Sahir Doshi, Joonyoung (Jase) Jeon, Rutuja Kharate, Serena Theobald, Qinglin (Jason) Yang

Jetson Leder-Luis

BA222

December 6, 2022

Project 3: Regression Prediction

Our project creates a regression prediction model to examine how likely contacted customers are to subscribe to a term deposit account – a savings account into which customers put money for a fixed amount of time. To clean the data, we removed all "unknown" values, and we converted several categorical variables from a "yes/no" configuration to a "0/1" notation, where 0 = "no" and 1 = "yes". Furthermore, one of the variables we chose, *contact*, was converted into a dummy variable. The selection and reasoning behind each variable in our model is as follows:

- *contact* - People usually ignore "spam" calls, but still look at text messages, so we assumed different contact methods would result in different levels of clients reading promotions
- *pdays* - if a client was contacted more recently by a previous campaign, they are more aware of the term deposit account, and therefore likely to subscribe to one
- *emp.var.rate* - high employment > people have stable income > ability to save for the future
- *cons.conf.idx* - a low CCI equates to a generally negative sentiment for the economy's future, prompting individuals to save more and theoretically open new savings accounts
- *cons.price.idx* - a high CPI equates to inflation, meaning goods and services cost more, so individuals would save more (and often via savings accounts)

Fitted Equation:

y = *a* + (*b* * telephone) + (*c* * pdays) + (*d* * emp.var.rate)

+ (*e* * cons.conf.idx) + (*f* * cons.price.idx)

Our R^2 value was: **0.204 (rounded)**

Our Out-of-sample R^2 value was: **0.195 (rounded)**

Regression Output of Model on Training Data:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.204
Model:	OLS	Adj. R-squared:	0.203
Method:	Least Squares	F-statistic:	158.5
Date:	Sun, 04 Dec 2022	Prob (F-statistic):	3.08e-150
Time:	18:58:42	Log-Likelihood:	-555.03
No. Observations:	3090	AIC:	1122.
Df Residuals:	3084	BIC:	1158.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-17.4267	1.505	-11.576	0.000	-20.378	-14.475
cons.price.idx	0.1955	0.016	12.072	0.000	0.164	0.227
cons.conf.idx	0.0103	0.001	8.413	0.000	0.008	0.013
emp.var.rate	-0.0947	0.006	-17.123	0.000	-0.106	-0.084
pdays	-0.0003	2.79e-05	-10.908	0.000	-0.000	-0.000
telephone	-0.1261	0.014	-8.857	0.000	-0.154	-0.098

Omnibus:	1179.864	Durbin-Watson:	1.906
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4107.768
Skew:	1.937	Prob(JB):	0.00
Kurtosis:	7.112	Cond. No.	2.83e+05

Contributions:

Sahir did the primary coding, cleansing the data, deducing the final regression model, and formatting the Jupyter Notebook. Serena and Rutuja also assisted in programming by running various other models that were not chosen. Jase and Jason typed the majority of the report. All 5 team members participated in productive conversations about how to clean the data, which variables to use (both in unused models and the used model), and interpreting the chosen variables in the context of the task.