Journal of Big Data

RESEARCH Open Access

Supervised machine learning predictive analytics for alumni income

Daniela A. Gomez-Cravioto, Ramon E. Diaz-Ramos*, Neil Hernandez-Gress, Jose Luis Preciado and Hector G. Ceballos

*Correspondence:
a01133921@itesm.mx
School of Engineering
and Sciences, Tecnologico
de Monterrey,
64849 Monterrey, NL, Mexico

Abstract

Background: This paper explores machine learning algorithms and approaches for predicting alum income to obtain insights on the strongest predictors and a 'high' earners' class.

Methods: It examines the alum sample data obtained from a survey from a multicampus Mexican private university. Survey results include 17,898 and 12,275 observations before and after cleaning and pre-processing, respectively. The dataset comprises income values and a large set of independent demographical attributes of former students. We conduct an in-depth analysis to determine whether the accuracy of traditional algorithms can be improved with a data science approach. Furthermore, we present insights on patterns obtained using explainable artificial intelligence techniques.

Results: Results show that the machine learning models outperformed the parametric models of linear and logistic regression, in predicting alum's current income with statistically significant results ($p < 0.05$) in three different tasks. Moreover, the later methods were found to be the most accurate in predicting the alum's first income after graduation.

Conclusion: We identified that age, gender, working hours per week, first income and variables related to the alum's job position and firm contributed to explaining their current income. Findings indicated a gender wage gap, suggesting that further work is needed to enable equality.

Keywords: Machine learning, Income prediction, Alumni survey analysis, Knowledge Discovery, Explainable Artificial Intelligence

Introduction

Higher education institutions seek to boost their alumni outcomes after graduation. To validate whether this goal is being accomplished, there is value in collecting data from their alumni and identifying patterns between those who achieved their expected outcomes and those who did not. The results from this analysis can help guide stakeholders' decisions to support future alumni. In this study, we exploit the data from Tecnologico de Monterrey's alumni survey to obtain two main insights: assessing students' economic outcomes and validating gaps with relation to diverse backgrounds (e.g.

gender, education and socioeconomic diversity). Understanding the factors that may favour some alumni will help give them equal opportunities to achieve their economic objectives.

Many institutions survey their graduates to collect information on their post-graduation outcomes, such as their income and socioeconomic status [1–3]. These studies are beneficial to evaluate an institution's effectiveness and support institutional planning and future students' achievements. Unfortunately, the actions taken to analyze the results rarely include data mining to obtain insights regarding features that can have a higher relationship with the outcome. This is especially true for actionable features that can be boosted with activities performed during students' lives on campus. In this work, a data-based model is built for understanding the main factors that can influence alumni income prediction. The study uses data science, advanced analytics and machine learning techniques. While career success can be evaluated as intrinsic or extrinsic [4], this study will focus solely on extrinsic success; based explicitly on the objective rating of salary.

The data on which this work is focused comes from a survey carried out in 2018 by Tecnologico de Monterrey in the university's approach to measure their graduates' social and economic impact. The survey was sent through email to the total alumni population who graduated from 1953 and 2017, and advertisement for this survey was promoted on social media. The overall response rate was 7% of the total population, accounting for 17,896 former students. The obtained data set provides an excellent opportunity to supplement the university with knowledge about previously hidden trends and patterns regarding the factors that affect alumni salary attainment. This study's primary purpose is to identify if factors such as age, gender, major, graduate studies, the overall grade achieved, and parent's education and occupation can influence the alumnus's first income after graduation and their current monthly salary. Furthermore, we aim to identify the variables that also impact the first income after graduation, which have resulted in a significant predictor for the former.

The contribution of this study is threefold. Firstly, it contributes to the modern field of machine learning research applied to econometric studies by exploring income distribution and comparing traditional econometric techniques, such as Quantile Regression, Linear Regression and Logistic Regression with machine learning non-parametric tree-based algorithms, Random Forest and Gradient Boosting to find the best method for approaching the problem of income prediction. Secondly, the study adds to the existing literature in Educational Data Analytics with a data-driven approach and machine learning algorithms applied to an alumni impact survey dataset. Finally, the study adds to the application of Knowledge Discovery in Data and Explainable Artificial Intelligence by identifying rule-based patterns in the dataset, identifying feature importance with Shapley Additive exPlanations (SHAP) values, and performing a sensitivity analysis on the variables detected as having the most important relationships with income.

This paper is organized as follows. The remainder of section "Introduction" is composed of prior work performed for income prediction and the description of the dataset used in the study. Section "Methodology" presents the methodology for the data preparation, exploratory analysis, model building, tuning, and evaluation. Section "Results" presents the results of the experiments and the explanation of the predictions made by the model. Section "Discussions" present the discussion, contributions and limitations. Finally, section Conclusions and future work presents the study's conclusions and considerations for future paths.

Related work

Over the past several decades, many studies have estimated how the final grades, college major, demographics, and occupation characteristics affect individuals' income. However, very few studies have combined all these characteristics in a single model. This research builds on previous works that examined college students' future income to determine the most important features and use machine learning as a tool to assess these features. A table showing the most recent studies on individual income prediction with a multivariate model can be seen in Table 1.

Alina Lazar [5] proposed the use of Support Vector Machines (SVM) to predict income. She used the Current Population Survey (CSP) from the U.S. Census Bureau as a database for her study. This dataset contained social, demographic and economic characteristics of U.S. citizens 16 years and older. The author used Principal Component Analysis (PCA) to reduce the number of features in the dataset and then fed this to a Support Vector Machine classifier. With this, Lazar achieved an accuracy score as high as 84%.

The study from Hartog and Webbink [6] analysed both expectation and realisation of incomes from former students who graduated from high schools or universities in the

Table 1 Recent studies on income prediction summary recollection

Source	Task	Methods	Results
Lazar [5]	Classification	SVM	Acc = 0.84
Hartog and Webbink [6]	Regression	OLS	R2 = 0.14

Lee and Lee [7]	Quantile regression	5th
		25th
		50th
		75th
		95th
			Pseudo-R2 = 0.29
			Pseudo-R2 = 0.33
			Pseudo-R2 = 0.34
			Pseudo-R2 = 0.34
			Pseudo-R2 = 0.32

Oehlrein [8]	Regression	OLS	R2 = 0.37
Stran and Truong [9]	Regression	Lasso OLS	USD $6,394.64 (RMSE)

Figueiredo and Fontainha [10]	Quantile regression	10th
		50th
		90th
			Pseudo-R2 = 0.27
			Pseudo-R2 = 0.45
			Pseudo-R2 = 0.50

Sharath et al. [11]	Classification	NB
		C4.5
		Boosted C4.5
			Acc = 0.48
			Acc = 0.51
			Acc = 0.53

Khongchai and Songmuang [12]	Multi-class classification	DT
		SVM
		MLP
		KNN
		NB
			Acc = 0.73
			Acc =0.43
			Acc =0.38
			Acc = 0.84
			Acc =0.43

Chen et al. [13]	Multi-class classification	SVM
		DT
		LR
		RF
		GBM
		NN
		LSTM
		DNN
			Acc = 0.74
			Acc = 0.74
			Acc = 0.72
			Acc = 0.71
			Acc = 0.70
			Acc = 0.68
			Acc = 0.65
			Acc =0.65

Netherlands. The variables analysed included background variables (gender, age, parent's education, parent's income), higher education variables (year of education, student's status), and secondary education variables (school marks), potential work experience (time since graduation). One of the experiments conducted in this study included a prediction of realised earnings. This model was performed with Ordinary Least Squares (OLS) regression and achieved a 16% R2.

The study from Lee and Lee [7] investigated the wage determinants in the Korean labour market. The researchers used quantile regression methods. They indicated that the advantage of quantile regressions is that it allows examining a more comprehensive picture for different quantile wage groups. The results obtained from their study showed that age is the most important factor for wage determination. The authors also found that female workers are significantly underpaid compared to their male counterparts.

Oehrlein [8] attempted to determine the aspects of college that impacted students' future income. He focused on deciding whether or not their GPA was an influencer. In this study, OLS regression was used, and we obtained a prediction R-squared score of 0.374. The author's findings include that grades, natural ability, and major significantly affect income. He found that the highest paying major was engineering and that the attribute female was negatively correlated with income.

The research study from Stran and Truong [9] evaluated different demographic features to predict earnings by comparing the results of students graduating from several colleges. The most important features identified in this study were the percentage of students who received a Pell grant, the number of female students, the rate of first-generation students, and the percentage of students who had sent a FAFSA application to multiple schools before entering. The best performance, considering the MSE, was the one from Lasso Regression and Random Forest.

The research performed by Figueiredo and Fontainha [10] studied the distinct wages for men and women in Portugal with an OLS and a quantile regression approach. The results from this study showed that quantile regression obtained better results than OLS. The findings indicated that the levels of education have a higher impact on wage determination. Also, the variables that contributed the most in the model were related to the firm, while those related to family only contributed to explaining men's wages. Finally, the study indicated a significant difference between men's and women's wages, indicating that further studies are required to explain the gender wage gap.

Sharath et al. [11] performed a machine learning study with the US Census Bureau dataset. The focus of the study was to obtain insights into the financial status of the people in the US. The results obtained showed inequality in society due to a gender wage gap. They showcased one of the root causes of these inequalities by determining the relationship between income and education level. Furthermore, we obtained a classification model to predict economic class categories with an accuracy of 53%.

Khongchai and Songmuang [12] presented their work using classification to predict future students' income. Their initial dataset contained 108 attributes obtained from graduate student history data collected for 10 years from a university in Thailand. The features included gender, faculty-student ratio, programme, workplace type, work experience, certifications, total Grade Point Average (GPA), and salary. The best model obtained by the authors was the KNN with an 84% accuracy.

During this research, the most recent work was the one from Chen, Sun, and Thakuriah in 2019 [13]. The authors used many metadata in the web and relational attributes such as job descriptions, locations, job content and job-related features to predict individuals' salaries. They compared Support Vector Machines, Decision Trees, Logistic regression, Random Forest, Gradient Boosting, Graph Convolutional Networks and Deep Learning to classify six predefined categories (0–20,000; 20,000–30,000; 30,000–40,000; 40,000–50,000; and >50,000). The best overall accuracy obtained was SVM with a 0.74 accuracy. The metadata features had a significant contribution to reaching this accuracy.

Dataset

This study analyses data from an alumni impact study survey conducted by Tecnologico de Monterrey university. The survey invitation was electronically sent to all former students since the inception of the university in 1943 (269,482 individuals). From the total population of alumni who graduated between 1953 and 2017, 7% responded to the survey; this accounts for 17,896 graduates across different generations. Tecnologico de Monterrey provided the original dataset collected from the survey for this study. The dataset contains no personally identifiable information, and the dataset contains all the salary figures normalised and reported in Mexican Pesos.

The records include 72 columns with demographic information from the alumni such as major, gender, graduation date, campus, age, occupation, level of education attained, parents' education, parents' occupation, as well as information related to their accomplishments such as businesses created, type of business, salary and score reported based on their satisfaction in their professional lives, as well as other variables.

Regarding the unintended bias inherent in this dataset, the first bias we identified was the one towards younger adults, specifically for those between 27 and 60 years old. Since the survey was sent by email, it accounted for fewer elder respondents (older than 60 years old). Hence, it is essential to note that our results will not account for alumni older than 60.

Methodology

The methodology followed in this study is an adaptation of the Cross-Industry Standard Process for Data Mining (CRISP-DM). The CRISP-DM consists of a general model of a data mining project. This was developed in 1996 [14, 15] and has been widely used since then. The steps followed in this project to transform raw data into insights are shown in Fig. 1. This diagram shows specific actions performed during the application of the CRISP-DM process.

Methods

Quantile Regression (QR) can be used when asymmetries and heavy tails exist in data distributions. The advantage of QR over linear regression is that this method is more robust to outliers and more flexible to the linear assumptions. The main difference between these two is that while least-squares regression is focused on minimising the sums of squared residuals to estimate models for conditional mean functions, QR models the conditional quantile of the response variable for some quantity of

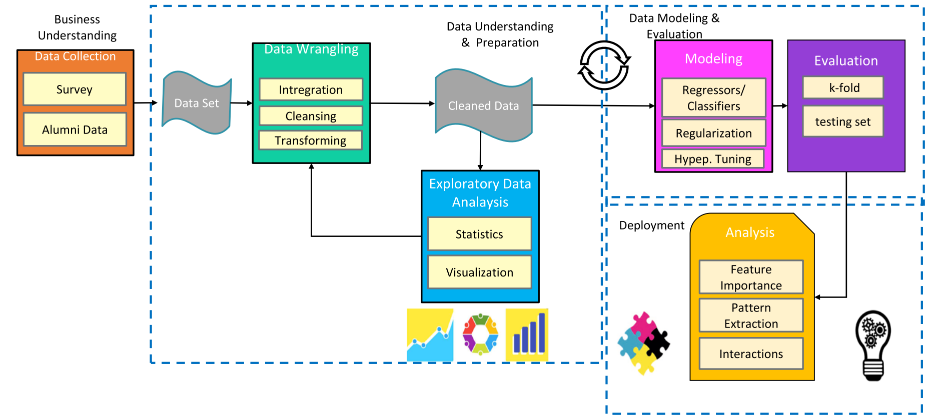

Fig. 1 The Knowledge Discovery Process flow diagram employed in this work and linked to the CRISP-DM methodology

$$\tau \in (0, 1),$$

where $\tau = 0.5$ is the median [16]. For example, when trying to predict income in countries where the income is highly skewed, we can predict the median or the quantile instead of the mean. For this reason, the QR method is highly used in econometrics studies for wage determinants, discrimination effects and income inequality trends.

Ensemble Methods successful approaches to counteract the decision tree issues of stability [17] and accuracy [18], are the ensemble of decision trees. The ensemble approach integrates multiple predictors and is built by two specific methods: bagging and boosting [19]. One of the best performing applications of the bagging method is Random Forest (RF). A practical algorithm based on the boosting notion is the widely used ensemble method Gradient Boosting (GB).

Random Forest algorithm consists of building B random samples, and for each of these samples, building a decision tree model f_b [20]. The final prediction is obtained by taking into account the vote of each of the models for a classification task 1 and the average prediction for a regression task 2.

$$\hat{C}(x) = majority\,vote\{\hat{C}_b\} \tag{1}$$

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f_{b(x)} \tag{2}$$

The advantages of RF are mainly inherited from the decision trees, previously explained. For instance, they can be used for classification and regression tasks; their nature enables them to handle categorical predictors; they are non-parametric models, so they do not need a formal distribution assumption. Additionally, they can manage non-linear relationships between the covariates and target variables and perform feature selection automatically. However, unlike decision trees, random forests are harder to interpret, as the model is built with multiple decision trees, making it hard to visualise in a plot.

Another limitation of this model is that it can become highly computationally complex when having many trees.

Gradient Boosting combines multiple simple decision trees. The trees are joined sequentially, each tree trying to amend the errors of the previous one, $f_i^{(j)}$ (3). Frequently, this method has a better performance than Random Forest while having similar properties; however, careful tuning is required to avoid over-fitting the data [21].

$$f_i = f_i + \alpha f_i^{(j)} \tag{3}$$

Data preparation

The data present in the survey results contain several missing values and an excessive number of attributes. To give the data the proper format for data mining, we performed a series of steps to clean the data.

1. *Data Integration* First, the original dataset was compiled with a dataset with information from the university's planning department, which contained students' data upon graduation. The data included final GPA, number of semesters in which the student was involved in co-curricular activities (sports, leadership, and cultural activity), their English score, and whether they had previous work experience before graduating (internships). In this step, we noted that most campuses track recent students' participation in co-curricular activities and store their scores in a database; however, not all campuses held this information for alumni from older cohorts; this was among the fields with the largest percentage of missing values.
2. *Correcting Inconsistencies* Subject matter expertise was needed to correct errors/inconsistencies since they were present in the survey. We first cleaned the data by translating all the questions to variable names and translating all the data to English. Many different words in the responses referred to the same term, so we grouped them in a single word. We corrected typos and misspellings. Finally, we removed punctuation such as commas, apostrophes, quotes, question marks and others.
3. *Handling Missing Values* The next step performed was the handling of missing values. We eliminated all the records which had no information regarding the target variable. Then, we eliminated follow-up questions with more than 80% of missing values, as they were not of central importance to our analysis. Variables for extra-curricular activities and work experience previous to graduation had more than 80% of missing values. However, as these variables were of interest for our analysis, we split the dataset into two. The first split was all graduates' information; we later used this to predict their current income. The second one was a subset of the original dataset; we preserved all records with information regarding their school activities (co-curricular activities, internships, etc.); we used this later to predict the First Income after graduation. The age of the respondents from the subset is exclusively between 21 and 28 years old at the time of the survey. Hence, this analysis was exclusive to recent graduates (alumni who graduated between 2012 and 2017). After this step, we had less than 40% missing values in both datasets and no values missing for the target variables. We then performed a missing values imputation. We completed the imputation by using a Nearest Neighbours (NN) imputation, considering three dis-

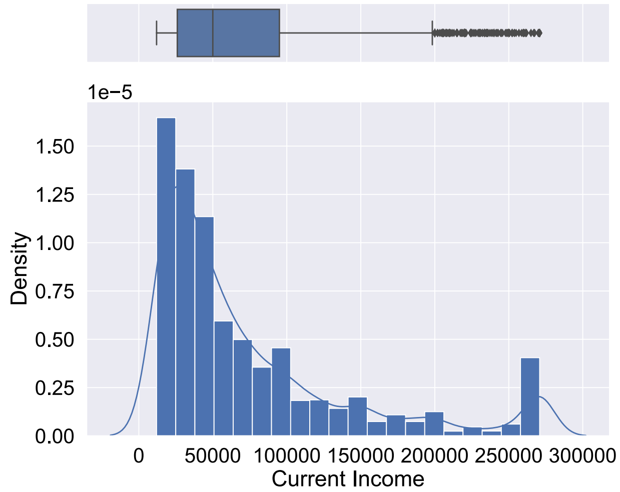

Fig. 2 Density bar plot and boxplot of the winsorized result of the Current Income variable

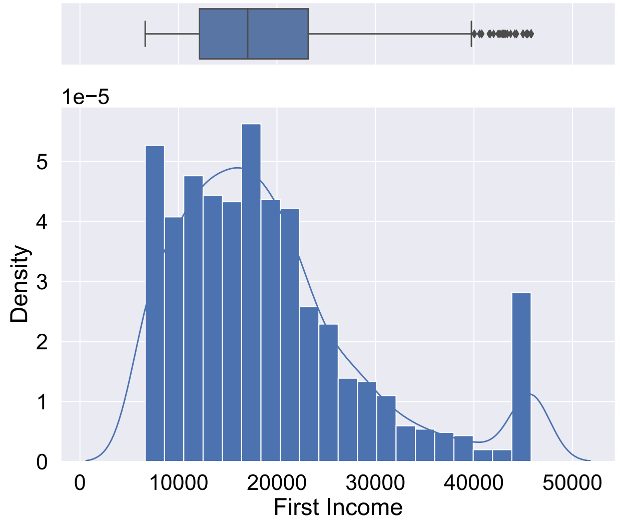

Fig. 3 Density bar plot and boxplot of the winsorized result of the First Income variable

tinct neighbours. We selected a K-NN model for this process as it has proven to be a useful technique for predicting missing values; it has surpassed the efficacy of average or median imputation in previous research [22]. The missing values are imputed using the mean value from the three distinct neighbors' weighted average closer in space and the Euclidean distance metric. We identified extreme values in both target variables, current income and first income after graduation. In this regard, a winsorization method was used to mitigate the effect of the extreme values. The difference between just trimming the data and winsorizing it is that the latter will retain the observations but changes the numeric outliers to fall on the edge of the distribution [23]. We bounded the data to the 0.05 and 0.95 percentiles with the winsorisation. The resulting distribution of the target variables, after the winsorisation process (Current Income and First Income after graduation), are observed in Figs. 2 and 3.

We can see that there are still some outliers in the distribution. However, as these observations are ascertained as genuine, they are not removed. The data should be transformed for its use in data analysis, specifically in parametric models, as these require symmetric distribution. This change is performed in the transformation step in this section. From the figures above, we can observe that both distributions are highly skewed to the right. To use these variables in the linear regression, we must perform an additional transformation to the data. In this step, the transformation that we approached was a box-cox transformation that could approximate normality assumptions. Even though the resulting transformations exhibits symmetry, they do not resemble a normal distribution. Both of the histograms show "heavy tails" a common topic in income distributions [24–27]. Since having fat tails makes it interesting to understand the distribution, we decided to start modelling with a QR model (experiment A) instead of linear regression. The QR and the non-parametric machine learning models explored in this study make no assumptions about the distribution of the residuals; hence, they can be used when asymmetries and heavy tails exist in data distributions. The transformations that yielded the most symmetric distributions were then used for the linear regression model.

4. *Data Binning* Since the unequal representation of the different groups could lead to unfair outcomes towards individuals or demographics, in this step, we seek to drop this difference by binning categories and reduce this imbalance as much as possible. We used data binning to deal with unrepresented groups during the pre-processing of the data. For this, we used an unsupervised discretization method, an equal-frequency binning, for exploring the data in the variables and grouped the predictor variables that contained more than five categorical labels. This was performed to guarantee that every bin had roughly the same amount of data. The labeling of the bins was performed based on business acumen. The process was a labor-intensive activity that could potentially be reduced with automation. An example of this was the variable "Campus." Initially, the variable contained 33 categories, we reduced these to only six based on the economic regions in which the "Campus" are located across Mexico and an additional one for the "Virtual Campus," which represent those students that did not have a physical Campus but took all their courses online. The categories for the variables "Current Location" and "Pre-Study Location" were also binned based on these economic regions and an additional label for those living overseas.
5. *Dealing with Multicollinearity* To determine whether the independent variables were highly correlated, we used the Variance Inflation Factor (VIF) and chose the score five as a threshold. VIF measures the correlation inflation between the independent variables. When the score was above the threshold, we dropped the variable from the dataset. The final results indicated that the remaining variables do not have a VIF above 5, implying no multicollinearity issues in the dataset.
6. *Categorical Encoding* Concerning variable encoding, for ordinal categorical variables, the assignment was done with incremental ordering, starting with the lowest category (i.e. 0 years was assigned a 0, 1 to 3 years was set to one, and more than three years was given 2). We transformed the categorical variables without a natural order-

ing into dummy variables with one- hot-encoding. To deal with the "dummy variable trap" [28], we dropped one of the dummy variables from each categorical feature.

7. *Data Standardization* A standardisation of the data was performed in numerical variables using the standard normal or z-score normalisation method. This process is necessary so that the machine learning models treat all variables equally, and a variable is not considered more important because it has a higher range of values [29].

Exploratory data analysis

After the data wrangling step, the "Current Income" dataset contains 12,275 observations and 65 continuous and categorical variables. Six of these variables are numeric, and 59 are categorical; however, they are now numeric values. On the other hand, the "First Income" dataset contains 2264 observations and 39 variables; 2 are numeric and 37 categorical. The first step for exploring these variables was to analyse correlation to identify those variables having a higher linear relation with the target variables. A heatmap showing the resulting Spearman coefficients for the continuous and ordinal variables is presented in Fig. 4. The results showed a moderate relation ($0.3 \leq |r| \leq 0.5$) between the target variable and "Age," "First Income," "People in Charge," and "Years Worked Foreign." Then, we obtained statistics and visualisations to measure the marginal effect of variables of interest as per previous studies and their relation with the target variable. The money currency for the current and first income variables is in Mexican Pesos (MXN).

Salary Based on Gender The aggregation table in Table 2 depicts a comparison between "Gender" and "First Income" and "Gender" and "Current Income." Overall, there is a gap between the results obtained for each gender; however, the gap is more prominent in the latter target variable. The gap for "First Income" and "Current Income"

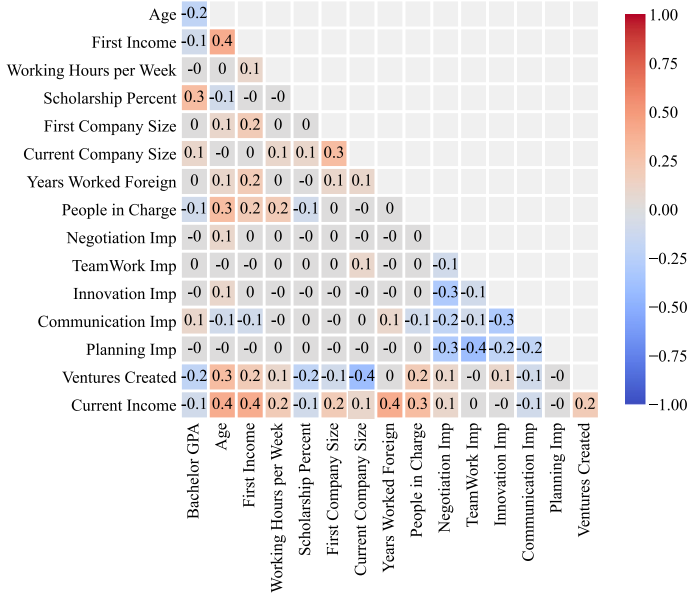

Fig. 4 Current Income Spearman correlation coefficient matrix of the continuous variables of the dataset

Table 2 First Income and Current Income median, mean, and standard deviation statistics by Gender

Category	Mean	Std	Median
First Income			
F	$17,335.73	$8,699.79	$15,615.00
M	$20,940.73	$10,494.72	$18,766.00
Current Income			
F	$55,933.13	$54,427.00	$37,155.00
M	$89,726.13	$73,106.86	$64,000.00

is $3151 and $26,845 respectively. When looking closer into the results with the plot in Fig. 5 and after performing a Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction hypothesis testing, we determined that the medians are significantly different, with a significance level of 0.01% in both analyses.

Salary Based on School The salary variable was examined concerning the school groups. Table 3 shows that the alumni who graduated from "Engineering" have a higher median than the other categories in both cases. The plot in Fig. 6 exhibits a significant difference between the School Variable medians. The difference between "Engineering" and "Business" is not that significant in the "First Income" result; however it becomes more critical in the latter score, where all of the comparisons obtained a significance level of 0.01%.

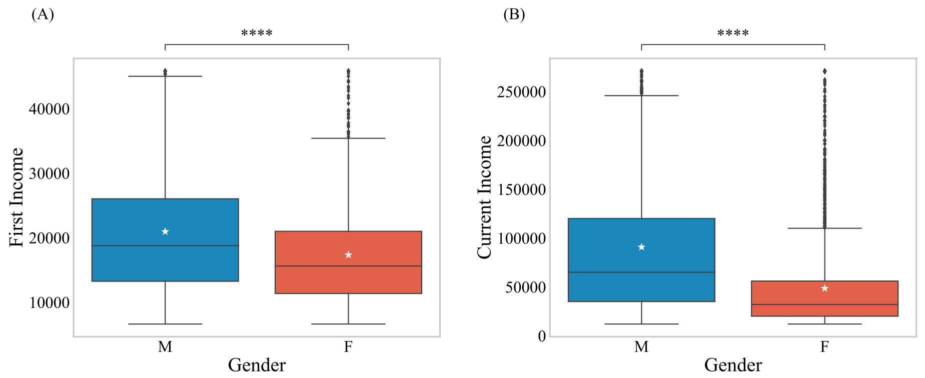

Fig. 5 First Income and Current Income boxplot based on Gender (male M, and female F) distribution count

Table 3 First Income and Current Income median, mean, and standard deviation statistics by School variable

Category	Mean	Std	Median
First Income			
Business	$ 18,398.68	$ 9,164.97	$ 16,558.00
Engineering	$ 21,555.90	$ 10,068.39	$ 19,870.00
Other	$ 15,922.15	$ 8,990.18	$ 13,533.00
Current Income			
Business	$ 78,838.27	$ 68,905.98	$ 53,707.50
Engineering	$ 81,871.54	$ 69,803.65	$ 56,000.00
Other	$ 55,154.67	$ 57,171.84	$ 34,657.00

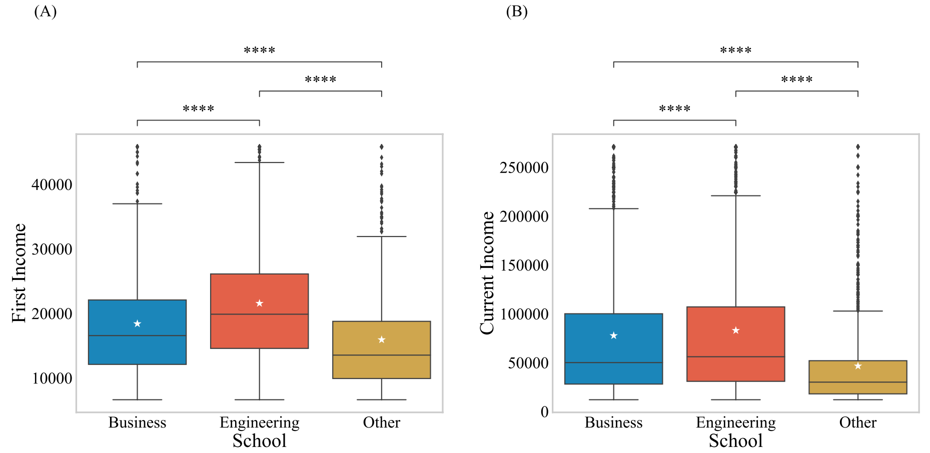

Fig. 6 First Income and Current Income boxplot based on School distribution count

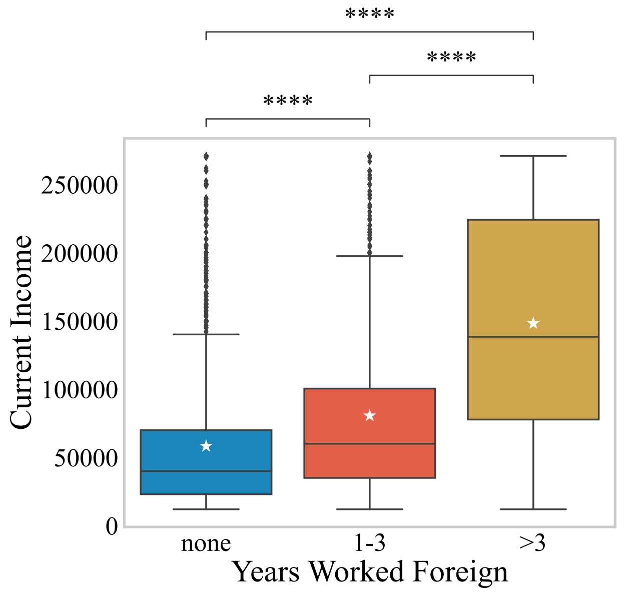

Fig. 7 First Income and Current Income boxplot based on Years Worked Foreign variable distribution count

Salary based current employment characteristics

The box-plots in Figs. 7 and 8 presents a comparison between the most critical variables identified in the correlation analysis, which describe ‘current employment’ characteristics. These variables were not evaluated in the ‘First Income’ analysis as these are characteristics of the alumnus’s current status. In this analysis, we can see a significant difference between the number of years that the alumni have lived in a foreign country (outside of Mexico), showing higher values for those that have lived (and presumably worked) outside the longest (Table 4). The study did not present information about where these alumni have lived outside of Mexico. However, based on previous academic descriptive analysis, it was determined that 70% of the former students have migrated to North America. Finally, whether the former student has obtained a graduate degree or

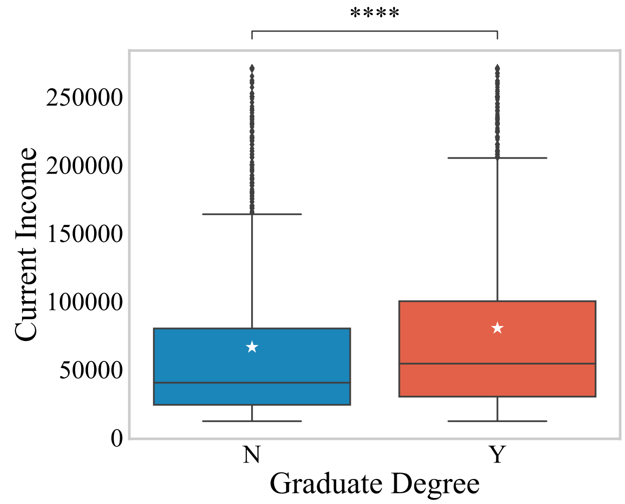

Fig. 8 First Income and Current Income boxplot based on Graduate Degree variable distribution count

Table 4 First Income and Current Income median, mean, and standard deviation statistics by Years Worked Foreign variable

Current Income			
Category	**Mean**	**Std**	**Median**
None	$ 62,883.47	$ 59,282.52	$ 40,100.00
1–3	$ 80,345.03	$ 65,678.08	$ 60,000.00
> 3	$ 137,887.29	$ 81,580.06	$ 117,362.00

Table 5 First Income and Current Income median, mean, and standard deviation statistics by Graduate Degree variable

Current Income			
Category	**Mean**	**Std**	**Median**
No	$ 66,801.01	$ 63,662.08	$ 42,000.00
Yes	$ 83,083.64	$ 70,611.37	$ 58,036.00

not has a significant difference (Table 5), showing a positive outcome for those that have achieved higher educational attainment (a Master's or Ph.D. degree).

Modeling

In the previous section, the marginal analysis provided us a general picture of the interrelation between selected variables in the survey with income. The results are limited, as they only provide a descriptive statistic of bivariate association; they do not reflect relationships between covariates and their impacts. Thus, a multivariate analysis is explored to give us more precise assertions and greater predictive power. To this end, this section presents multiple experiment configurations using both statistical and machine learning techniques performed to analyze the alumni monthly income.

Since the most popular techniques used in econometric studies for income prediction are QR and linear regression, this research starts by evaluating these techniques and then comparing them with modern non-parametric machine learning algorithms, RF QR and GB QR. Subsequently, to explain the most important factors related to alumni income, the study proposes exploring the data through a classification setting. This is done by discretizing the dependent variables with multiple quartile categorizations and using a median split.

The modeling experimentation for this analysis was performed in four different configurations. The experiments were labeled with letters that go from A to D. *Experiment A* involves a Quantile Regression; *Experiment B* involves a Traditional Ordinary Least Squares (OLS) Regression, *Experiment C* includes a Multi-Class Classification, and finally, *Experiment D* consists of a Binary Classification.

The learning algorithms employed in each one of the experiments are: QR, Multiple-Linear Regression, Logistic Regression, RF and GB. For the binary classification, we also integrated other machine learning models to compare their performance with other state-of-the-art methods. This evaluation contemplated a total of eight classifiers: Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbours (KNN), a simple C4.5 Decision Tree (DT), Support Vector Machines (SVM), Naive Bayes Classifier (NB) Random Forest Classifier (RFC) and Gradient Boosting Classifier (GBC).

Regularization and feature selection

To avoid overfitting the models, for the linear, QR, and LR models, we used a lasso regularization. For the selection of the lambdas used in the penalty term, we performed a 10-fold-cross validation in each of the models and used the loss function as the absolute error metric. While regularization was obtained with this process, we found that the number of variables selected was still too many to design concrete policies based on them. Therefore, to further reduce the identified variables, we conducted a Sequential Forward Floating Selection (SFFS) algorithm and evaluated the performance of the model with the top twenty most important variables. The SFFS is a floating variant of the traditional stepwise variable selection method [30]. It involves searching for the best subset of variables by adding and removing features at each step and evaluating the loss function.

On the other hand, for the tree-based methods, we performed the Recursive feature elimination (RFE) [31] method with a 10-fold-cross-validation in the training set. This technique implements a backward selection; it starts with a model that contemplates all predictors and continuously evaluates the model's score when removing each one of them. Those features with less importance are then removed from the final model. This method is frequently used with tree-based ensemble models since they can leverage the RF and GB internal methods for measuring feature importance [32].

With the SFFS (Fig. 9) and RFE (Fig. 10 methods, we were able to select the most important variables for the model; twenty variables were selected for the 'Current Income' model and 16 for the 'First Income' model. The subset of features was selected based on the optimizing the loss function.

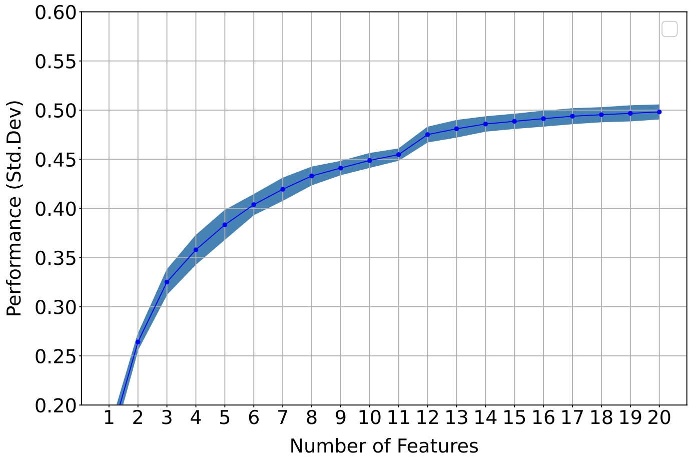

Fig. 9 Sequential forward floating selection linear regression with top 20 most important variables

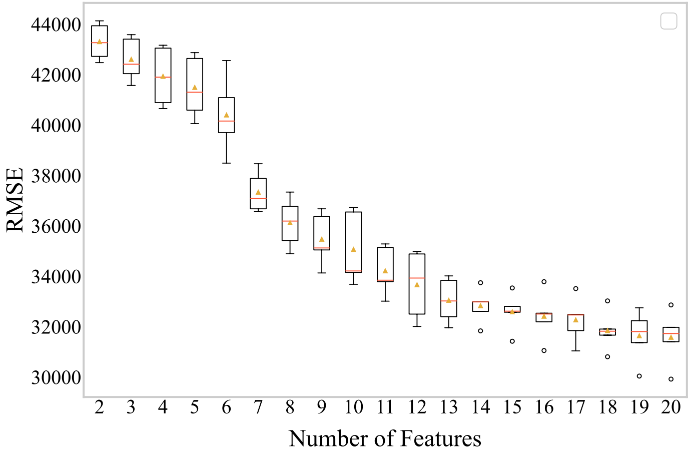

Fig. 10 Recursive feature elimination of gradient boosting model for the 20 most important variables

Cross-validation and evaluation metrics

For model development evaluation purposes, the datasets were split into a training set and a testing set, with 80% and 20% of the data. The latter set was left out for the last evaluation process, and the training set was split into multiple partitions with the use of a stratified 10-fold cross-validation. This split was done to estimate the regressors and classifiers' performance and to carry out the hyper-parameter tuning. The complete dataset was stratified uniformly so that there were all different types of attributes' values in both the Training set and the Test set.

The metrics evaluated with the cross-validation method were accuracy and Area Under the Curve (AUC) for the classification task and Root Mean Squared Error (RMSE) and adjusted-R2 for regression.

When working with a sample with high dimensionality, it is preferable to use the adjusted-R-squared-statistic as it penalizes the use of predictors that are not helping explain the variation of the dependent variable [33, 34]. Equation 4 describes the adjusted-R-squared statistic, where n represents the sample size and k the number of features for the given observations in the analysis.

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1} \tag{4}$$

For the use of R^2 in quantile regression models, we use the pseudo-R-squared (Eq. 5) defined by Koenker and Manchado in 1999 [35]. This metric allows the measurement of variability for a particular quantile defined by τ. $\hat{V}(\tau)$ represents the pseudo R-squared for an unrestricted quantile regression model, while $\tilde{V}(\tau)$ is an intercept-only model. The pseudo-R-square metric value, such as in the traditional R-square, ranges between [0,1]. Still, it is a local measure of how well a particular quantile fits the model, not a global measure of goodness of fit for the total distribution [36].

$$R^1(\tau) = 1 - \frac{\hat{V}(\tau)}{\tilde{V}(\tau)} \tag{5}$$

To evaluate the loss function for linear regression, we measured the RMSE (Eq. 6). To measure this in quantile regression, the quantile-loss error is used [37, 38]. This is also called the pinball loss and is similar to the Mean absolute Error (MAE) loss; however, it is not based on the mean but in the conditional quantile. The formula is shown in Eq. 7.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{6}$$

$$L(y,t) := \begin{cases} (1-\tau)(t-y) & \text{if } y < t \\ \tau(y-t) & \text{if } y \geq t \end{cases} \tag{7}$$

To measure classification models' performance, the confusion matrix and the following metrics are computed: overall accuracy (Eq. 8), and the AUC. The latter measures the two-dimensional area that is underneath the receiver operating characteristic curve (ROC). The ROC curve is the graph that counts the number of correct positive classification gains in each of its thresholds; the curve plots the True Positive Rate (TPR) and the False Positive Rate (FPR) as defined below in Eqs. 9 and 10.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$TPR = \frac{TP}{TP + FN} \tag{9}$$

$$FPR = \frac{FP}{FP + TN} \tag{10}$$

Model interpretation methods

By utilizing complex methods, such as ensemble methods or deep learning, more complex interactions can be found by the algorithms; and potentially a higher accuracy can be obtained. The problem with these, however, is that interpreting the results of complex machine learning methods is not straightforward as it is in simple models, such as linear regression. For interpreting these, there exists a collection of post hoc interpretability methods. These methods seek to convert 'black-box models' to 'glass-box models' and are referred to as Explainable Artificial Intelligence or XAI techniques [39].

In this study, to obtain the most important variables from the best performing model for all of the different experiments we did the following: if the best performing model was a tree-based model, we used Shapley Additive exPlanations (SHAP), and if it was a LR model, we obtained the most important features based on each variable's weighted coefficients.

SHAP helps explain how complex machine learning models make predictions and provides global interpretability using game theory and providing each feature with a SHAP value. SHAP values were introduced by Shapley [40]. They provide a way to distribute contributors' total gain (attribute's marginal contribution), assuming that all features contribute. The greater the Shapley value, the more positive effect it has on the objective function. SHAP values give feature attribution to each future with the classical Shapley values from game theory.

For the last approach presented, we performed two XAI methods to explain the alumni income results. For this, we employed the binary model since categorization can ease the presentation of variable effects. The first strategy consisted of visualizing the interactions between two variables and their relation with the target variable with the use of SHAP dependence plots. SHAP dependence plots are a popular visualization technique to summarize model predictions. This method is similar to the Partial Dependence Plot (PDP) introduced by Friedman [41]. They show how a feature relates to the model's target value. In the SHAP dependence plot, each observation is plotted as a scatter-plot point; the y-axis corresponds to the SHAP value and the x-axis to the attribute's value. By defining a different color for each feature and showing them in a 2-D graph, we can visualize two variables' interaction effects.

For this study, the SHAP values were calculated and plotted in log-odds. Log-odds create a logistic transformation to the function, which provides visual attractiveness. When plotting the prediction's log-odds, we can see the effect between the feature inputs and the output value. With this unit, we can observe the change in the value of the target value when the predictor analyzed is changed by one log-odd, and all the other variables are fixed. When the ratio is greater than 1, it indicates that the event is more likely to happen as the independent variable increases. In contrast, when the odds ratio is less than 1, the event is less likely to occur as the independent variable increases.

As a second strategy to interpret the results, we used a mining rule-based patterns. The algorithm used to perform this was the PBC4cip algorithm [42]. We used this method to identify insights regarding the decision rules identified for better discrimination of the classes.

The PBC4cip algorithm is a model-specific method that uses an ensemble of decision trees and converts them into multivariate decision rules. As an example, a multivariate contrast pattern for income prediction could be the following: [IF Marital Status = Married AND Gender = Female AND Education = High School, THEN Class = Low].

Table 6 Regression models results of pseudo R2, quantile loss, adjustes R2, and root mean squared error for Current Income variable

	A			B	
	Pseudo R2	**Q-Loss**		**R2-adj**	**RMSE**
QR50	0.23	18,659.52	OLS	0.44	50,431.45
QRF50	0.37	14,719.26	RFR	0.51	47,325.67
QLGB50	0.38	14,301.58	GBC	0.54	45,892.69

Table 7 Classification models results of accuracy and area under the curve for Current Income variable

	C			D	
	Accuracy	**AUC**		**Accuracy**	**AUC**
LR	0.48	0.749	LR	0.79	0.870
RFC	0.50	0.762	RFC	0.82	0.890
GBC	0.53	0.796	GBC	0.83	0.910

During the training phase, the PBC4cip algorithm weights the sum of the supports in each of the classes as stated in Eq. 11, where C represents the number of instances belonging to the class c, T the number of instances in the training dataset, P the set of patterns found for the class c, and $Sup(p, c)$ the support of the pattern p into the class c.

$$w_c = \frac{1 - \frac{C}{T}}{\sum_{p \in P} Sup(p,c)} \tag{11}$$

Then in the classification stage, the sum of each class's supports is multiplied by the weight w_c of its corresponding class. This is done to punish the high sum of supports computed by the majority class. Then, the instance evaluated is classified based on the class with the highest value according to Eq. 12.

$$w(p,c) = w_c \sum_{p \in P} Sup(p,c) \tag{12}$$

Finally, in order to select the most relevant patterns, filtering is performed based on two constraints: support difference and confidence above a minimal defined threshold. If the support difference or the confidence is not large enough, it is assumed that the pattern is not worthy of consideration. The rules obtained from this model were filtered by considering only those which had a support difference between both classes 40% or higher and confidence above 65% to ensure the rules were relevant for the prediction task.

Results

In this section, we present the results for the four different experiments conducted in this paper for the prediction of current income.

Tables 6 and 7 summarizes the results of the different methods. All of the methods were tuned with respect to their specific parameters and hyper-parameters by using Grid Search, and they all considered a subset of the topmost important features selected in a pipeline process, using the SFFS method for the linear models and the RFE for tree-based methods.

When looking at the results, we identified that the performance was, in general, better for the GB Model, both for the regression and classification tasks. However, in the QR model's statistical analysis, we identified that the improvement of this algorithm was not significant.

To determine the significance, we used the data of the quantile loss obtained from the 10-fold CV for the three different models and performed a post hoc test to identify the pair of algorithms that do not have equal performance. For this statistical analysis, we used the post hoc Tukey HSD test. For Experiment A, the results of the Tukey HSD test showed there are no significant differences between the performance of the following models: QR50 and QGB50, we could not reject the hypothesis (p-value < 0.05) in any of the quantiles; thus, there was no sufficient statistical evidence to confirm that the results have a different distribution. Hence, by the parsimony theorem [43], we recommend using the traditional QR model for the QR approach. In contrast, the significance of the GB model in the traditional regression, the multi-class classification and the binary classification was significantly better than the rest of the models.

We integrated additional machine learning techniques in the binary approach and compared their accuracy scores in the 10-fold cross-validation. The current income model results for this approach are shown in Fig. 11. Based on the post hoc Tukey HSD test, we infer no significant differences within the following groups: GB and RF; SVC and LDA; KNN, DT and LR. All other differences were significant.

It can be observed in the results of the first income model shown in Tables 8 and 9, and in the results of the post hoc Tukey HSD test for the binary models (Fig. 12), that for this model our hypothesis that non-parametric methods can perform better in this data does not hold. The results indicated that the linear and logistic regression were the most adequate to describe the variables' relationship with first income after graduation.

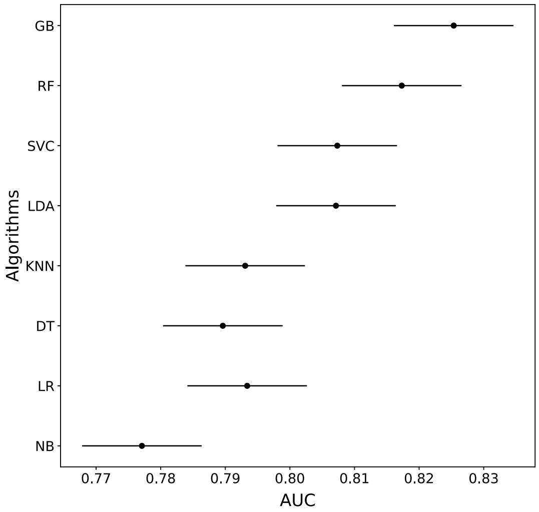

Fig. 11 Tukey HSD Post Hoc Test for the Binary Current Income classification models area under the curve results

Table 8 Regression models results of pseudo R2, quantile loss, adjusted R2, and root mean squared error for First Income variable

	A			B	
	Pseudo R2	**Q-Loss**		**R2-adj**	**RMSE**
QR50	0.13	2,266.01	OLS	0.20	8,554.05
QRF50	0.10	3,167.97	RFR	0.19	8,609.08
QLGB50	0.13	3,090.62	LGBR	0.17	8,710.96

Table 9 Classification models results of accuracy and area under the curve for First Income variable

	C			D	
	Accuracy	**AUC**		**Accuracy**	**AUC**
LR	0.42	0.681	LR	0.69	0.75
RFC	0.37	0.6377	RFC	0.66	0.72
LGBC	0.37	0.6335	LGBC	0.65	0.71

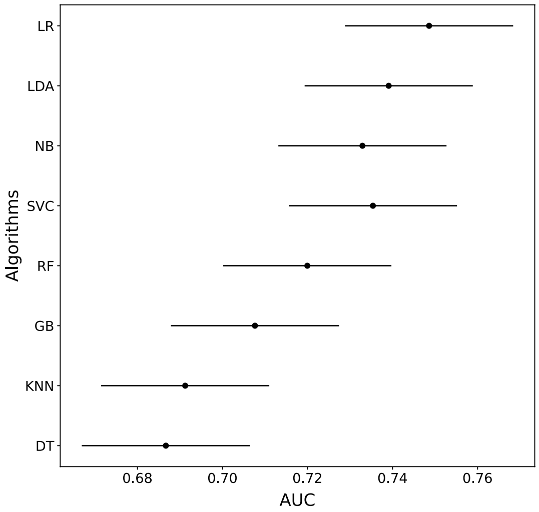

Fig. 12 Tukey HSD Post Hoc Test for the Binary First Income classification models area under the curve results

Feature importance

For the 'Current Income' model, the ranking of the most important features and their overall contribution was plotted in a SHAP-values graph. This graph shows in red the variables that negatively impact the model and in green the ones that impact positively. The graphs in Fig. 13 show the features that impact the class 'High'. This technique is similar to obtaining the coefficients in a linear model and can bring transparency to our machine learning model. In this graph, we can see how 17 of the subsets of variables impact the model positively. For instance, 'Age' is the most important variable for 'Current Income' and impacts in a positive way; the 'Gender' variable follows this, the number of 'Years worked Foreign' and the 'First Income' variable. On the other hand, working in the 'Tertiary Industry' sector is impacting negatively. An interesting insight

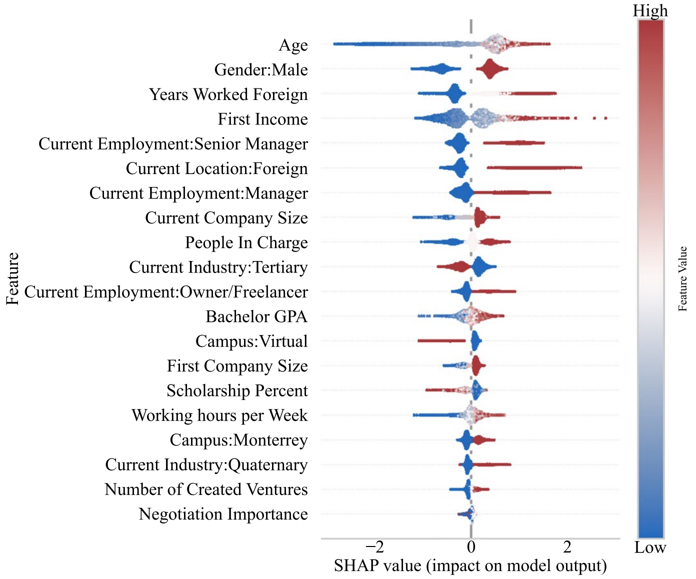

Fig. 13 GB Feature Importance with SHAP values

that can be noted is that having a high 'Scholarship' percentage during the alumni studies impacts negatively in their 'Current Income'. However, this is affected by the proportion of the people with a scholarship vs. alumni without a scholarship; therefore, there might be an unfair bias for this variable.

On the other hand, something positive that can be observed is that 'Bachelor GPA' affects positively in many scenarios. Finally, a proxy variable is showing up in this graph, the *Negative Importance* variable, which shows that overall, giving importance to negotiation skills can boost the income of the alumni.

Regarding the 'First Income' model, a feature importance plot was obtained based on the estimated coefficients of the 13 selected features for the LR model. The plot is shown in Fig. 14. When taking a close look at the coefficients, we can see that the features which impacted the model the most were the size of the company and having attended Engineering school, and having worked Foreign in their first job, all of these variables impacted positively, whereas working in the quaternary sector, having lived in Mexico in a region different from the North or Central area, and being a First-Generation student impacted negatively in the income-class prediction.

Exploring feature interactions

The SHAP partial dependence plots exhibit the marginal effect between two features on predicting the target variable. This visually shows covariates' relationships with the target variable besides being linear, monotonic or more complex. This section used SHAP partial dependence plots to show the stronger covariate relationships with income for the GB model for 'Current Income'.

Figure 15 shows a strong interaction between the 'First Income' feature and Age. When looking at the observations, we can also notice this interaction, which is stronger for those former students between 35 and 40, seconded by those between 30 and 35. The

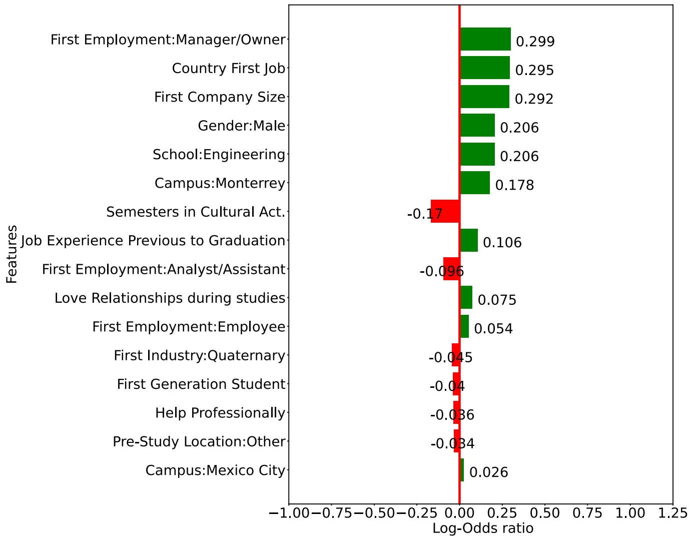

Fig. 14 First Income LR Feature Importance with coefficient weights

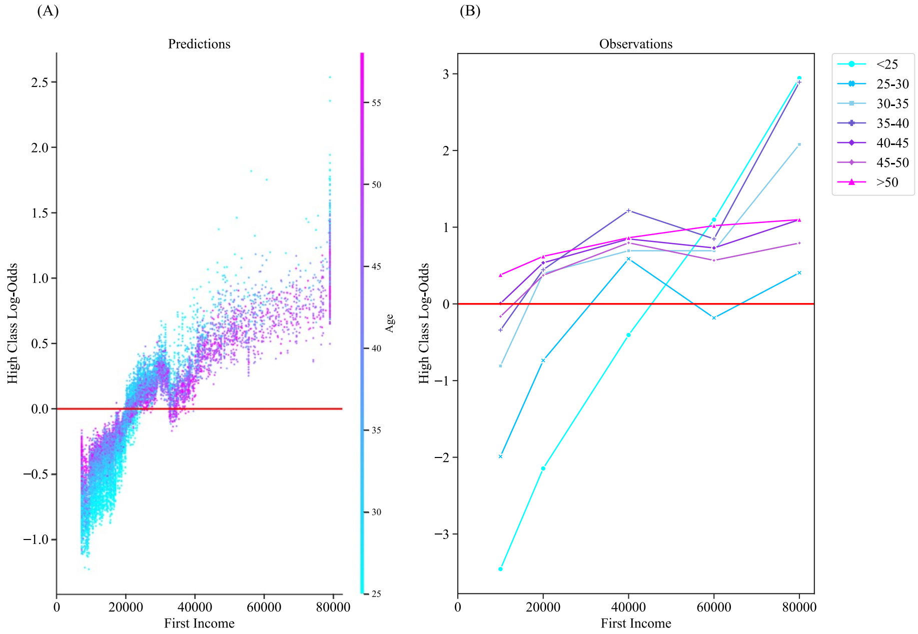

Fig. 15 First Income and Age SHAP dependency plot of the gradient boosting model, where **A** are the predictions and **B** are the real observations

interactions seem to be very weak for those older than 50 years old, which seems logical as these are people that graduated a long time ago, and thus the impact of their 'First Income' is not so relevant. Thus there is much likely a mix of other features not shown in this dataset that explain their variation.

The second interaction implies the relation between Working Hours and the Current Location of the alumnus. We can observe from Fig. 16 hat there is a linear relationship for working hours per week, which reaches a plateau. However, there is an interaction

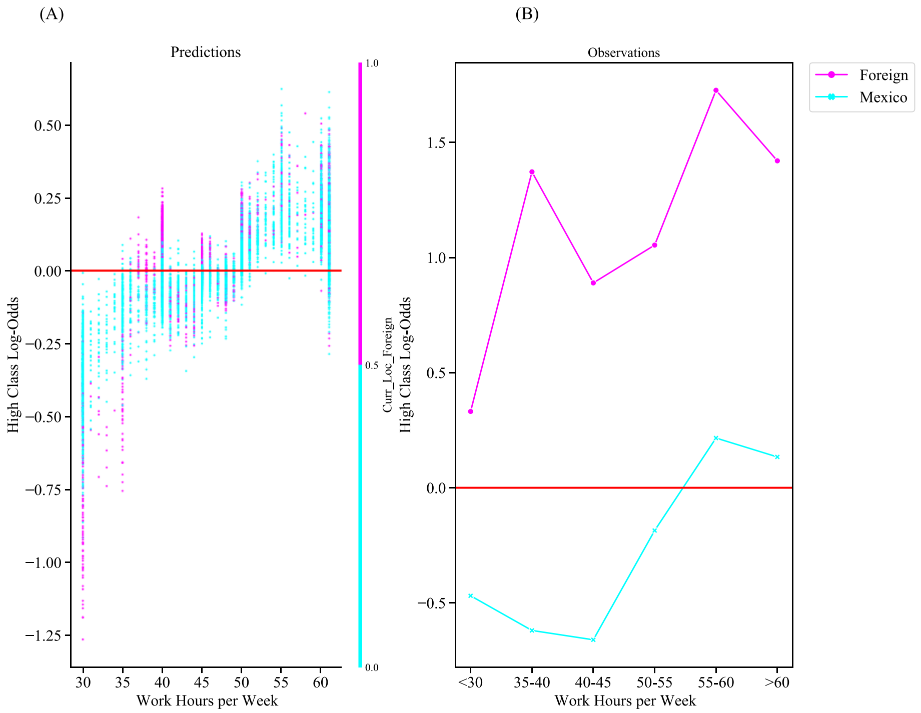

Fig. 16 Working hours per week and Current Location SHAP dependency plot, where **A** shows the predictions of the gradient boosting model and **B** the current observations

of the alumni's location. The predictions show that those working between 35 and 40 h reach the peak if they live outside Mexico; yet, those living in Mexico do not reach this peak until working between 55 and 60.

Another interesting insight that can be explained with this model is the interaction between gender and age. Graph B of Fig. 17 depicts how female alumni have a positive linear interaction when they are between 20 and 30 years old. Then this plateaus for the following years. On the other hand, graph A obtained from the model predictions shows how after 30, the variable of Gender impacts negatively in female alumni. Furthermore, the graph can be interpreted as gender having a negative effect on female alumni. We can see that the gap between Male and Female odds increases as the alumnus is older.

With the graphs presented in this section, we have identified that the most important variables for Income prediction identified by our model do not affect solely but interact with other covariates. The graph presented show the primary interaction relationships for the predictions along with a comparison with the observed data points versus log-odds.

Mining contrast patterns

The last technique used to analyze the 'Current Income' prediction model results was the contrast-pattern extraction with PBC4cip. In this section, an experiment was conducted to obtain contrast patterns that could give additional insight regarding the data analyzed.

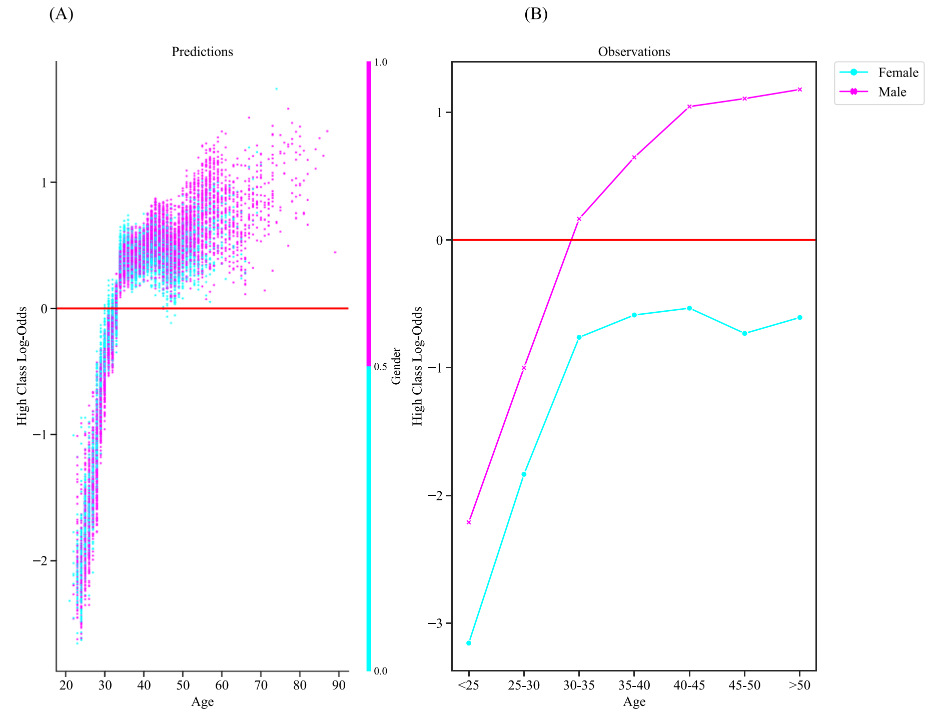

Fig. 17 Gender and Age SHAP dependency plot of the gradient boosting model, where **A** are the predictions and **B** are the observations

One of the advantages of machine learning techniques over linear models is accounting for interactions between features. PBC4cip constructs rules by decomposing decision trees in a RF model, and any path that leads to a node can be transformed into a decision rule. The advantage of this is that the rules created are easy to interpret because, in our problem, they are binary decision rules. A limitation of data mining is that although it can identify patterns that are not obvious from data, not all of the patterns extracted may be useful. This is the most reason why data mining requires human intervention.

The rules obtained from this model were filtered by considering only those which had a support difference between both classes 40% or higher and confidence above 65%. This ensures that the rules are relevant for the prediction task. In addition, the redundant atoms obtained in the extracted patterns were removed with the automation filter in PBC4cip. Three patterns were obtained that complied with these constraints and are shown in Table 10. This table shows that the set of patterns extracted each contain three features. To better comprehend the mathematical representation of the contrast patterns obtained, we used bar plots to visualize these three variables' impact on the 'High' class. The bar plots are shown in Figs. 18, 19 and 20.

In the first visualization, we can observe how studying an Engineering or Business bachelor degree, having a job title different from Employee and being older than 28 years old gives the alumnus a higher probability for the 'High' class. As noted in Table 10, the support of this pattern for the 'High' class is 74%. This means that the pattern describes 74% of the observations with class 'High' (from the total dataset of 12,275 observations,

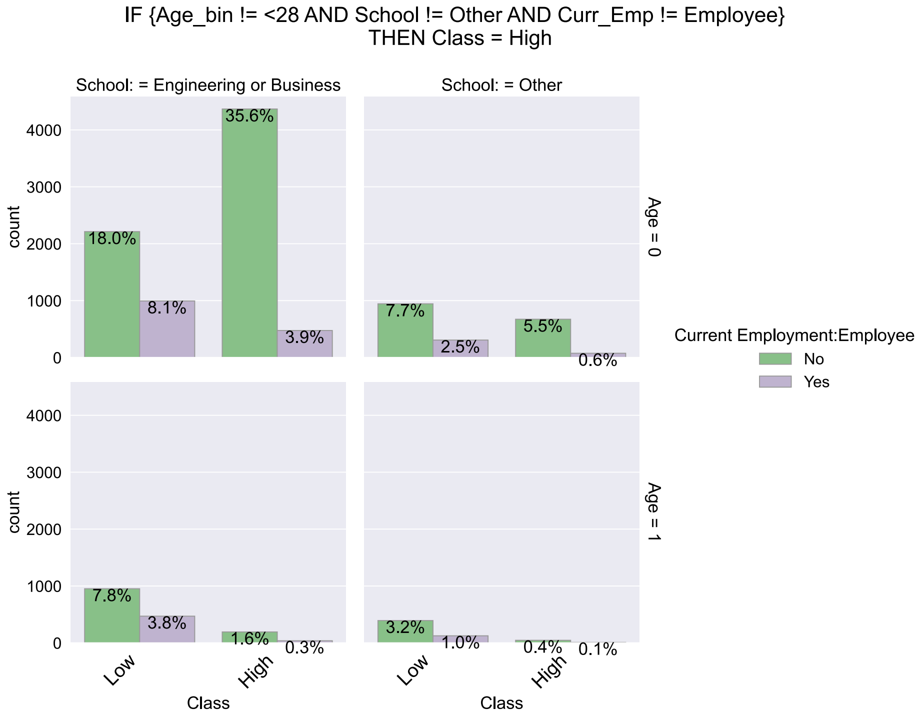

Fig. 18 Visualization of the first contrast pattern of Age, School, and Current Employment variables in bar plot count

where 5877 belong to class 'High', 4396 objects comply with this pattern). This pattern has a great coverage since the observation that it describes represents 35.6% of the overall observations in the dataset. Furthermore, the pattern confidence indicates that the probability that an object fulfills the property class 'High' given that the object fulfills the pattern is 64%.

Next, the second visualization shows how being a Male alumnus, having a job title different from Employee, and being older than 28 years old, gives an alumnus a higher probability for the 'High' class. As noted in Table 10, the support of this pattern for the 'High' class is 61%, which indicates that the pattern describes 61% of the observations in class 'High'. There is also an extensive coverage since these observations represent 29.3% of the overall observations. The confidence given to this pattern is slightly higher than the previous one; it is 70%.

Finally, the visualization in Fig. 20 depicts that having People in Charge, being older than 28 years old having a job title different from Employee also gives alumnus higher probabilities for the 'High' class, with support of 70%. Moreover, this pattern covers 33.7% of the observations in the dataset. The confidence for this pattern is 66%.

In this analysis, we mined three important patterns to contrast the two classes in our target variable, 'Current Income'. The variables that became evident in the obtained patterns were: School of Bachelor Degree, Job Title, Age, having people in charge and Gender. While the first three are understandable variables to explain income, the latter variable has made evident the gender bias for the 'High' Class in the alumni population.

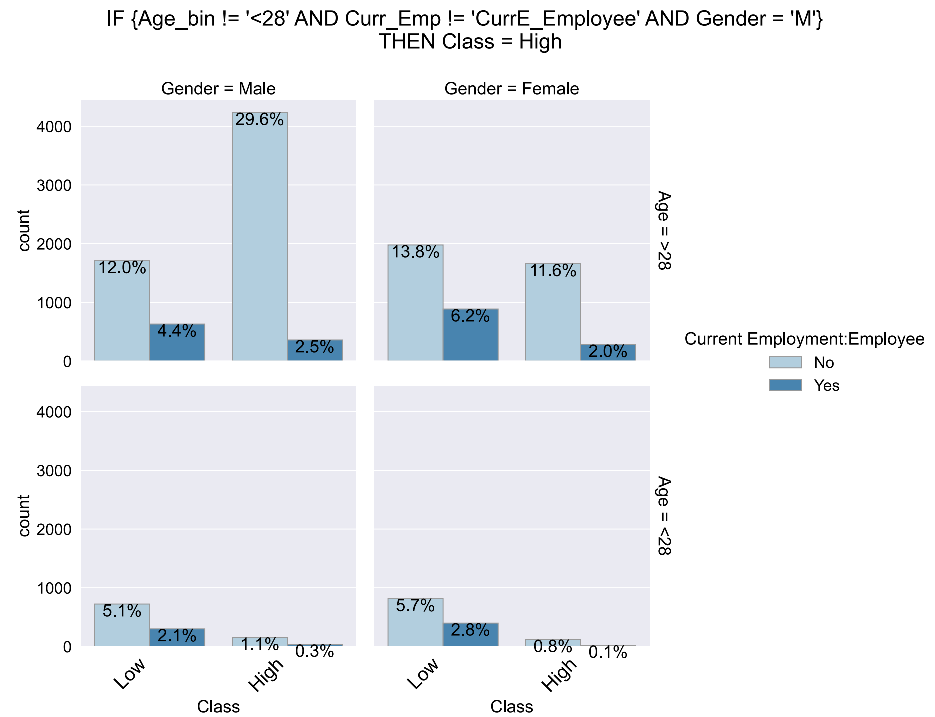

Fig. 19 Visualization of the second contrast pattern of binarized Age, Current Employment, and Gender variables in bar plot count

We note that the factors Age and the Job Title appear in all the patterns, and each one has a distinct variable. We can also see that the pattern which receives the most confidence is pattern 2, is where the distinct variable is gender; this shows us the importance that gender has been for the tree-based miner to determine the class.

Discussions

When comparing the related work's results with the results from this thesis, we can see that in the QR approach, our results achieved better pseudo-R2 results than those obtained by Lee and Lee [7]. In contrast, the results from Figueiredo and Fontainha [10] had considerably better results. The main variables that were detected by the researchers and that were not available in the data set used in this thesis were: marital status, children status, the observation's firm's foreign capital, and the years of tenure at the current employer. Including these variables in the analysis as future work could serve positively to our model's performance.

For the traditional regression model approach, the results obtained by this study for the current income model were significantly better than the literature analyzed. Both of the analyzed related work used OLS to predict income, and with this study, we have shown that the decision-tree ensembles can yield significantly better results.

Unfortunately, in relation to the multi-class classification model, our results were worse than the literature analyzed. Khongchai and Songmuang [12] obtained the best results using K-Nearest-Neighbours and Chen et al. [13] using Decision Trees. The

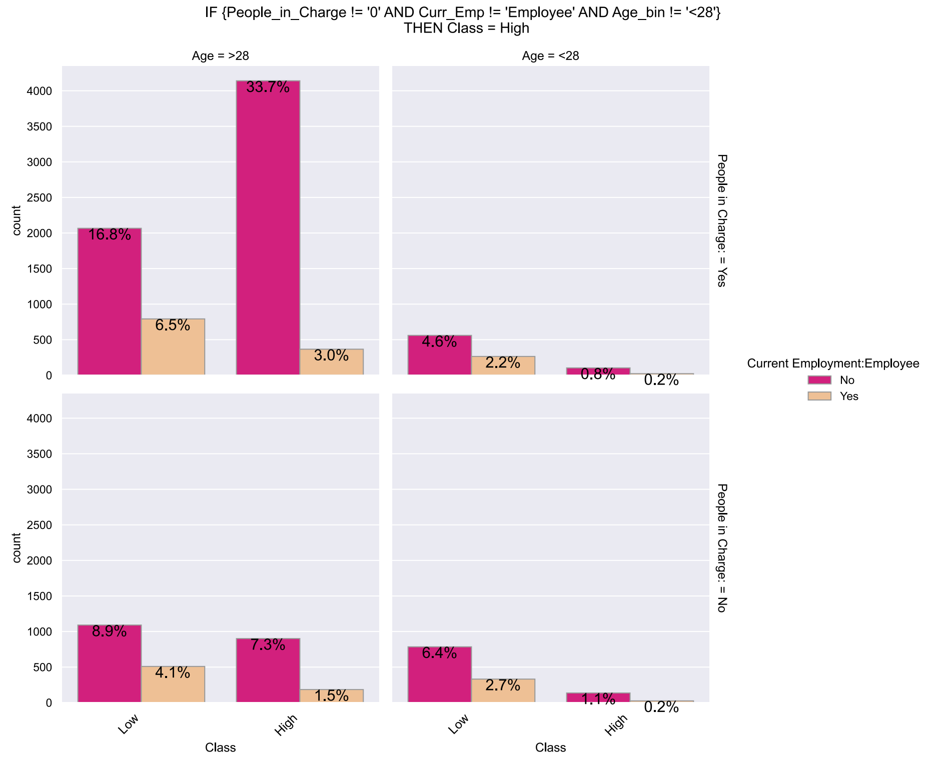

Fig. 20 Visualization of the third contrast pattern of People in Charge, Current Employment, and binarization of Age variables in bar plot count

Table 10 PBC4cip three contrast patterns for current income high class

ID	Pattern	Support by Class		Confidence
		Low	High	
CP1	IF Age_bin != '<28' AND School != 'School_Other' AND Curr_Emp != 'CurrE_Employee' THEN Class = 'High'	0.32	0.74	0.64
CP2	IF Age_bin != '<28' AND Curr_Emp != 'CurrE_Employee' AND Gender = 'M' THEN Class = 'High'	0.21	0.61	0.70
CP3	IF People_in_Charge != '0' AND Curr_Emp != 'CurrE_Employee' AND Age_bin != '<28' THEN Class = 'High	0.3	0.7	0.66

former research used these additional dependent variables, which were not available in the dataset analyzed in this thesis: specific degree program and type of work performed in the company. While similar variables are included in this study, the analyzed study variables consider more characteristics about the type of work that the students performed; other patterns could have been identified with the specific degree and work type. Therefore, this thesis hypothesis that follow-up research with more data can build models based on the specific degree and the type of work of the alumni. The latter research considered features provided by job descriptions from job posting sites. This included more work-related features such as location, contract versus permanent type, job content and job relationship features. While this work's objective was different from this thesis, it provided insights into how the detail of the job that the individual performs can be effectively used to predict their income.

Finally, for the binary model, this thesis obtained very similar results to related work. Lazar [5] showed that SVM could achieve high performance when predicting income. The author used the following predictors that were different from those used in this thesis: work class, marital status, race, capital gain, and capital loss. Further work can be done by including these variables in our GBC model to improve the performance. On the other hand, Sharath [11] achieved good results with boosted trees and various demographics as predictors; however, our study achieved significantly better performance with the GBC and the variables identified as the most important for income prediction.

Conclusions and future work

With the appearance of the digital transformation and the big data era, advanced analytics and data science has been increasingly used in many industries. In education, it has been used to improve the learning process and evaluate academic institutions' efficiency. In econometric sciences, these techniques have been used to explain the links between economic, financial and social effects. The differences between data analytics and data science are mainly that the latter makes use of machine learning techniques. These methods can provide more accurate predictions than the traditional statistical models used in data analytics. Nonetheless, these methods' do not provide a clear interpretation of individual factors compared to conventional statistical methods. Consequently, data analytics continues to dominate in education and econometric studies because of the ease of interpretation and the ability to distinguish variable effects.

In this study, we show an application of the data science project life cycle to predict and identify the variables with a strong relationship with alumni income. For this, we use 'the CRISP-DM methodology'. We followed the standard steps in the strategy and implemented additional steps to explain the results. Given this, we illustrate the flexibility that CRISP-DM can provide to data science projects based on the business's needs or research.

We showed the importance of cleansing and transforming the data during this project's data understanding and preparation phase. Before modelling, we showed the importance of an exploratory analysis to understand the data, detect bias and identify specific pre-processing needs through the cleansing and transformation process. The data exploration included descriptive statistics, visualizations through box-plots, correlation analysis, and the application of hypothesis testing for comparing two-factor levels and determining marginal effects of the independent variables with income.

We compared different modelling techniques based on a distinction between parametric and non-parametric models during the modelling phase, and utilized XAI techniques to interpret the results. The purpose of the study was to investigate the relationship between the target variables 'Current Income' and 'First Income' with demographical attributes obtained from an alumni survey. For this purpose, this research created and analyzed several machine learning methods to predict the first income after graduation and former students' current income.

This study identified that for the best performing classification task, which discerns between low and high earners, the top most important variables were: years worked foreign, first income, age, employment title, gender, employer's characteristics (company size, industry), the number of people in charge, the bachelors GPA, and the working hours per week. While most of these variables are control variables, we identified the following actionable variables: bachelor's GPA, years worked foreign, working hours per week and first income after graduation. Hence, these variables can be paid more attention by those students seeking to achieve a high expected salary. Furthermore, this study's insights can be used to influence changes in the work sector and academic institutions, mainly to drive salary transparency and reduce the gender wage gap.

There are some interesting directions in which this work could be extending:

1. In this study, we only focused on comparing traditional econometric algorithms with ensemble tree-based algorithms; it will be interesting to learn the performance of neural networks and explore the power of XAI techniques in deep learning.
2. Other educational institutions can use the methodology followed in this study to perform a similar analysis to evaluate their alumni outcomes, identify bias, and provide them additional opportunities for obtaining their expected earnings.
3. Future work can consider the variables identified as more important in this study and augment the variables provided in the related studies analyzed, such as marital status, children status, as well as more job-related characteristics.

Abbreviations

SHAP: Shapley Additive exPlanations; CSP: Current Population Survey; PCA: Principal Component Analysis; OLS: Ordinary Least Squares; R2: Coefficient of determination; SVM: Support Vector Machines; NB: Naive Bayes; C4.5: C4.5 Machine Learning Algorithm; DT: Decision Tree; MLP: Multilayer Perceptron; KNN: K-Nearest Neighbors; LR: Logistic Regression; RF: Random Forest; GBM: Gradient Boosting Machines; GPA: Grade Point Average; MSE: Mean Square Error; QR: Quantile Regression; LSTM: Long Short-Term Memory; NN: Nearest Neighbours; DNN: Deep Neural Network; FAFSA: Free Application for Federal Student Aid; US: United States; CRISP-DM: Cross-Industry Standard Process for Data Mining; τ: Quantile; $\hat{C}(x)$: Classification given x; $\hat{C}_b$: Classification for sample b; $\hat{f}(x)$: Regression given x; *B*: Number of times the bagging aggregation is applied; $f_{b(x)}$: Predictor for b samples; α: Learning rate; $f_i^{(j)}$: The ith predictor; VIF: Variance Inflation Factor; MXN: Mexican Pesos; *r*: Correlation coefficient; F: Female; M: Male; LDA: Linear Discriminant Analysis; RFC: Random Forest Classifier; GBc: Gradient Boosting Classifier; SFFS: Sequential Floating Selection; RMSE: Root Mean Square Error; RFE: Recursive Feature Elimination; AUC: Area Under the Curve; *k*: Number of features; *n*: Sample size; R^1: Pseudo-R-square; $\hat{V}$: Pseudo-R-square for an unrestricted quantile; $\tilde{V}$: Intercept-only model; $L(y, t)$: The pinball loss function; *y*: Real value; *t*: Quantile forecast; ROC: Receiver Operating curve; TPR: True Positive Rate; FPR: False Positive Rate; TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative; XAI: Explainable Artificial Intelligence; PDP: Partial Dependence Plot; CV: Cross-Validation; HSD: Honestly significant difference; Q-Loss: Quantile Loss; PBC4cip: Machine learning classifier package based on contrast patterns; CP1: Contrast Pattern 1.

Acknowledgements

The authors express their thanks to the School of Engineering and Science, Tecnológico de Monterrey, Monterrey, México, for providing the data for the analysis and to the PNPC CONACYT program with its financial support for DGC and RDR to study their master's degree.

Authors' contributions

DAG analyzed and interpreted the alumni data regarding their income, performed the experiments reported, and was a major contributor in writing the manuscript. RED performed a review and examination of the state of the art, and contributed in writing the manuscript. JLP provided valuable guidance for the statistical analysis and machine learning modeling. HGC and NH-G collected the data used in the manuscript and provide feedback regarding the writing of the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The data that support the findings of this study are available from Tec de Monterrey, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. The authors, however, may provide a sample with synthetic data with the features used in this study upon request.

Declarations

Ethics approval and consent to participate

This article does not contain any studies with human participants or animals performed by any authors.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Received: 9 September 2021 Accepted: 10 January 2022
Published online: 24 January 2022

References

1. Pace CR. Measuring outcomes of college. Fifty years of findings and recommendations for the future. Washington: ERIC; 1979.
2. Delaney AM. Ideas to enhance higher education's impact on graduates' lives; alumni recommendations. Tert Educ Manag. 2004;10(2):89.
3. Volkwein JF. Assessing alumni outcomes. New Dir Inst Res. 2010;2010(S1):125.
4. Rode JC, Arthaud-Day ML, Mooney CH, Near JP, Baldwin TT. Ability and personality predictors of salary, perceived job success, and perceived career success in the initial career stage. Int J Select Assess. 2008;16(3):292.
5. Lazar A. Income prediction via support vector machine., In: ICMLA. Citeseer; 2004, p. 143–149.
6. Webbink D, Hartog J. Can students predict starting salaries? Yes! Econ Educ Rev. 2004;23(2):103.
7. Lee BJ, Lee MJ. Quantile regression analysis of wage determinants in the Korean labor market. J Korean Econ. 2006;7(1):1.
8. Oehrlein P. Determining the future income of college students. Undergrad Econ Rev. 2009;5(1):7.
9. Strand M, Truong T. Predicting Student Earnings After College Predicting Student Earnings After College
10. Figueiredo MdC, Fontainha E. Male and female wage functions: a quantile regression analysis using LEED and LFS Portuguese Databases, ISEG-Departamento de Economia. 2015.
11. Sharath R, Nirupam KN, Sowmya BJ, Srinivasa KG. Data analytics to predict the income and economic hierarchy on census data, In 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS) (IEEE, 2016), pp. 249–254.
12. Khongchai P, Songmuang P. Improving students' motivation to study using salary prediction system, In 2016 13th International Joint Conference on Computer Science and Software Engineering, JCSSE 2016 2016. https://doi.org/10.1109/JCSSE.2016.7748896
13. Chen L, Sun Y, Thakuriah P. Modelling and Predicting Individual Salaries in United Kingdom with Graph Convolutional Network, In: International Conference on Hybrid Intelligent Systems. Springer; 2018, p. 61–74.
14. Larose DT. Discovering knowledge in data: an introduction to data mining. 2005. https://doi.org/10.1002/0471687545.
15. Leventhal B. An introduction to data mining and other techniques for advanced analytics. J Direct Data Digit Market Pract. 2010;12(2):137.
16. Koenker RW. Quantile regression (Econometric Society Monographs; No. 38). Cambridge University Press; 2005.
17. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. Newton: O'Reilly Media; 2019.
18. Kotsiantis SB. Decision trees: a recent overview. Artif Intell Rev. 2013;39(4):261.
19. Burkov A. The hundred-page machine learning book. Quebec: Andriy Burkov Canada; 2019.
20. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. New York: Springer Series in Statistics; 2001.
21. Efron B, Hastie T. Computer age statistical inference. Cambridge: Cambridge University Press; 2016.
22. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520.
23. Dixon WJ, Yuen KK. Trimming and winsorization: a review. Statistische Hefte. 1974;15(2–3):157.
24. Embrechts P, Schmidli H. Modelling of extremal events in insurance and finance. Zeitschrift für Oper Res. 1994;39(1):1.
25. Ibragimov R. Heavy-tailed densities. London: The New Palgrave Dictionary of Economics Online; 2009.
26. McNeil AJ, Frey R, Embrechts P. Quantitative risk management: concepts, techniques and tools-revised edition. Princeton: Princeton University Press; 2015.
27. Ibragimov R, Prokhorov A. Heavy tails and copulas: topics in dependence modelling in economics and finance. Singapore: World Scientific; 2017.
28. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. Melbourne: OTexts; 2018.
29. Myatt GJ, Johnson WP. Making sense of data II: a practical guide to data visualization, advanced data mining methods, and applications. Hoboken: Wiley Online Library; 2009.
30. Efroymson MA. Multiple regression analysis, Mathematical methods for digital computers. 1960;191–203.
31. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1–3):389.
32. Kuhn M, Johnson K. Feature engineering and selection: a practical approach for predictive models. Boca Raton: CRC Press; 2019.

33. Miles J. R squared, adjusted R squared. Wiley StatsRef: Statistics Reference Online; 2014.
34. Aggarwal CC. Data mining: the textbook. Berlin: Springer; 2015.
35. Koenker R, Machado JAF. Goodness of fit and related inference processes for quantile regression. J Am Stat Assoc. 1999;94(448):1296.
36. Kurzawa I, Lira J. The application of quantile regression to the analysis of the relationships between the entrepreneurship indicator and the water and sewerage infrastructure in rural areas of communes in Wielkopolskie Voivodeship. Metody Ilościowe w Badaniach Ekonomicznych. 2015;16(2):33.
37. Steinwart I, Christmann A. Estimating conditional quantiles with the help of the pinball loss. Bernoulli. 2011;17(1):211.
38. Zhang F, Fan X, Xu H, Zhou P, He Y, Liu J. Regression via Arbitrary Quantile Modeling, arXiv preprint arXiv:1911.05441 2019.
39. Rai A. Explainable AI: from black box to glass box. J Acad Market Sci. 2020;48(1):137.
40. Shapley LS. A value for n-person games. Contributions to the theory of games. Ann Math Stud. 1953;2(28):307.
41. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;1189–1232.
42. Loyola-González O, Medina-Pérez MA, Martínez-Trinidad JF, Carrasco-Ochoa JA, Monroy R, García-Borroto M. PBC-4cip: a new contrast pattern-based classifier for class imbalance problems. Knowl Based Syst. 2017;115:100.
43. Bordens KS, Abbott BB. Research design and methods: a process approach. New York: McGraw-Hill; 2002.

Publisher's Note

Terms and Conditions