Direct Marketing

"The data is related to direct marketing campaigns of a Portuguese banking institution. Predict if client will subscribe for term deposit."

Data Summary

- Train Data Set Volume - 41188
- Test Data Set Volume - 4119
- Number of Input feature - 20
- categorical_vars- 10 - (job marital ,education ,default ,housing ,loan ,contact ,month ,day_of_week , poutcome)
- continuous_vars- 10 - (age duration, campaign, pdays ,previous , emp.var.rate ,cons.price.idx ,cons.conf.idx euribor3m ,nr.employed)

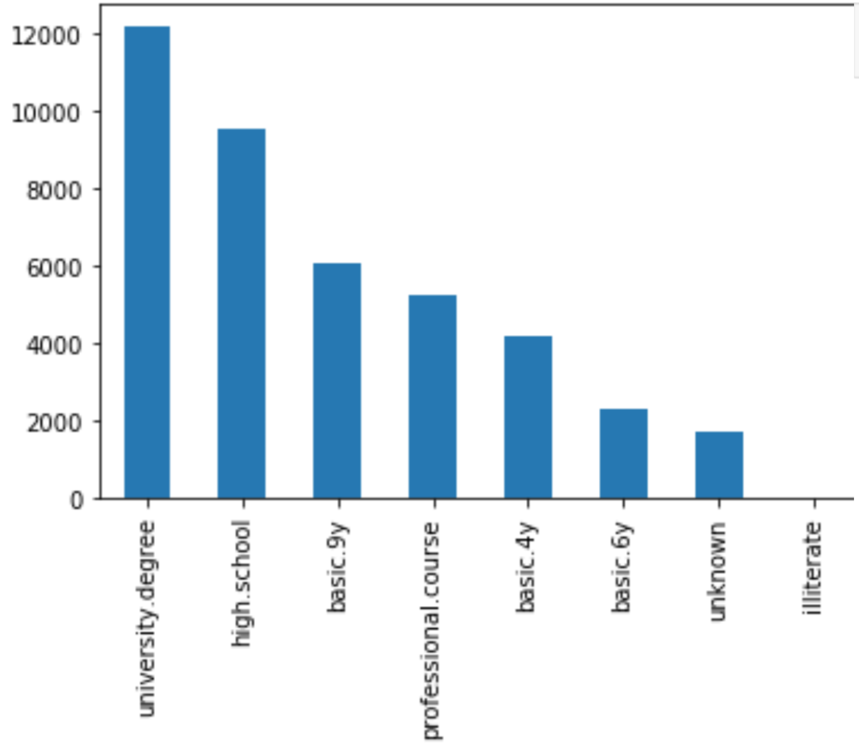
pdays

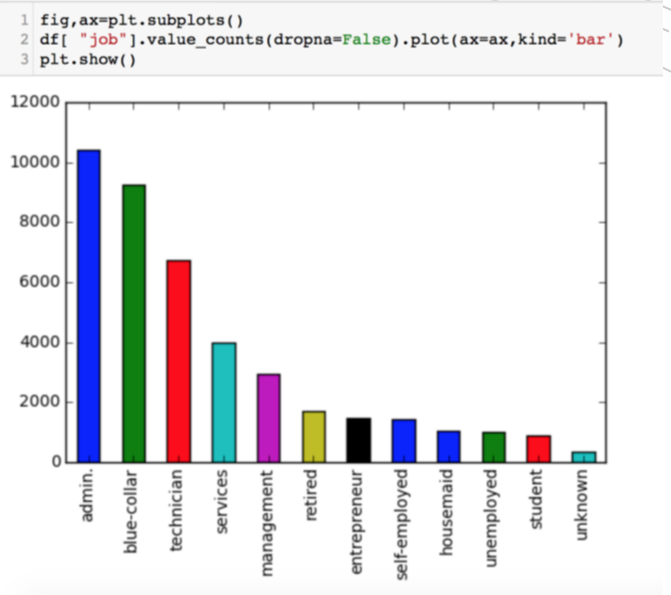
- df.pdays[(df['pdays'] == 999) & (df['previous'] != 0)].shape
  - 4110 : which is 10% of the overall data

- Introduction of a possible error
  - Poutcome has a conclusive result & previous indicates contact with the customer
  - Pdays however shows that the person was not contacted

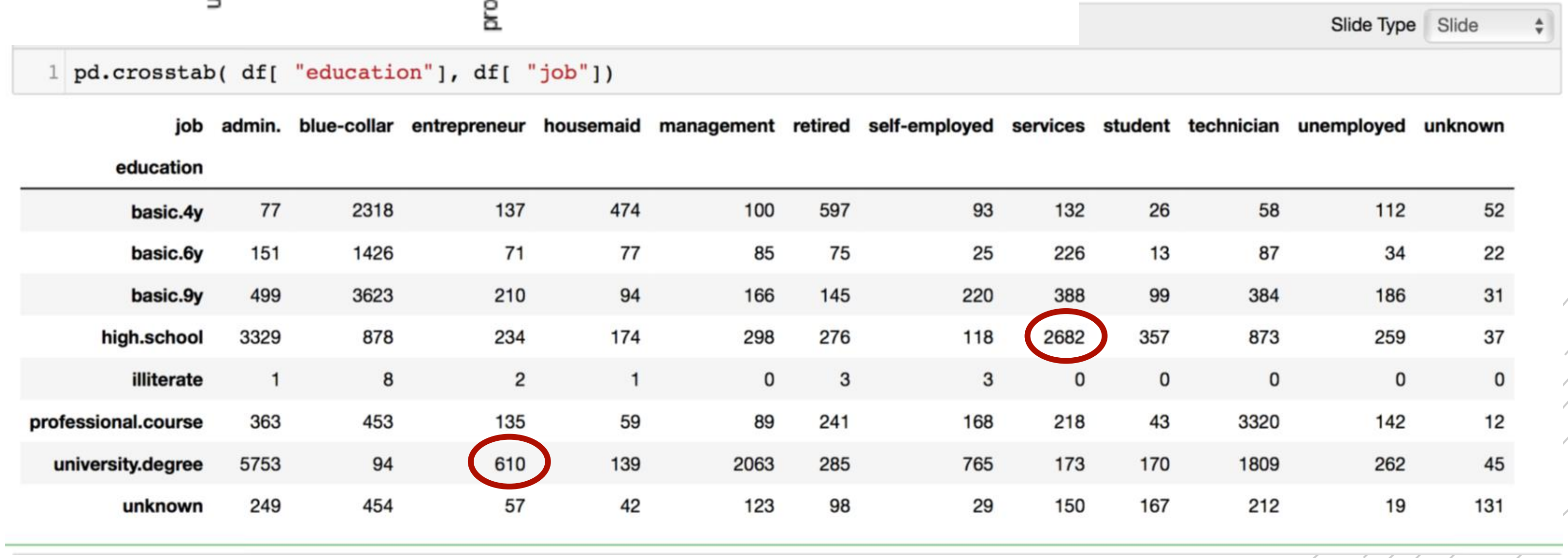
- Highly co-related to previous & poutcome, hence column removed

Imputation methods





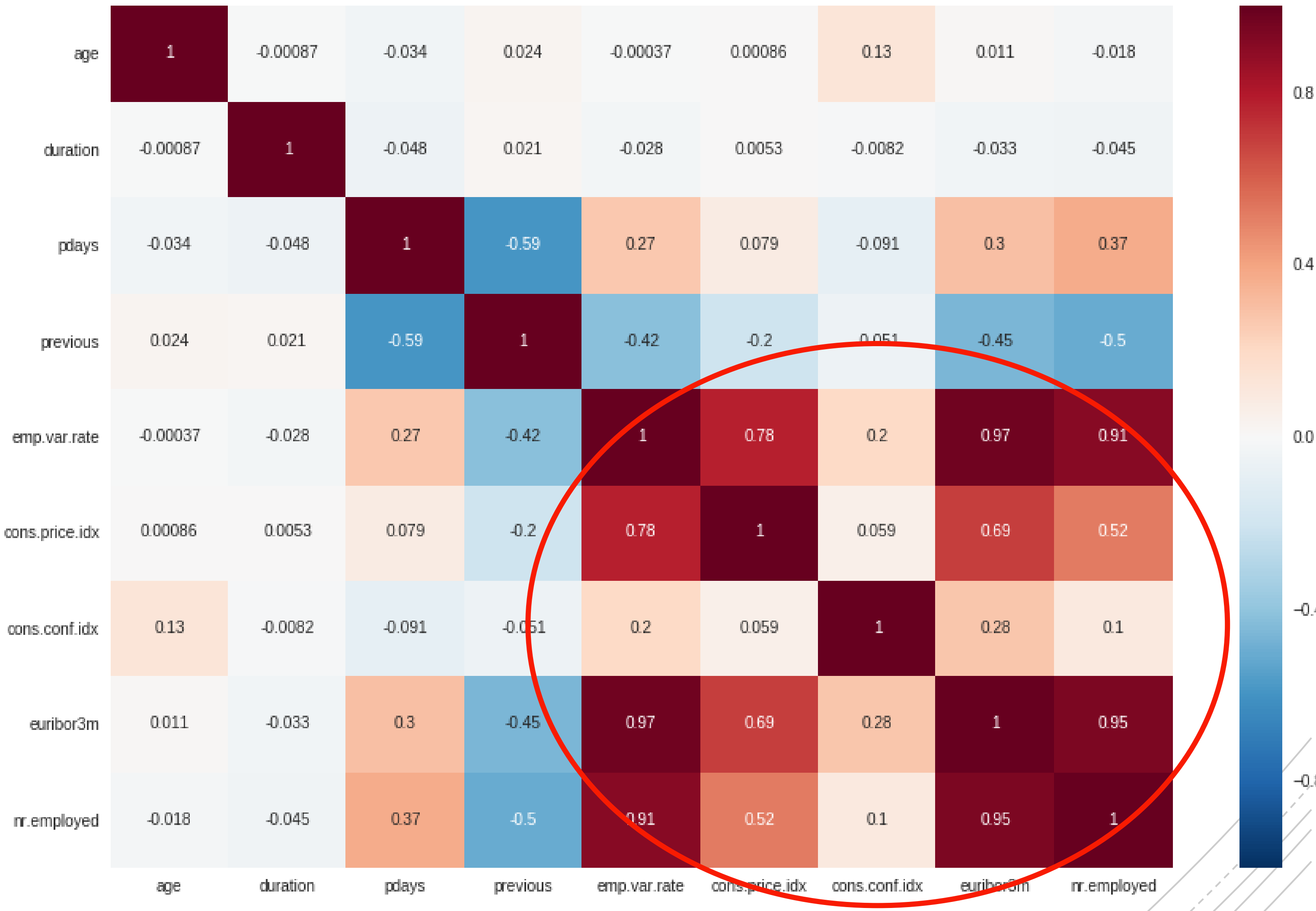
```
fig,ax=plt.subplots()
df[ "job"].value_counts(dropna=False).plot(ax=ax,kind='bar')
plt.show()
```



```
pd.crosstab( df[ "education"], df[ "job"])
```

job education	admin.	blue-collar	entrepreneur	housemaid	management	retired	self-employed	services	student	technician	unemployed	unknown
basic.4y	77	2318	137	474	100	597	93	132	26	58	112	52
basic.6y	151	1426	71	77	85	75	25	226	13	87	34	22
basic.9y	499	3623	210	94	166	145	220	388	99	384	186	31
high.school	3329	878	234	174	298	276	118	2682	357	873	259	37
illiterate	1	8	2	1	0	3	3	0	0	0	0	0
professional.course	363	453	135	59	89	241	168	218	43	3320	142	12
university.degree	5753	94	610	139	2063	285	765	173	170	1809	262	45
unknown	249	454	57	42	123	98	29	150	167	212	19	131

Indices



	age	duration	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1	-0.00087	-0.034	0.024	-0.00037	0.00086	0.13	0.011	-0.018
duration	-0.00087	1	-0.048	0.021	-0.028	0.0053	-0.0082	-0.033	-0.045
pdays	-0.034	-0.048	1	-0.59	0.27	0.079	-0.091	0.3	0.37
previous	0.024	0.021	-0.59	1	-0.42	-0.2	-0.051	-0.45	-0.5
emp.var.rate	-0.00037	-0.028	0.27	-0.42	1	0.78	0.2	0.97	0.91
cons.price.idx	0.00086	0.0053	0.079	-0.2	0.78	1	0.059	0.69	0.52
cons.conf.idx	0.13	-0.0082	-0.091	-0.051	0.2	0.059	1	0.28	0.1
euribor3m	0.011	-0.033	0.3	-0.45	0.97	0.69	0.28	1	0.95
nr.employed	-0.018	-0.045	0.37	-0.5	0.91	0.52	0.1	0.95	1

EDA - Highlights

- Classifier problem
- Hot encode the features that have few variables[yes/no] and where we need to remove unknowns
- Scale numeric data
- Remove highly co-related feature [eg. duration]

Walk through

The model selection & parameter tuning

Key Results

- Trade-off between cost & opportunity
  - High cost of more contacts
  - Opportunity of a potential being classified as non-interested
- The market performance indices has a major influence on subscription rather than user demographics
- Age group, day of the week and month of the year also influence the campaign

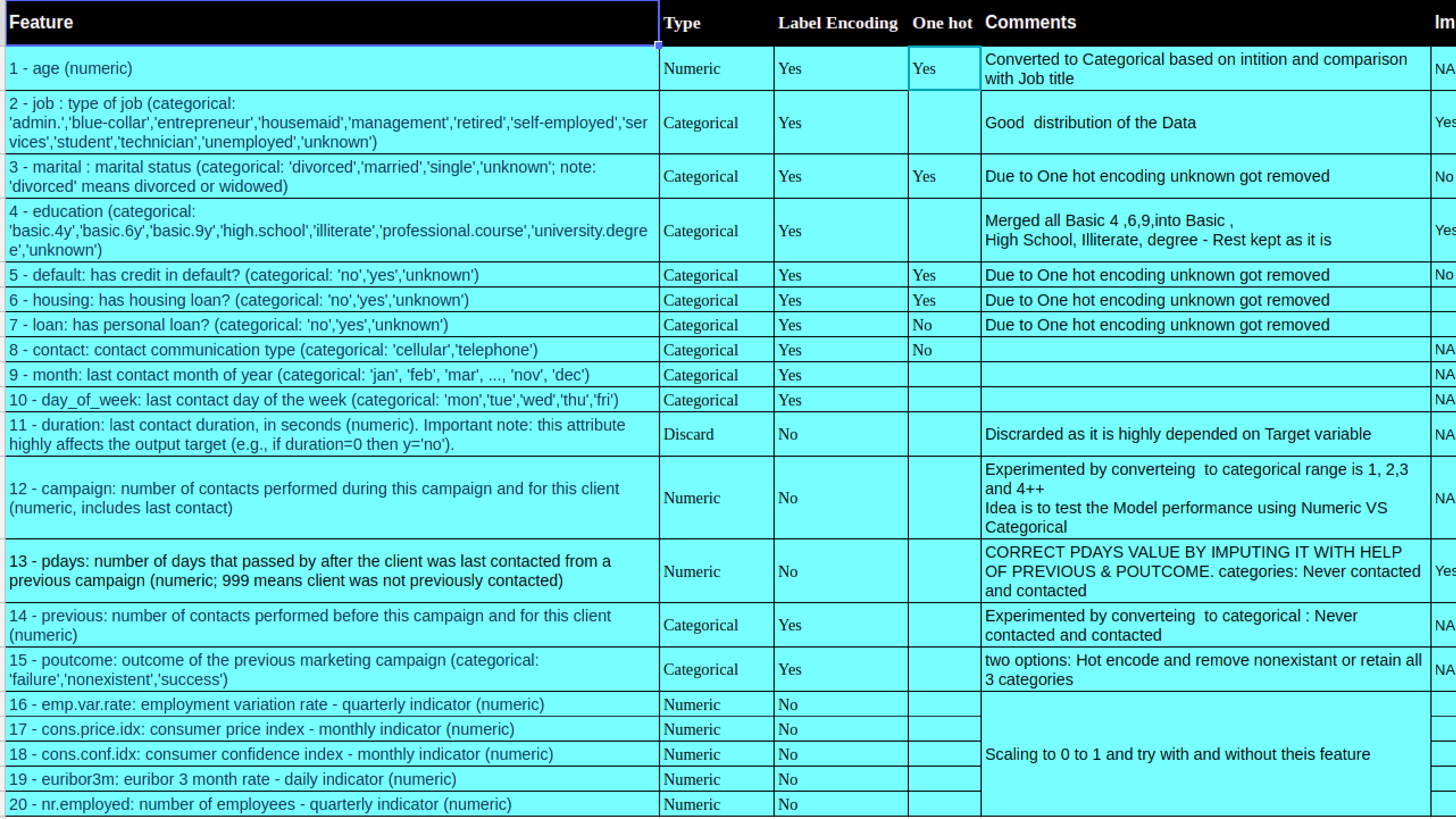
Future Scope

- Study the Type 2 error
  - Observe which feature is most co-related to the predicted values
- Ensemble with other models to improve score and reduce the increase in cost vs lost opportunity
- Combine the highly co-related numeric features

References

- Slide Type-SlideSub-SlideFragmentSkipNotes
- Writeup and sites refered too-
- For Eurobor def-http://www.mymoney.lu/3-questions-to-help-you-understand-euribor/?lang=en
- For Merging Highly co-related features-https://www.quora.com/Given-several-highly-correlated-variables-how-can-I-pick-what-is-the-best-predictor-for-the-others
- https://stats.stackexchange.com/questions/116853/combining-merging-correlated-variables
- Understanding Importance of hot- Encoding and label Encoder - https://datascience.stackexchange.com/questions/9443/when-to-use-one-hot-encoding-vs-labelencoder-vs-dictvectorizor
- Steps to follow for Model Building hyper parameter tuning for classfication Model -http://blog.kaggle.com/2016/07/21/approaching-almost-any-machine-learning-problem-abhishek-thakur/

Appendix



Feature	Type	Label Encoding	One hot	Comments	Im
1 - age (numeric)	Numeric	Yes	Yes	Converted to Categorical based on intition and comparison with Job title	NA
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')	Categorical	Yes		Good distribution of the Data	Yes
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)	Categorical	Yes	Yes	Due to One hot encoding unknown got removed	No
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')	Categorical	Yes		Merged all Basic 4 ,6,9,into Basic , High School, Illiterate, degree - Rest kept as it is	Yes
5 - default: has credit in default? (categorical: 'no','yes','unknown')	Categorical	Yes	Yes	Due to One hot encoding unknown got removed	No
6 - housing: has housing loan? (categorical: 'no','yes','unknown')	Categorical	Yes	Yes	Due to One hot encoding unknown got removed	
7 - loan: has personal loan? (categorical: 'no','yes','unknown')	Categorical	Yes	No	Due to One hot encoding unknown got removed	
8 - contact: contact communication type (categorical: 'cellular','telephone')	Categorical	Yes	No		NA
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')	Categorical	Yes			NA
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')	Categorical	Yes			NA
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no').	Discard	No		Discrarded as it is highly depended on Target variable	NA
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)	Numeric	No		Experimented by converteing to categorical range is 1, 2,3 and 4++ Idea is to test the Model performance using Numeric VS Categorical	NA
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)	Numeric	No		CORRECT PDAYS VALUE BY IMPUTING IT WITH HELP OF PREVIOUS & POUTCOME. categories: Never contacted and contacted	Yes
14 - previous: number of contacts performed before this campaign and for this client (numeric)	Categorical	Yes		Experimented by converteing to categorical : Never contacted and contacted	NA
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')	Categorical	Yes		two options: Hot encode and remove nonexistant or retain all 3 categories	NA
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)	Numeric	No		Scaling to 0 to 1 and try with and without theis feature	
17 - cons.price.idx: consumer price index - monthly indicator (numeric)	Numeric	No			
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)	Numeric	No			
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)	Numeric	No			
20 - nr.employed: number of employees - quarterly indicator (numeric)	Numeric	No			