IE School of Human Sciences & Technology

Master's in Big Data & Business Analytics

Course Name:

Machine Learning I (Supervised Assignment)

Supervised by:

Prof. Álvaro José Méndez López

Submitted by:

MANUEL MARINA HERRERA

EXECUTIVE SUMMARY

The purpose of this assignment is to determine and predict income of prospective customers of a bank, since this data will not be available anymore in the future. The bank management provided us a dataset of existing customers, containing 32 features of 499 different clients. With our model, we were able to explain 71.6% of income variation in the dataset. The most significant variables are “retire”, “employ” and “ed” (years of education). The implications are that being employed and having more years of education positively impacts “income” while being retired significantly lowers the customers predicted income.

DATA PREPARATION

In order to be able to include categorical variables in our regression analysis in a valid format, we created dummy variables for the following variables:

- Card: amex, visa, master, discover, other.
- Jobcat: cat1, cat2, cat3, cat4, cat5, cat6.
- Polview: ex-lib, sl-lib, lib, moderate, sl-con, con, ex-con.
- Jobsat: highly dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, highly satisfied.

Even though job satisfaction (jobsat) is an ordinal variable and therefore on a scale, we decided to create dummy variables, because there is no meaningful interpretation between each unit of the variable.

As “reason” has mostly missing values, the variable was not part of the regression analysis and got dropped.

After creating dummy variables, we ran a step-wise method to obtain a set of significant variables and with those we ran a linear regression model. At first sight it seemed like income could be best described with a linear model as the R^2 was 0.91. Nevertheless, the scatter plot of the dependent variable did not look linear (Appendix, Figure 1) and R^2 of the validation dataset was only 0.51. Therefore, we removed a clear outlier that could “hide” a nonlinear relationship. In order to see the scatterplot without the outlier, we run a new regression with the same significance variables. The scatter plot of predicted and actual income values clearly show a nonlinear relationship (Appendix, Figure 2 & 3).

We are aware of the fact that multicollinearity is not a problem for prediction models. Nevertheless, we checked for correlation between all significant variables by calculating a correlation matrix. No high correlation was found (Appendix, Figure 4), therefore MLR 3 (no perfect collinearity) holds.

Knowing that we have a nonlinear regression, it is mandatory to convert it to a linear one, manipulating our dependent variable, income, into log(income). As we have a new dependent variable we run the stepwise method again with log(income) as the dependent variable.

MODEL DESCRIPTION

The relevant variables for our model are: creddebt (credit debt in thousands), retire (retired), employ (years with current employer), ed (years of education), address (years at current address), cardspent (amount spent on primary card last month), cat4 (agricultural and natural resources), hometown (home ownership), amex (American Express credit card), cat 1 (managerial and professional), cat 2 (sales and office), highly satisfied (job satisfaction) (Appendix, Figure 5).
Before developing the model we split the dataset into 80% model (training) data and 20% test (cross-validation) data.
Using Dataiku, we selected "Auto ML Prediction" from the Lab including only the significant variables that we found previously.
Regarding missing values, instead of dropping them, we imputed the mean. For the training data, we followed the same approach as in class:

- Policy: explicit extracts from the dataset.
- Train set: with random(approx. ratio) with 80%.
- Test set: with random (approx. ratio) with 20%.

After setting these features, we trained the model. In order to validate it we used an iterative process to get the highest R^2 (where the variation of the model is explained by the exogenous variables), but also one that's interpretable. As we know R^2 is an estimator that may not reflect "goodness of fit" for the whole dataset (there's a risk of overfitting). Therefore, we cannot only rely on the value of R^2, but the model needs to have some business interpretation. To find the best model, our algorithm of choice was the Ordinary Least Squares (OLS) method. Our final regression model with variables and settings as described above scored an R^2 of 0.788 (Appendix, Figure 6).

VALIDATION

The first evaluation step is to check how well our trained model makes predictions on the test-set of the training data. The histogram of the error indicates a normal distribution around the median and mean of zero (Appendix, Figure 7). The scatter plot appears to have a random distribution around the regression line (Appendix, Figure 8). Therefore, important MLR assumptions hold (MLR 4: zero conditional mean and MLR 6: normality of error terms). Once we were confident with our model, we cross-validated it with the 20% reserved to see how well it predicted our dependent variable log(income). We used the evaluate option in Dataiku, creating as output two data frames, one with the predicted log(income) and another one with the metrics of the evaluation. Our evaluated model had an R^2 of 0.63 showing a prediction really similar to our trained model. Also, we computed a histogram of the error and the scatter

plot prediction of the dependent variable (Appendix, Figure 9 & 10). Again, we got an error histogram following a normal distribution around zero and a linear regression plot with data points around the ideal regression line. The combination of $R^2$and the graphs show that our trained model is predicting well when it is applied over a new dataset so we can accept our model. As an additional check, we evaluated the whole initial dataset getting similar results (0.716 R^2, a normalized error histogram and data distributed around the ideal regression line).

INTERPRETATIONS & CONCLUSIONS

Our Linear Regression analysis enabled us to identify the different variables that help predict the income of future customers. Based on our findings, below, we provide an understanding of each variable's impact. The top three most relevant variables (largest absolute t-student test value) for our model are retire, employ and ed (Appendix, Figure 12). The variables with the biggest effect (largest standardized coefficient in absolute terms) on the dependent variable are retire, employ, creddebt and ed (Appendix, Figure 13).

The following interpretations are made on average, holding everything fixed (c.p.) and within a 95% confidence interval.

Ed: One extra year of education increases log(income) by 1.8% to 2.8%. This makes sense, as one more year of education means more qualification and expertise in the respective field.

Jobcat:

- Cat1 (managerial positions) increases log(income) by between 4.1% to 12.5%.
- Cat2 (sales and office job positions) increases log(income) by between 2.1% to 11.4%.
- Cat4 (agricultural and natural resources work) decreases log(income) by between 5% to 20.9%.

Job category and the amount earned is related to the amount of skills needed to perform a job. A person working in a low-skill job such as cat4 (agricultural and natural resources) will earn less than someone working in an occupation that requires more skills and therefore more years of education, such as cat1 (managerial and professional).

Employ: One extra year of employment with the same employer increases log (income) by 1.8% to 2.4%. Remaining loyal to the company increases the likelihood to climb the corporate ladder and increase responsibility and salary, therefore, it makes sense to expect greater income with greater company loyalty.

Retire: Being retired decreases log(income) by 58.7% to 69.9%. Naturally, whenever someone retires, the expectancy to generate income decreases, which is why being retired seems to indicate that lower income is to be expected.

Creddebt: Having 1000 additional euros of credit debt increases log(income) by between 1.6% to 2.4%.

Cardspent: The amount spent on the primary card in the last month seems to be irrelevant to increase the dependent variable as our analysis shows that it has almost no impact on log(income) (0.02%). This variable is still significant, even though the confidence interval goes from 0.000 to 0.000, because we can only see three decimals.

Jobsat (job satisfaction): Having a high satisfaction at your job increases log(income) by between 0.7% to 9.4%. Naturally, having a higher satisfaction at your workplace has a positive impact on productivity which in return increases the likelihood to generate greater income.

Homeown (home ownership): Owning your house increases log (income) by between 3.7% to 10.4%. Buying a house is usually done if the owner is in a comfortable and stable financial and professional situation, thus, it seems logical to expect higher income when owning a house.

Address: Living one more year at the same address increases log (income) by 0.2% to 0.5%. This could be a case of endogeneity (explanatory variable correlated with the error term): living longer in the same house could be a sign for a stable social environment and a stable job - factors that usually relate to a higher income. All these influences may be hard to measure and data on this was not in the data set. Therefore, these influences are being picked up by the error term. This is the reason why address may be correlated with the error term (= endogeneity).

Amex: Owning an American Express credit card decreases log(income) by between 1.8% to 9.5%.

APPENDIX

Figure 1: Scatter plot of the first regression

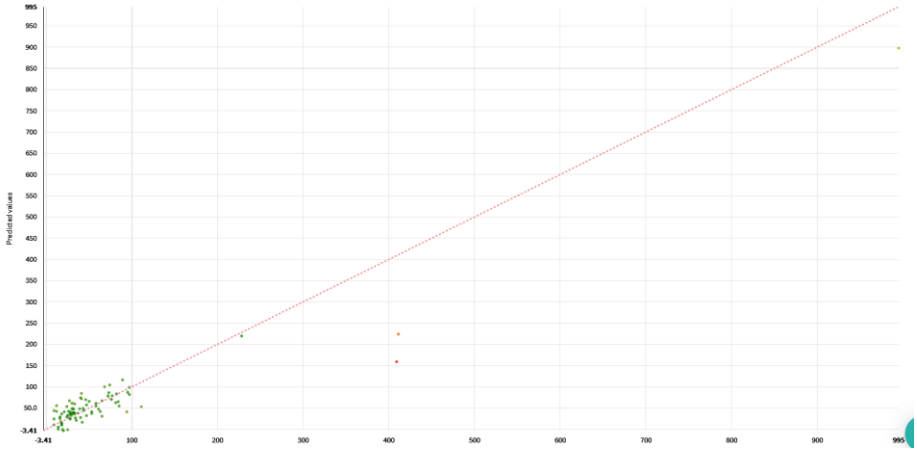


Figure 2: Scatter plot without the outlier

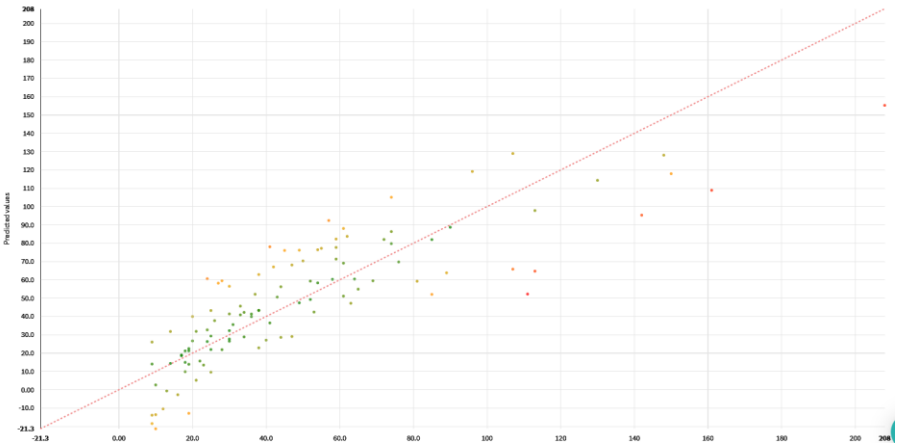


Figure 3: Scatter plot regression without the outlier

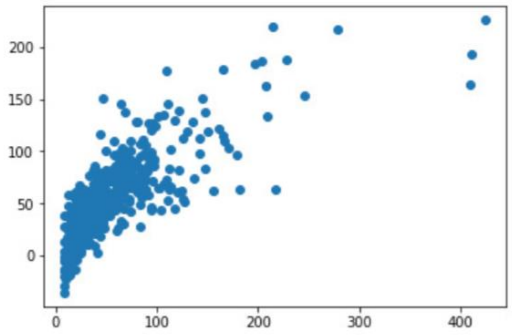


Figure 4: Correlation matrix

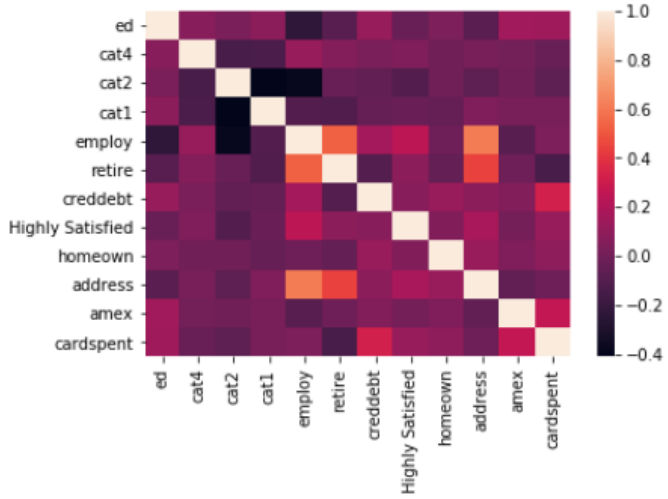


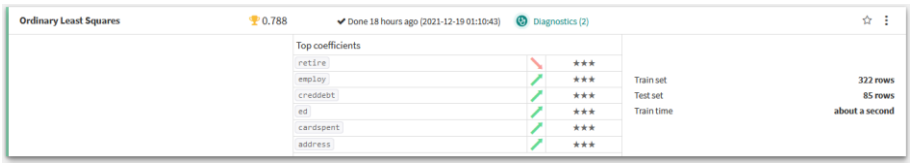
Figure 5: Significance variables obtained by step-wise method with log(income)

```
# Fills X (removing variables with missing values)
selected_fields=df.drop(labels=["polview","custid","card", "incom", "jobcat","jobsat", "reason","Log(income)"],axis=
# Sets y
y = df['Log(income)']

result = stepwise_selection(selected_fields, y)
```

```
Add  creddebt                  with p-value 3.42626e-32
Add  retire                    with p-value 1.37702e-20
Add  employ                    with p-value 3.50813e-47
Add  ed                        with p-value 2.48205e-17
Add  address                   with p-value 2.82793e-09
Add  cardspent                 with p-value 4.46005e-07
Add  cat4                      with p-value 0.00010819
Add  homeown                   with p-value 0.000371719
Add  amex                      with p-value 0.00396667
Add  cat1                      with p-value 0.00722139
Add  cat2                      with p-value 0.0050241
Add  Highly Satisfied          with p-value 0.0223848
```

Figure 6: Performance metrics



Detailed metrics

Explained Variance Score	**0.78795**	Root Mean Squared Error (RMSE)	**0.14988**
Mean Absolute Error (MAE)	**0.11791**	Root Mean Squared Logarithmic Error (RMSLE)	**0.057682**
Mean Absolute Percentage Error	**7.58%**	Pearson coefficient	**0.88774**
Mean Squared Error (MSE)	**0.022463**	R2 Score	**0.78795**

Figure 7: Histogram of errors of the test-set of the training data

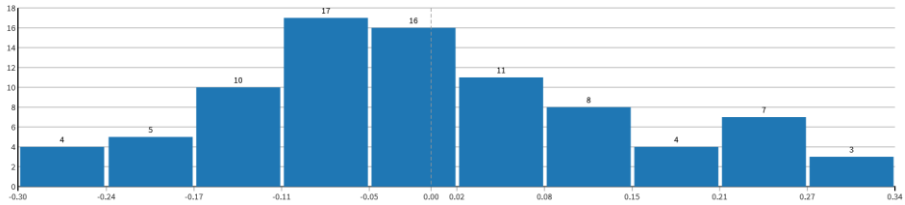


Figure 8: Scatter plot of the test-set of the training data with log(income)

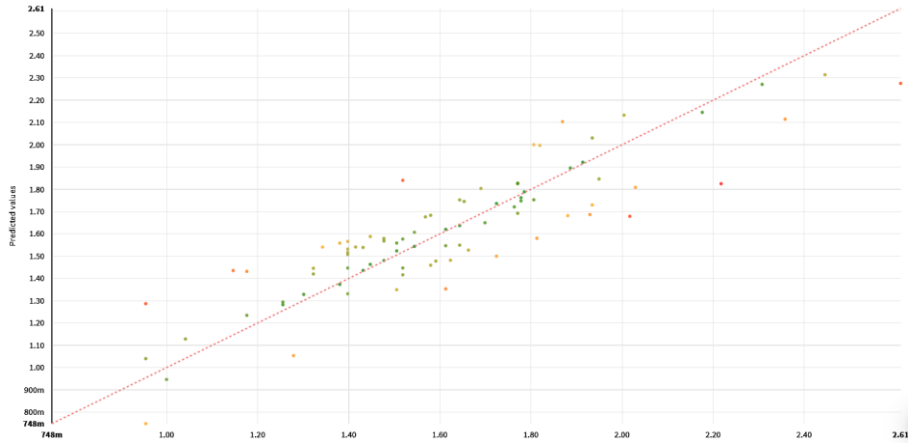


Figure 9: Scatter plot of the 20% test (validation) dataset

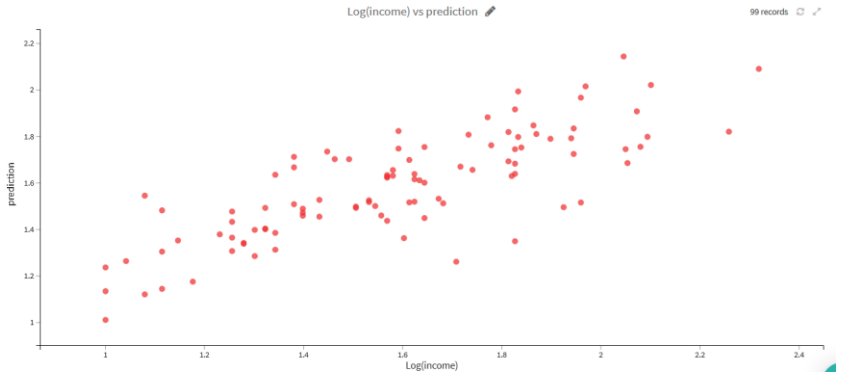


Figure 10:Histogram of errors of the 20% (validation) dataset

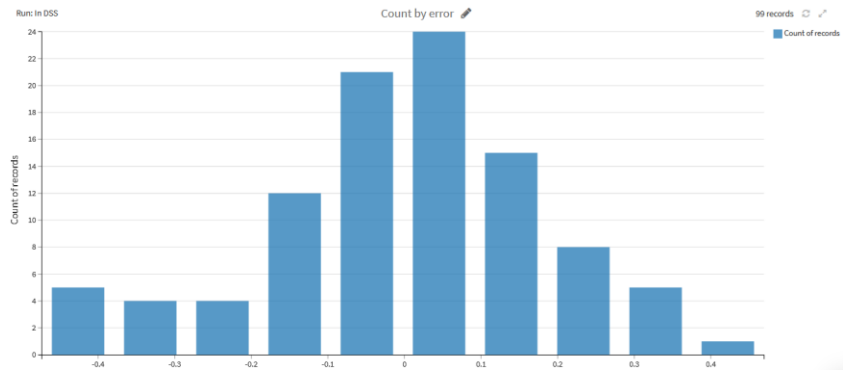
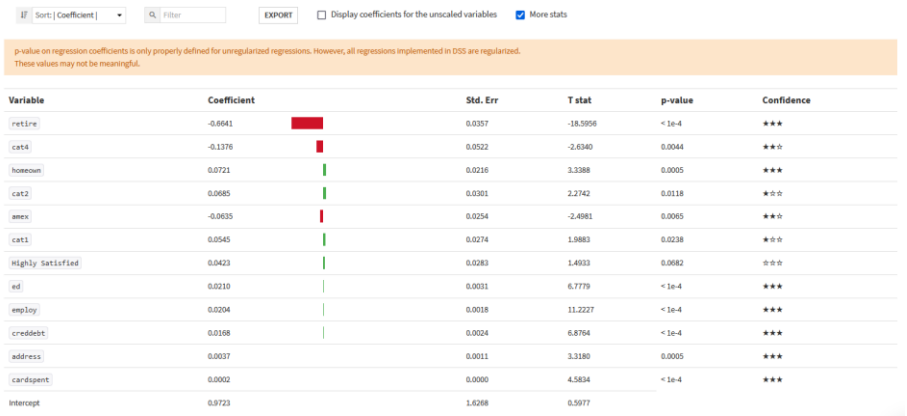


Figure 11: Non-standardized coefficients



Sort: | Coefficient | Filter EXPORT Display coefficients for the unscaled variables More stats

p-value on regression coefficients is only properly defined for unregularized regressions. However, all regressions implemented in DSS are regularized.
These values may not be meaningful.

Variable	Coefficient	Std. Err	T stat	p-value	Confidence
retire	-0.6641	0.0357	-18.5956	< 1e-4	★★★
cat4	-0.1376	0.0522	-2.6340	0.0044	★★☆
homeown	0.0721	0.0216	3.3388	0.0005	★★★
cat2	0.0685	0.0301	2.2742	0.0118	★☆☆
amex	-0.0635	0.0254	-2.4981	0.0065	★★☆
cat1	0.0545	0.0274	1.9883	0.0238	★☆☆
Highly Satisfied	0.0423	0.0283	1.4933	0.0682	☆☆☆
ed	0.0210	0.0031	6.7779	< 1e-4	★★★
employ	0.0204	0.0018	11.2227	< 1e-4	★★★
creddebt	0.0168	0.0024	6.8764	< 1e-4	★★★
address	0.0037	0.0011	3.3180	0.0005	★★★
cardspent	0.0002	0.0000	4.5834	< 1e-4	★★★
Intercept	0.9723	1.6268	0.5977		

Figure 12: OLS regression results

OLS Regression Results

Dep. Variable:	Log(income)	R-squared:	0.716
Model:	OLS	Adj. R-squared:	0.709
Method:	Least Squares	F-statistic:	102.3
Date:	Sun, 19 Dec 2021	Prob (F-statistic):	1.36e-124
Time:	11:17:48	Log-Likelihood:	170.96
No. Observations:	499	AIC:	-315.9
Df Residuals:	486	BIC:	-261.2
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.9323	0.044	21.089	0.000	0.845	1.019
ed	0.0227	0.002	9.099	0.000	0.018	0.028
cat4	-0.1293	0.040	-3.201	0.001	-0.209	-0.050
cat2	0.0674	0.024	2.857	0.004	0.021	0.114
cat1	0.0830	0.021	3.874	0.000	0.041	0.125
employ	0.0209	0.001	13.938	0.000	0.018	0.024
retire	-0.6428	0.029	-22.457	0.000	-0.699	-0.587
creddebt	0.0199	0.002	8.945	0.000	0.016	0.024
Highly Satisfied	0.0504	0.022	2.291	0.022	0.007	0.094
homeown	0.0702	0.017	4.107	0.000	0.037	0.104
address	0.0034	0.001	3.734	0.000	0.002	0.005
amex	-0.0566	0.020	-2.877	0.004	-0.095	-0.018
cardspent	0.0002	3.68e-05	5.196	0.000	0.000	0.000

Figure 13: Standardized coefficients

Sort: | Coefficient | Filter EXPORT Display coefficients for the unscaled variables More stats

p-value on regression coefficients is only properly defined for unregularized regressions. However, all regressions implemented in DSS are regularized.
These values may not be meaningful.

Variable	Coefficient	Std. Err	T stat	p-value	Confidence
retire	-0.2327	0.0125	-18.5956	< 1e-4	★★★
employ	0.1962	0.0175	11.2227	< 1e-4	★★★
creddebt	0.0765	0.0111	6.8764	< 1e-4	★★★
ed	0.0722	0.0107	6.7779	< 1e-4	★★★
cardspent	0.0519	0.0113	4.5834	< 1e-4	★★★
address	0.0448	0.0135	3.3180	0.0005	★★★
homeown	0.0343	0.0103	3.3388	0.0005	★★★
cat2	0.0310	0.0136	2.2742	0.0118	★☆☆
cat4	-0.0271	0.0103	-2.6340	0.0044	★★☆
amex	-0.0261	0.0104	-2.4981	0.0065	★★☆
cat1	0.0247	0.0124	1.9883	0.0238	★☆☆
Highly Satisfied	0.0154	0.0103	1.4933	0.0682	☆☆☆
Intercept	1.6258	0.0098	165.1480		