continued with the variable selection. The key difference between this model's variable selection and the previous one lied in the specific statistical tests applied to each type of covariate. While the first model utilized the chi-squared test for all categorical variables, this code employed the Spearman rank correlation test for numeric and ordinal covariates, and the chi-squared test exclusively for nominal covariates. This tailored approach acknowledged the distinct nature of each covariate type and ensured appropriate statistical testing for variable selection.

The main difference in the approach to multicollinearity issue compared to the previous model lies in the computation of VIF. Here, we explicitly calculated VIF for the selected set of covariates, which included numeric, ordinal, and nominal variables. We set the threshold of 10 to identify and exclude variables with a high degree of multicollinearity. Afterwards, we computed the K/N ratio, providing insights into the proportion of selected covariates relative to the sample size. In this case, the K/N ratio was 8.726 which proves that we have successfully handled the issue.

We followed with estimating the logistic regression model using the Probit link function, performed in three iterations. After each iteration, variables were categorized based on their p-values. Those with p-values greater than or equal to 0.05 are deemed insignificant, while those with p-values less than 0.05 were considered significant. After that, the model was validated similarly to the first model, with the result that can be seen below.

Information Criteria

Criterion	Intercept Only	Intercept and Covariates
AIC	1587.551	1293.307
BIC	1592.595	1333.659
-2LogL	1585.551	1277.307

Model Convergence Status

Optimizer	newton
Starting parameters	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
Max. iterations	1000
Tolerance rate	1e-08
Req. iterations	6
Convergence Status	True

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr>ChiSq
Likelihood Ratio Test	308.2434	7	0.0
Wald	200.0594	7	0.0

In this model, we can see that according to the values of AIC, BIC and Likelihood Ratio Test, adding coefficients still improved the model's fit. However, we can also see that the second model has higher values for each information criteria, which suggests that Model 1 is a better-fitting model.

The model specification can be defined as:

Job satisfaction among a sample of 916 individuals from Spain.

The model reads as,

$$y_i = \begin{cases} 1 \text{ if individual i is satisfied with their job} \\ 0 \text{ otherwise.} \end{cases}$$

$\pi_i = \Phi(\beta_0 + \beta_1$ * partner_family_fed_pressure_job_often$_i$ +
β_2 * how_emotionally_attached_country$_i$ +
β_3 * how_happy$_i$ +
β_4 * too_tired_work_enjoy_things_like_home_often$_i$ +
β_5 * line_manager_gives_workrelated_help_likely$_i$ +
β_6 * line_manager_supports_employees_balancing_work_much$_i$ +
β_7 * gender_2_Female_region_Basque_Country_Navarra_and_Extremadura$_i$ +
B_8 * fathers_occupation_respondent_Unskilled_Workers_what_type_organisation_work_worked_Miscellaneous$_i$)

Where:

π_i: Probability of individual i being satisfied with their job.

Φ: The cumulative distribution function of the standard normal distribution, reflecting the Probit model.

β_0: Intercept.

β_1 to β_8: Coefficients for each of the predictors and interaction terms.

The predictors include various individual and work-related aspects, such as pressure from partner/family, emotional attachment to the country, happiness levels, tiredness from work, likelihood of receiving work-related help from the line manager, and the line manager's support in balancing work and life.

Interaction terms account for complex relationships between variables, like the combined effect of gender and region, or the father's occupation and the type of organisation where the respondent worked.

This model was estimated using Maximum Likelihood Estimation (MLE) and focuses on the influence of specific interaction terms along with primary predictors on job satisfaction, indicating a more nuanced approach to understanding the factors contributing to job satisfaction among employees. The model was generated after several iterations testing all interaction terms and reducing using multicollinearity checks and significance of variables.

Model 3

Finally, we worked on the third model, which this time had original variables with transformations. In the creation of the richer set of covariates for the third model, several transformations and non-linear effects were applied to the original numerical features. These transformations included squaring, cubing, taking the logarithm, and exponentiating each numerical feature. While generating these transformations, special attention was given to handle potential issues, such as dropping columns that could result in infinity or NaN values. After creating these non-linear effects, the resulting features were concatenated with the original numerical features. Subsequently, the entire dataset, including both numerical and nominal features, underwent standardization using a Standard Scaler to ensure that different features are on a comparable scale. As a result, the third model, with its richer set of covariates and non-linear effects, differs significantly from the previous models in the complexity and diversity of features introduced through these transformations.

In the variable selection process, the Chi-square test of independence was employed to evaluate the significance of each feature in relation to the target variable since this test was suitable for categorical data. In addressing multicollinearity concerns we examined the features using the Variance Inflation Factor (VIF). The outcome of this analysis was a reduced set of features, which included 26 columns, demonstrating lower multicollinearity based on the VIF criterion.

Furthermore, we initiated model estimation for the third model. After fitting the initial logistic regression model, convergence was checked. A second iteration dropped insignificant variables, leading to a refined logistic regression model. We finalised with model validation, which has a result that can be seen below.

Information Criteria

Criterion	Intercept Only	Intercept and Covariates
AIC	1587.551	1302.841
BIC	1592.595	1343.194
-2LogL	1585.551	1286.841

Model Convergence Status

Optimizer	newton
Starting parameters	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
Max. iterations	35
Tolerance rate	1000
Req. iterations	5
Convergence Status	True

According to this result, we can see that each information criteria was higher for the Model 3 than for the previous models. Therefore, we can conclude that Model 1 is the best-fitting model out of the models we have estimated.

The model specification can be defined as:

Job satisfaction among a sample of 916 individuals from Spain.

$$y_i = \begin{cases} \text{1 if individual i is satisfied with their job} \\ \text{0 otherwise.} \end{cases}$$

The model reads as,

$$\begin{aligned}
&\pi_i = \Lambda(\beta_0 + \beta_1 * \text{parent_lives_household_1_Yes}_i + \\
&\beta_2 * \text{region_Basque_Country_Navarra_and_Extremadura}_i + \\
&\beta_3 * \text{region_Northern_and_Central_Regions}_i + \\
&\beta_4 * \text{partner_family_fed_pressure_job_often}_i + \\
&\beta_5 * \text{how_happy}_i + \\
&\beta_6 * \text{too_tired_work_enjoy_things_like_home_often}_i + \\
&\beta_7 * \text{line_manager_gives_workrelated_help_likely}_i + \\
&\beta_8 * \text{line_manager_supports_employees_balancing_work_much}_i)
\end{aligned}$$

Where:

π_i: Probability of individual i being satisfied with their job.

Λ: The logistic function, indicating the Logit model is used to model job satisfaction

β_0: Intercept.

β_1 to β_8: Coefficients for each of the predictors, reflecting the impact of living with parents, regional factors, pressure from partner/family, happiness levels, tiredness from work, likelihood of receiving work-related help from the line manager, and the line manager's support in balancing work and life on the probability of observing the binary outcome.

The model, estimated using Maximum Likelihood Estimation (MLE), indicates how various personal and work-related factors influence the likelihood of the binary outcome, with positive coefficients indicating an increase in probability and negative coefficients indicating a decrease.The model was generated through multiple iterations of transformations of variables, after checking for multicollinearity and doing variable selection from their significance level.

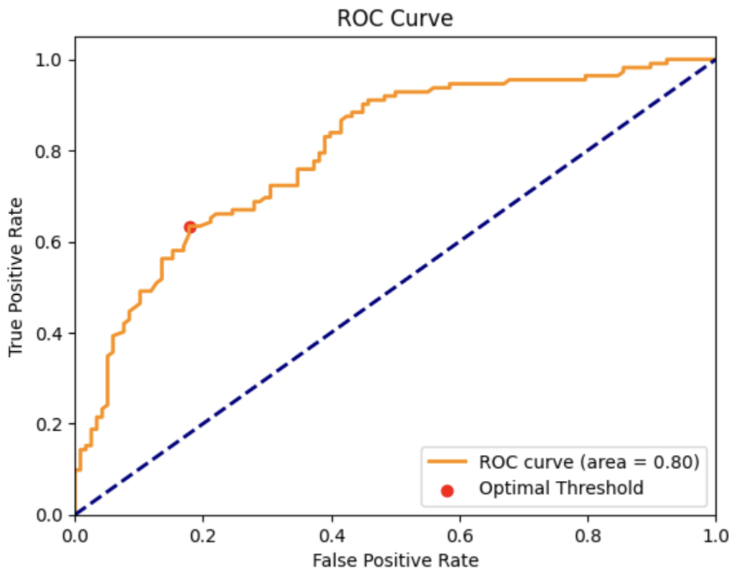
Model Selection

In this section, we perform model selection for each model. As part of this process, we split the dataset into training and testing sets, ensuring balanced proportions of 0s and 1s in both. The logistic regression model was trained on the training set and evaluated on the test set. Key performance metrics, including the confusion matrix, Area Under the Curve (AUC), R2 Efron, and classification accuracy were calculated and displayed.

The optimal threshold for classification was determined using the ROC curve. The ROC curve, AUC, R2 Efron, and the number of prediction errors were visualized and quantified to assess the model's predictive accuracy. Additionally, we extracted relevant summary statistics, such as AUC, AIC, BIC, and R2 Efron for comprehensive model evaluation.

Model 1

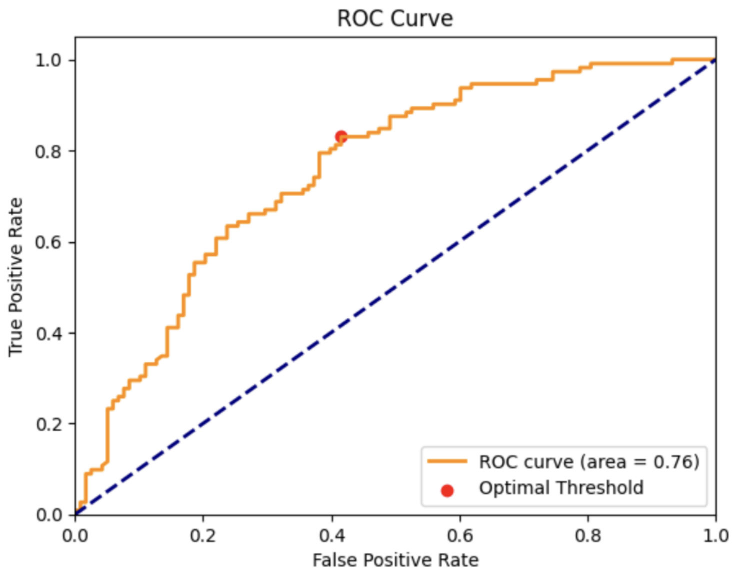
```
M1 Optimal Threshold: 0.590205014188407
M1 AUC: 0.7969506658595642
y_hat    0   1  All
y
0       97  21  118
1       42  70  112
All    139  91  230
```




```
AUC: 0.7970
R2 Efron: 0.2595570010330638
gamma: 0.590205014188407 - number of predictions errors: 63 - percentage of correctly classified: 0.7261
```

Model 2

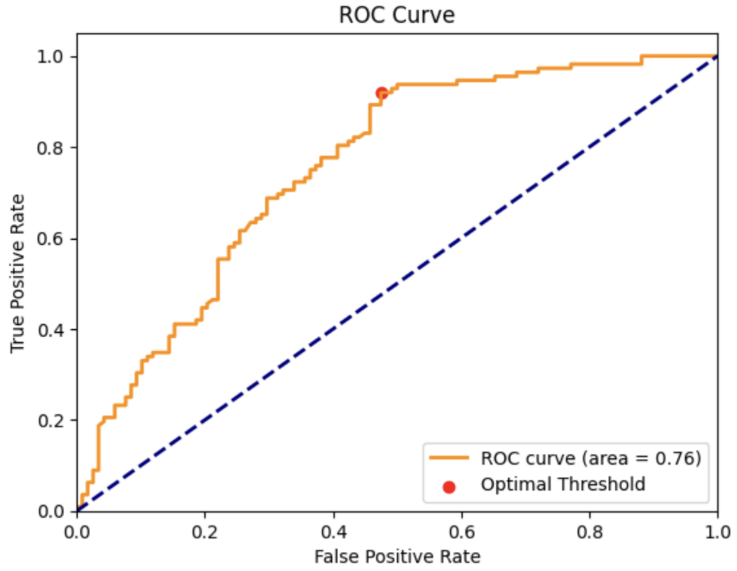
```
M2 Optimal Threshold: 0.378381988368963
M2 AUC: 0.7552209443099274
y_hat   0    1  All
y
0      69   49  118
1      20   92  112
All    89  141  230
```




```
AUC: 0.7552
R2 Efron: 0.18686813243081235
gamma: 0.378381988368963 – number of predictions errors: 69 – percentage of correctly classified: 0.7000
```

Model 3

```
M3 Optimal Threshold: 0.3306328451809499
M3 AUC: 0.7563180992736077
y_hat   0    1  All
y
0      62   56  118
1      10  102  112
All    72  158  230
```




```
AUC: 0.7563
R2 Efron: 0.19577265689763934
gamma: 0.3306328451809499 – number of predictions errors: 66 – percentage of correctly classified: 0.7130
```

Model Comparison

Criterion	Model 1 - Binned Model	Model 2 - Interaction Variables	M3 - Original Variables with transformations
AIC	985.378	1022.947	1037.296
BIC	1081.779	1061.507	1075.856
AUC	79.7%	75.52%	75.63%
R2_Effron	0.2596	0.1869	0.1958
R2	25.36%	20.5%	19.37%

Taking into account all of this information, we can conclude that the optimal threshold is quite similar for the Model 2 and Model 3, while it is significantly higher for the Model 1. Lower threshold increases the number of predicted positives, potentially leading to higher sensitivity but lower specificity, and vice versa. Furthermore, we looked at the values of AUC which measures the model's ability to distinguish between positive and negative cases. We can conclude that Model 1 has the highest AUC of 0.7970 which tells us that the model performs better in distinguishing between classes. Moreover, we analysed the values for R2 Efron, which measures the proportion of explained variation. This value also suggested that Model 1 has a better explanatory power since its value is the highest. Finally, the values for the number of prediction errors and percentage of correctly classified values also showed that Model 1 performs better then Model 2 and Model 3.

Taking into account all of the given information, we concluded that the best performing model was Model 1, the binned model.

Model Interpretation

Coefficients Signs

We continued our work by re-estimating the best model (Model 1) on the full dataset. With this, we were able to interpret the signs of the coefficients:

1. **β^how_emotionally_attached_country_Low: -0.9976 < 0.** An individual with a Low emotional attachment to their country is less likely to report job satisfaction than those with Very High attachment level (reference). This suggests that emotional attachment to one's country positively impacts job satisfaction.
2. **β^how_emotionally_attached_country_Medium: -0.6159 < 0.** An individual with Medium emotional attachment to their country is less likely to report job satisfaction than those with Very High attachment level (reference).
3. **β^how_happy_Low: -0.9723 < 0.** Individuals reporting Low happiness levels are less likely to be satisfied with their jobs than those with High happiness levels (reference). This highlights the significant positive effect of happiness on job satisfaction.
4. **β^how_happy_Medium: -0.6239 < 0.** Individuals reporting Medium happiness levels are less likely to be satisfied with their jobs than those with High happiness levels (reference).
5. **β^how_happy_Very High: 1.0090 > 0.** Individuals reporting Very High happiness levels are more likely to be satisfied with their jobs than those with High happiness levels (reference).
6. **β^line_manager_gives_workrelated_help_likely_Very Low: 0.7100 > 0.** For this category, the order is inverted. The very low category is actually Very High in the list of options. Individuals whose line manager respond Very High to provide work-related help are more likely to be satisfied with their job than those with who respond High (reference).
7. **β^line_manager_gives_workrelated_help_likely_High: -0.5431 < 0.** Individuals whose line manager respond Low to provide work-related help are less likely to be satisfied with their job than those with who respond High (reference).
8. **β^line_manager_gives_workrelated_help_likely_Very High: -1.2833 < 0.** Individuals whose line manager respond Very Low to provide work-related help are less likely to be satisfied with their job than those with who respond High (reference).
9. **β^line_manager_supports_employees_balancing_work_much_Medium: -0.7581 < 0.** Individuals who receive Medium support from line managers to balance work are less likely to be satisfied with their job, highlighting the importance of support for work-life balance.

10. **β̂number_people_living_regularly_member_household_Very Low: -0.3729 < 0.** Living in a household with very few members regularly decreases the likelihood of job satisfaction, suggesting that personal living arrangements may impact job satisfaction.
11. **β̂partner_family_fed_pressure_job_often_Very Low: 1.1960 > 0.** Individuals who rarely feel pressured by their partner or family about their job are more likely to be satisfied with their job than individuals who experience Medium pressure (reference class), indicating the stress from family expectations can negatively impact job satisfaction.
12. **β̂partner_family_fed_pressure_job_often_Low: 0.8606 > 0.** Individuals who respond Low in feeling pressured by their partner or family about their job are more likely to be satisfied with their job than individuals who experience Medium pressure (reference class), indicating the stress from family expectations can negatively impact job satisfaction.
13. **β̂too_tired_work_enjoy_things_like_home_often_Very High: -1.2955 < 0.** Frequently feeling too tired because of work to enjoy things at home significantly reduces job satisfaction, emphasizing the negative impact of work-related exhaustion on life satisfaction.
14. **β̂doing_last__days_unemployed_actively_looking_job_0_Not marked: 1.8101 > 0.** Individuals who did not mark their unemployment status and actively looking for a job in the last days, are more likely to be satisfied with their job than those who didn't mark it (reference class), suggesting a strong negative impact of recent unemployment on job satisfaction.
15. **β̂responsible_supervising_employees_1_Yes: 0.6336 > 0.** Being responsible for supervising employees increases job satisfaction compared to those who do not supervise others, highlighting the positive aspects of leadership roles.
16. **β̂occupation_isco_Drivers, Software Developers, and Miscellaneous: -1.3635 < 0.** Individuals in occupations such as drivers, software developers, and other miscellaneous categories are less likely to be satisfied with their jobs, indicating job-specific factors affecting satisfaction.
17. **β̂what_type_organisation_work_worked_Miscellaneous: -0.5945 < 0.** Working in miscellaneous types of organizations is associated with lower job satisfaction, compared to reference type of organization work, suggesting organizational factors play a role in job satisfaction.

18. **β^industry_nace_rev_Transportation and Miscellaneous: -1.3515 < 0.** Being employed in the transportation industry and miscellaneous sectors is linked to lower job satisfaction, compared to reference sectors, pointing to industry-specific influences.
19. **β^region_Northern and Central Regions: 0.3961 > 0**. Living in northern and central regions is associated with a higher probability of job satisfaction, compared to the reference region, indicating regional differences in job satisfaction.

Odd ratio

Furthermore, we analysed the odds ratios. Firstly, we noted that all variables are significant. The confidence intervals for all variables did not include 1, which meant that each variable has a different effect from the reference class.

OR(how_emotionally_attached_country_Medium vs Very High) = 0.54. An individual who reports Medium attachment to country has odds 0.54 times the odds of an individual who reports Very High (reference).In terms of variation rate: there is a decrease of 46% of the odd (chance to be satisfied with job rather than not be satisfied) when switching from Very High to Medium attachment.

OR(how_happy_Low vs High) = 0.378. An individual who reports Low happiness has odds 0.374 times the odds of an individual who reports High happiness (reference). In terms of variation rate: there is a decrease of 62.2% (chance to be satisfied with job rather than not be satisfied) when switching from High to Low happiness.

OR(how_happy_Medium vs High) = 0.536. An individual who reports Medium happinesshas odds 0.536 times the odds of an individual who reports High happiness (reference). In terms of variation rate: there is a decrease of 46.4% (chance to be satisfied with job rather than not be satisfied) when switching from High to Medium happiness.

OR(how_happy_Very High vs High) = 2.743. An individual who reports Very High happiness has odds 2.743 times the odds of an individual who reports High happiness (reference). In terms of variation rate: there is an increase of 174.3% (chance to be satisfied with job rather than not be satisfied) when switching from High to Very High happiness.

OR(line_manager_gives_workrelated_help_likely_Very High vs High) = 2.034.

An individual whose line manager is very likely to give work-related help has odds equal to 2.034 times the odds of an individual whose manager is High to offer workrelated help (reference). In terms of variation rate: there is an increase of 103.4% in the odds of job satisfaction for the switch from High to Very High to get help from a line manager.

OR(line_manager_gives_workrelated_help_likely_Low vs High) = 0.581.

An individual whose line manager scores Low to give work-related help has odds equal to 0.581 times the odds of an individual whose manager is High to offer workrelated help (reference). In terms of variation rate: there is a decrease of of 41.9% in the odds of job satisfaction for the switch from High to Low to get help from a line manager.

OR(line_manager_gives_workrelated_help_likely_Very_Low vs High) = 0.277.

An individual whose line manager is scores Very Low to give work-related help has odds equal to 0.277 times the odds of an individual whose manager scores High to offer workrelated help (reference). In terms of variation rate: there is a decrease of 72.3% in the odds of job satisfaction for the switch from High to Very Low to get help from a line manager.

OR(line_manager_supports_employees_balancing_work_much_Medium vs High) = 0.469. An individual who receives Medium support from line managers for balancing work has 0.469 times the odds of job satisfaction compared to someone with High support (reference), indicating a 53.1% decrease in the odds of being satisfied with their job if they switch from High to Medium.

OR(number_people_living_regularly_member_household_Very Low vs High) = 0.689. Individuals living in households with very few regular members have 0.689 times the odds of being job satisfied compared to those with High members (reference). This shows a decrease of 31.1% in the odds of job satisfaction for the switch from many family members to low family members.

OR(partner_family_fed_pressure_job_often_Very Low vs Medium) = 3.307. Individuals who report Very Low to feel job-related pressure from family or partners have 3.307 times the odds of being job satisfied compared to those who report to feel it Medium (reference). There is an increase of 230.7% in the odds of job satisfaction when pressure switches from Medium to Very Low.

OR(partner_family_fed_pressure_job_often_Low vs Medium) = 2.365. Individuals who report Low to feel job pressure from family or partners have 2.365 times the odds of being job satisfied compared to individuals who experience it at a Medium frequency (reference). This corresponds to an increase of 136.5% in the odds of being satisfied with their job.

OR(too_tired_work_enjoy_things_like_home_often_Very High vs High) = 0.274. Individuals who report Very High in being too tired from work to enjoy things at home have 0.274 times the odds of job satisfaction compared to those who report High (reference). There's a decrease of 72.6% in the odds of job satisfaction if the individual switches from High to Very High.

OR(doing_last__days_unemployed_actively_looking_job_0_Not marked vs Marked) = 6.111. Individuals who did not mark themselves as unemployed and actively looking for a job in the last days have 6.111 times the odds of being job satisfied compared to those who did mark themselves as such (reference). The odds of job satisfaction increase by 511.1% if individual switches from marked unemployed to not marked.

OR(responsible_supervising_employees_1_Yes vs No) = 1.884. Individuals responsible for supervising employees have 1.884 times the odds of job satisfaction compared to those who aren't supervisors (reference), indicating an 88.4% increase in the odds of job satisfaction when individual switches to have supervisory responsibilities from not having responsibilities.

OR(occupation_isco_Drivers, Software Developers, and Miscellaneous vs Other Occupations) = 0.256. Individuals in occupations such as drivers, software developers, and miscellaneous categories have 0.256 times the odds of being job satisfied compared to those in the reference occupations (reference), showing a decrease of 74.4% in the odds of job satisfaction when individual switches from other occupations to drivers...

OR(what_type_organisation_work_worked_Miscellaneous vs Other Types) = 0.552. Individuals working in miscellaneous types of organizations have 0.552 times the odds of job satisfaction compared to working in other types of organizations (reference). This reflects a 44.8% decrease in the odds of job satisfaction when the individual switxhes from other types of organisation work to miscellaneous type of work.

OR(industry_nace_rev_Transportation and Miscellaneous vs Other Industries) = 0.259. Individuals who are employed in the transportation industry and miscellaneous sectors have 0.259 times the odds of job satisfaction compared to other industries (reference), indicating a decrease of 74.1% in the odds of job satisfaction when the individual switches industries from others to transportation.

OR(region_Northern and Central Regions vs Other Regions) = 1.486. Individuals living in northern and central regions have an odd 1.486 times the odds of job satisfaction compared to other regions (reference), representing an increase of 48.6% in the odds of being satisfied with one's job if the individual switches from other regions to northern and central regions.

Marginal Effect

We continued our analysis by calculating the marginal effects for individual 1. According to our calculations, we got the following results:

- When the emotional attachment to country changes from Very High to Low, the probability to be job satisfied, compared to not be job satisfied, decreases by 10.44%.
- When the emotional attachment to country changes from Very High to Medium, the probability to be job satisfied, compared to not be job satisfied, decreases by 7.35%.
- When the happiness category changes from High to Low, the probability to be job satisfied, compared to not be job satisfied, decreases by 10.26%.
- When the happiness category changes from High to Medium, the probability to be job satisfied, compared to not be job satisfied, decreases by 7.43%.
- When the happiness category changes from High to Very High, the probability to be job satisfied, compared to not be job satisfied, decreases by 19.5%.
- When the line manager likelihood to give workrelated help changes from High to Very High, the probability to be job satisfied, compared to not be job satisfied increases by 12.81%.
- When the line manager likelihood to give workrelated help changes from High to Low, the probability to be job satisfied, compared to not be job satisfied decreases by 9.37%
- When the line manager likelihood to give workrelated help changes from High to Very Low, the probability to be job satisfied, compared to not be job satisfied decreases by 12.1%
- When the line manager supporting employees balance work changes from High to Medium, the probability to be job satisfied, compared to not be job satisfied decreases by 8.61%
- When the number of people living regularly in the household changes from High to Very Low, the probability to be job satisfied, compared to not be job satisfied decreases by 4.83%.

- When the job pressure fed from partner and family changes from Medium to Very Low, the probability to be job satisfied, compared to not be job satisfied increases by 23.9%.
- When the job pressure fed from partner and family changes from Medium to Low, the probability to be job satisfied, compared to not be job satisfied increases by 16.1%.
- When the frequency to be too tired from work to enjoy things like home changes from High to Very High, the probability to be job satisfied, compared to not be job satisfied, decreases by 26.4%.
- When the status changes from marked as unemployed and actively looking for a job to not marked,the probability to be job satisfied, compared to not be job satisfied, increases by 14.4%.
- When the status changes from not responsible for supervising employees to responsible for supervising employees, the probability to be job satisfied, compared to not be job satisfied, increases by 7.5%.
- When the occupation changes from others(reference) to Drivers and Miscellaneous, the probability to be job satisfied, compared to not be job satisfied, decreases by 12.6%.
- When the type of organisation changes from other (reference) to miscellaneous, the probability to be job satisfied, compared to not be job satisfied, decreases by 10.4%.
- When the industry changes from other (reference) to Transportation, the probability to be job satisfied, compared to not be job satisfied, decreases by 12.5%.
- When the region changes from other to Northern and Central Regions, the probability to be job satisfied, compared to not be job satisfied, increases by 5%.

Scorecard

	Category	Level	Score
0	Doing Last Days Unemployed Actively Looking Job	Marked	12.93
1	Doing Last Days Unemployed Actively Looking Job	Not Marked	0.00
2	How Emotionally Attached Country	Low	7.12
3	How Emotionally Attached Country	Medium	4.40
4	How Emotionally Attached Country	Very High	0.00
5	How Happy	Low	14.15
6	How Happy	Medium	11.66
7	How Happy	High	7.21
8	How Happy	Very High	0.00
9	Industry Nace Rev	Others	9.65
10	Industry Nace Rev	Transportation And Miscellaneous	0.00
11	Line Manager Gives Workrelated Help Likely	Low	8.95
12	Line Manager Gives Workrelated Help Likely	Very Low	5.07
13	Line Manager Gives Workrelated Help Likely	Very High	0.00
14	Line Manager Supports Employees Balancing Work...	Medium	5.41
15	Line Manager Supports Employees Balancing Work...	High	0.00
16	Number People Living Regularly Member Household	Very Low	2.66
17	Number People Living Regularly Member Household	High	0.00
18	Occupation Isco	Others	9.74
19	Occupation Isco	Drivers, Software Developers, And Miscellaneous	0.00
20	Partner Family Fed Pressure Job Often	Medium	8.54
21	Partner Family Fed Pressure Job Often	Low	2.40
22	Partner Family Fed Pressure Job Often	Very Low	0.00
23	Region	Northern And Central Regions	2.83
24	Region	Others	0.00
25	Responsible Supervising Employees	Yes	4.52
26	Responsible Supervising Employees	No	0.00
27	Too Tired Work Enjoy Things Like Home Often	Very High	9.25
28	Too Tired Work Enjoy Things Like Home Often	High	0.00
29	What Type Organisation Work Worked	Others	4.25
30	What Type Organisation Work Worked	Miscellaneous	0.00

Conclusion

The project aimed at performing an analysis between economic fluctuations, job satisfaction, and various socio-demographic factors in Spain, recurring to data from the European Social Survey. During the study, a methodology was followed encompassing exploratory data analysis, model generation, and extensive model selection.

To pursue our goal, three models were created. The model 1 was a binned model with categorical variables, emerging as the optimal choice. It effectively captured patterns in the data, outperforming models with interaction variables (model 2) and original variables with transformations (model 3). The interpretation of Model 1's coefficients, odds ratios, and marginal effect shows the influencing factors of job satisfaction.

Key findings show that emotional attachment to the country, happiness levels, support from line managers, family pressure, fatigue from work, and occupational and organizational factors significantly impact job satisfaction. The model's predictive accuracy was measured by AUC, R2 Efron, and classification accuracy, reinforced its effectiveness.

In practical terms, the insights from this project can inform targeted interventions and policies to enhance job satisfaction in Spain. Whether through fostering a positive work environment, addressing family-related stressors, or tailoring support for specific occupations, the findings offer actionable strategies for improving overall workplace well-being.