Supervised learning capstone
Bank telemarketing analysis

BY: *NUSAIR IMAM*

MENTOR: *ILYAS USTUN*

PROGRAM MANAGER: *JOSEPHINE PIKE*

Presentation Outline

1. Understanding the data
2. Data analysis, cleaning and preparation
3. Initial application of selected models
4. Feature engineering and parameter tuning
5. Conclusion and Future Considerations

The data - variables

Categorical (dtype = object)	Continuous (dtype = numerical)
Job	Age
Marital	Duration
Education	Campaign
Default	Pdays
Housing	Previous
Loan	Emp.var.rate
Contact	cons.price.idx
Month	Cons.conf.idx
Day_of_week	Euribor3m
Poutcome	Nr.employed

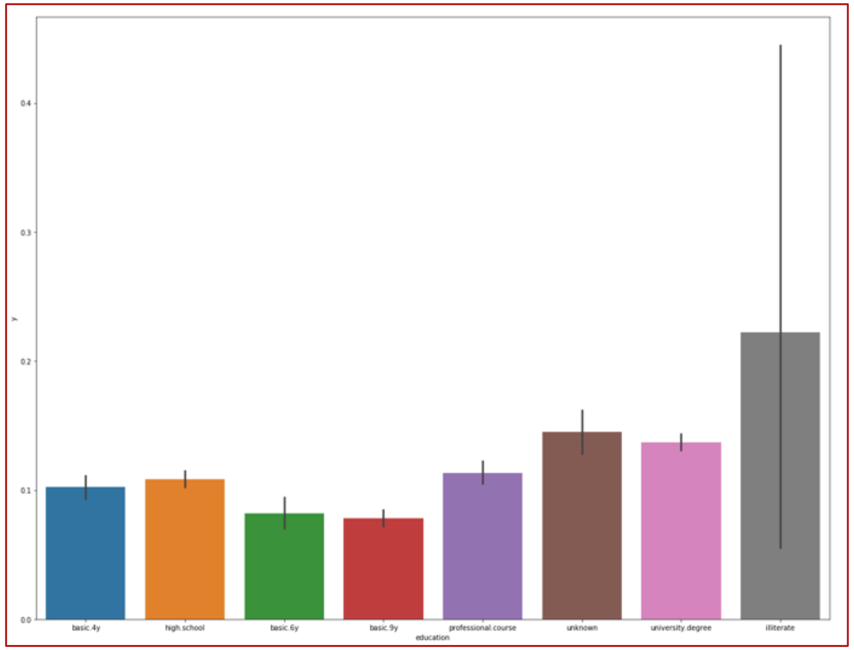
Object Variables

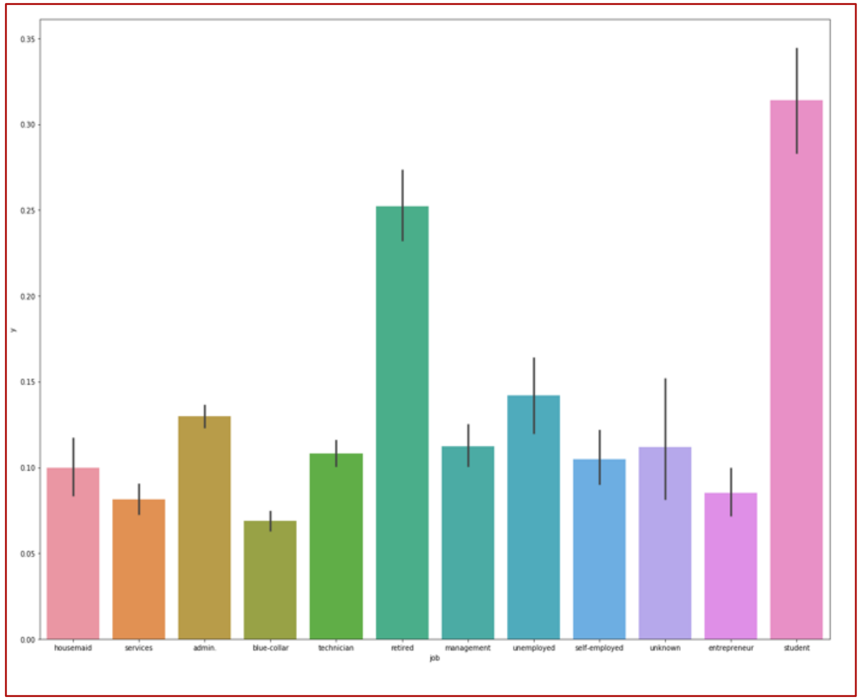
- Job : Occupational information
- Marital : Marital status
- Education : Education level
- Housing : Housing loan (Yes/No)
- Default : Credit default (Yes/No)
- Loan : Personal loan (Yes/No)
- Contact : Method of contact (telephone/cellular)
- Month : Month of the year (last contact)
- Day_of_week: Day (last contact)
- Poutcome: Outcome of previous campaign

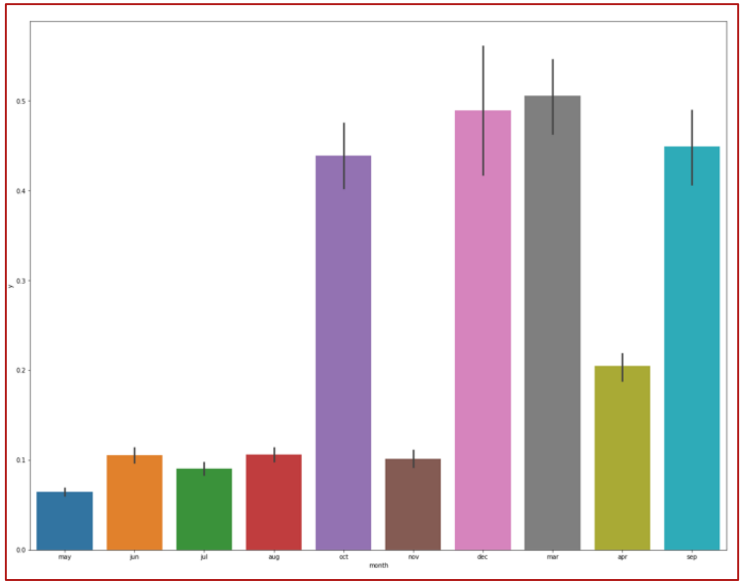
Object Variables

Key Takeaways

- Students and retired people have a greater chance of making a deposit
- People who identified as illiterate have a higher chance of making a deposit ; large error bar indicates presence of outliers
- Certain months have a higher success rate (March, April, Sept, Oct, Dec)
- Class imbalance ; only 11% of participants made a deposit

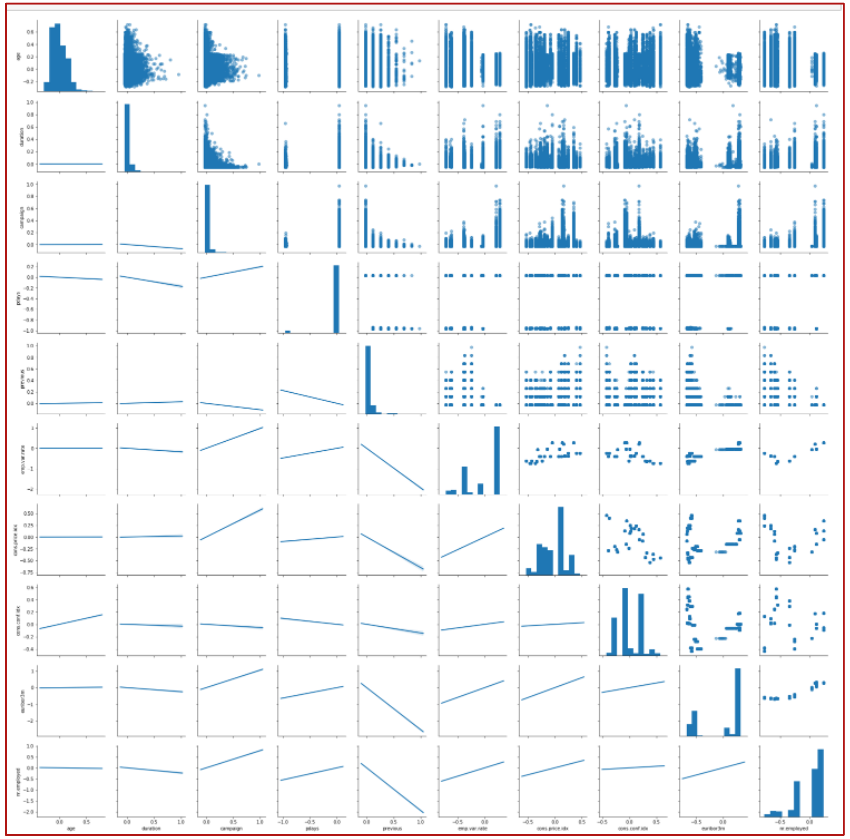






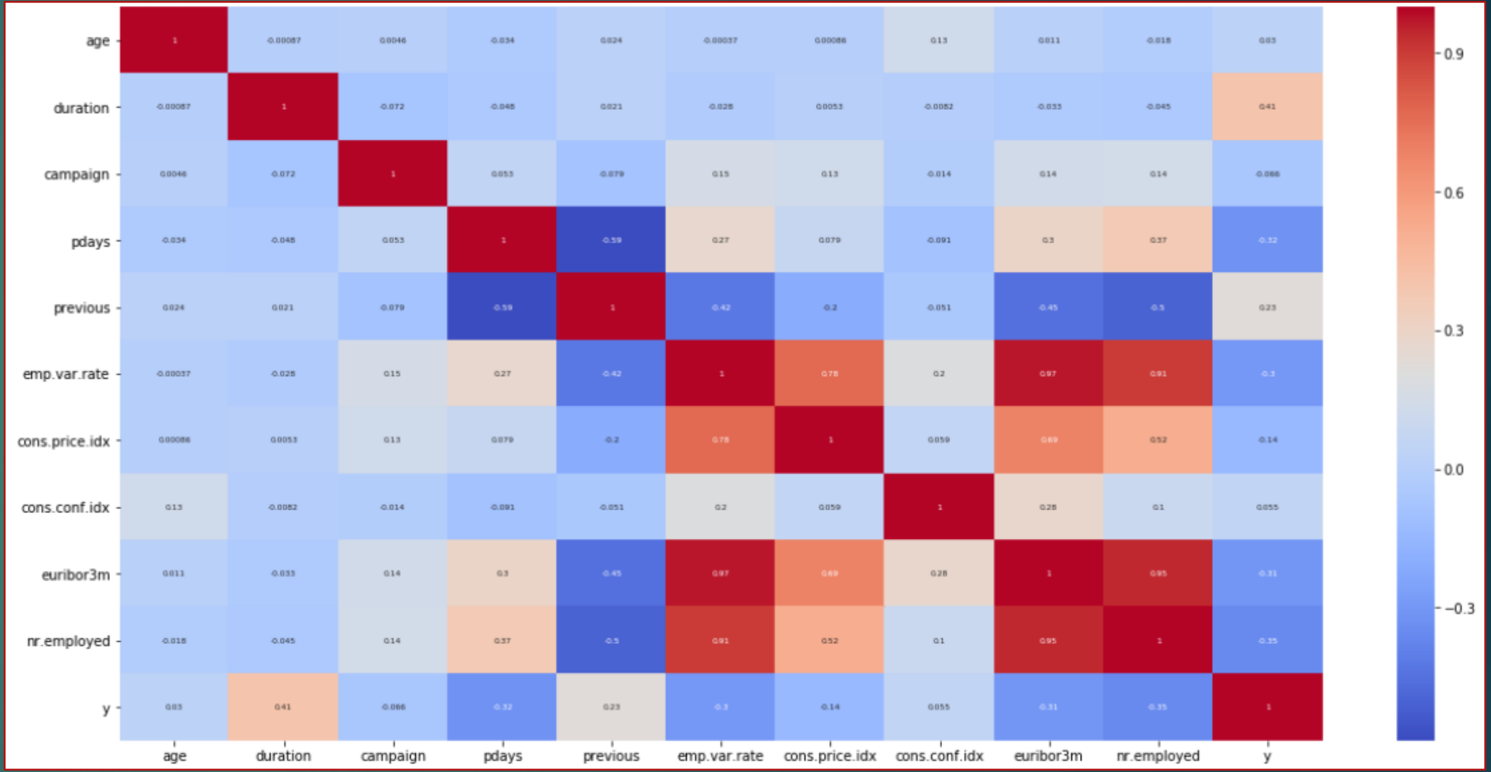
Numerical variables

- Age : Age of client
- Duration : Contact duration (seconds)
- Campaign : Number of calls made to client
- Pdays : Number of days since last contact (previous campaign)
- Previous : Number of calls in previous campaign
- Emp.var.rate : Employment variation rate (quarterly)
- cons.price.idx : Consumer price index (monthly)
- Cons.conf.idx : Consumer condfidence index (monthly)
- Euribor3m : Euro interbank interest rate (daily)
- Nr.employed : Number of employees (quarterly)



Numerical variables - heatmap

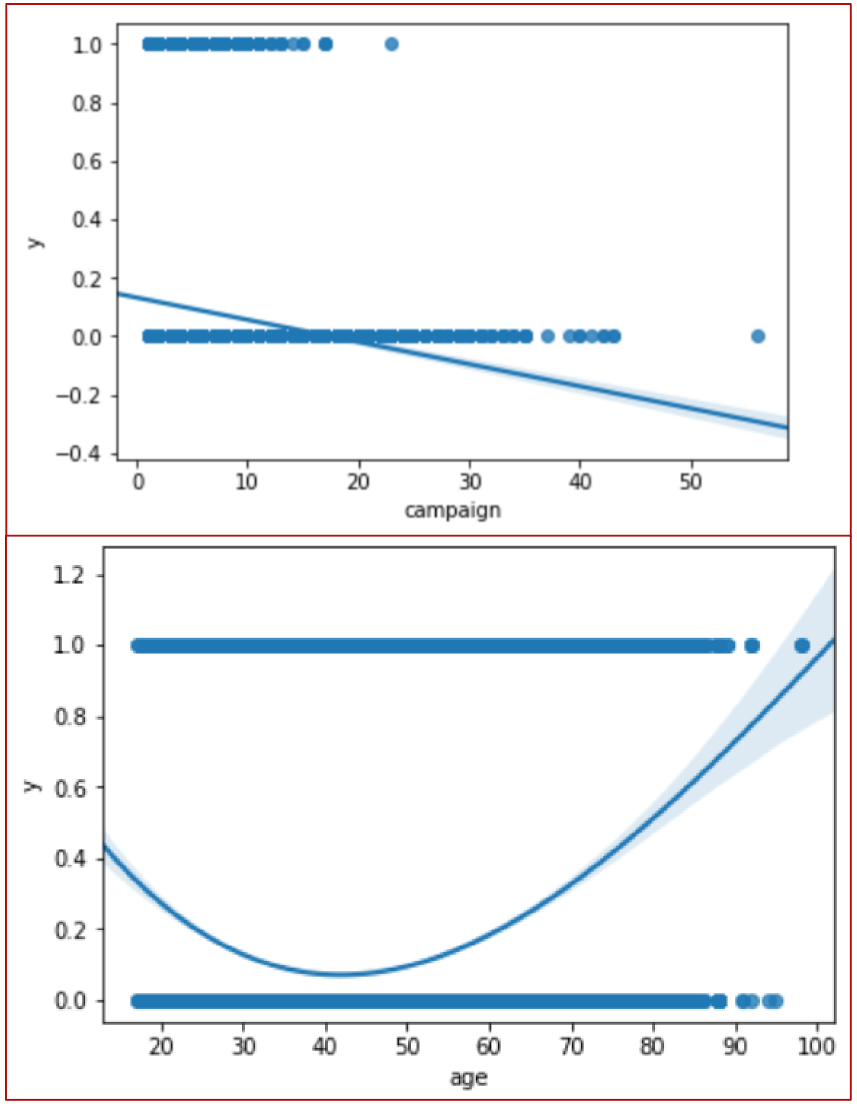
- High correlations between the output and the following variables:
  - Nr.employed
  - Euribor3m
  - Duration
  - Emp.var.rate
  - Pdays
  - Previous (outcome)



Numerical variables

Key Takeaways

- Rememeber – the variable ‘campaign’ is the number of calls rendered per client during the campaign
  - No person who was contacted more than 20 times subscribed to a fixed term deposit
- People at the ends of the age spectrum are more likely to subscribe
- High Correlations between fixed term deposits and



Data Cleaning

- No NaN values, missing data
- Converted output from Yes/No to numerical 1/0
- Creating dummy variables for object variables in order to have a numerical dataset with all the variables.
- Dealt with class imbalance (data skewed towards the negative outcome)
  - Oversampling via Sklearn's SMOTE

```
#checking for null values
df.isnull().sum()
```

```
age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
contact           0
month             0
day_of_week       0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
emp.var.rate      0
cons.price.idx    0
cons.conf.idx     0
euribor3m         0
nr.employed       0
y                 0
dtype: int64
```

Test, train methodology

- Data split on an 80:20 basis using Sklearn’s train_test_split function
- Models trained on train dataset and tested on both train and dataset
- *Initial modelling on Oversampled train dataset with 63 features*
- *Final modelling on reduced set of features based on feature importance from RFC, LogR and feature engineering*

Terms and definitions – *in context of what it means in this application*

- Type 1 error: FP
  - *Incorrectly identifying non-subscibers as subscribers*
- Type 2 error: FN
  - *Not identifying those that are predicted to subscribe*
- **Precision: TP/(TP+FP)**
  - **Out of those that we think will subscribe, what percentage actually did?**
- **Recall: TP/(TP+FN)**
  - **Out of those that did subscribe, what percentage did we predict would?**

Terms and definitions – *in context of what it means in this application*

- Precision: TP/(TP+FP)
  - Out of those that we think will subscribe, what percentage actually did?
    - Example Scenario : sending coupons to those that think are likely to subscribe – you do not want to waste coupons on those that will not subscribe!
- Recall: TP/(TP+FN)
  - Out of those that did subscribe, what percentage did we predict would?
    - Bank telemarketing – this is the primary KPI for this application – we do not want to misclassify people who end up subscribing

Models

- Random Forest Classifier
- K-nearest Classifier
- Logistic Regression
- SVM
  - SVC
  - Linear SVC

	RFC	LogR	KNN	SVC	lSVC
Accuracy (%)	99.3	91.0	93.1	95.9	88.4
Type 1 (%)	0.04	3.26	2.46	-	0.16
Type 2 (%)	0.63	6.65	4.41	-	10.99
Precision (%)	99.7	66.46	73.7	-	67.08
Recall (%)	94.45	41.23	61.0	-	2.90

Initial Results : Trained and tested on entire dataset w/o oversampling

Models – Initial training on unbalanced data

- Results biased towards dominant class (client does NOT make deposit)
  - SVC left out due to computational constraints (# feat = 63)
- Tendency to predict False negative (Type 2)
  - This manifests itself in the Recall

	RFC	LogR	KNN	SVC	ISVC
Accuracy (%)	90.9	91.2	90.5	88.9	89.1
Type 1 (%)	2.39	2.16	3.81	-	0.1
Type 2 (%)	6.69	6.62	5.66	-	10.8
Precision (%)	64.6	67.2	58.6	-	74.2
Recall (%)	39.5	40.1	48.8	-	2.53

**Initial Results : Test subset (after test/train split)*

Oversampled data

- SMOTE (Synthetic Minority Oversampling Technique)
  - Potentially important information may have been simulated (risk of overfitting)

```
#dealing with class imbalance
from imblearn.over_sampling import SMOTE
smote =  SMOTE(random_state = 1, ratio = 1.0)
xx = X
yy = Y

x_balanced,y_balanced = smote.fit_sample(xx,yy)

print(x_balanced.shape)
print(y_balanced.shape)

x_var = list(xx.columns)
bal_df = pd.DataFrame(data = x_balanced , columns = x_var)
bal_df.describe()
bal_df.describe()
```

Models – Initial training on *oversampled* data (using SMOTE)

- Linear SVC has an exceptional recall but much lower accuracy than the others
- Reduction in type 2 errors and increase in Recall

	RFC	LogR	KNN	SVC	lSVC
Accuracy (%)	87.7	86.8	84.7	-	63.0
Type 1 (%)	8.7	11.8	12.81	-	36.87
Type 2 (%)	3.56	1.40	2.50	-	0.12
Precision (%)	46.3	45.0	40.0	-	22.86
Recall (%)	67.8	87.4	77.4	-	98.9

**Initial Results : Test subset (after test/train split)*

Parameter Tuning - GridSearchCV

- Implemented on RFC, Logistic Regression
- Other model's left out due to computational constraints

```
knn = KNeighborsClassifier()

params = {'n_neighbors':[1],
          'leaf_size':[1,2,3,5],
          'weights':['uniform', 'distance'],
          'algorithm':['auto', 'ball_tree','kd_tree','brute'],
          'n_jobs':[-1]}
```

```
rfc = ensemble.RandomForestClassifier()

params = {"n_estimators": [10, 5, 15],
              "max_depth": [3, 4, 5],
              "min_samples_split": [2, 3],
              "min_samples_leaf": [5, 10, 20],
              "max_leaf_nodes": [20, 40],
              "min_weight_fraction_leaf": [0.0]}
```

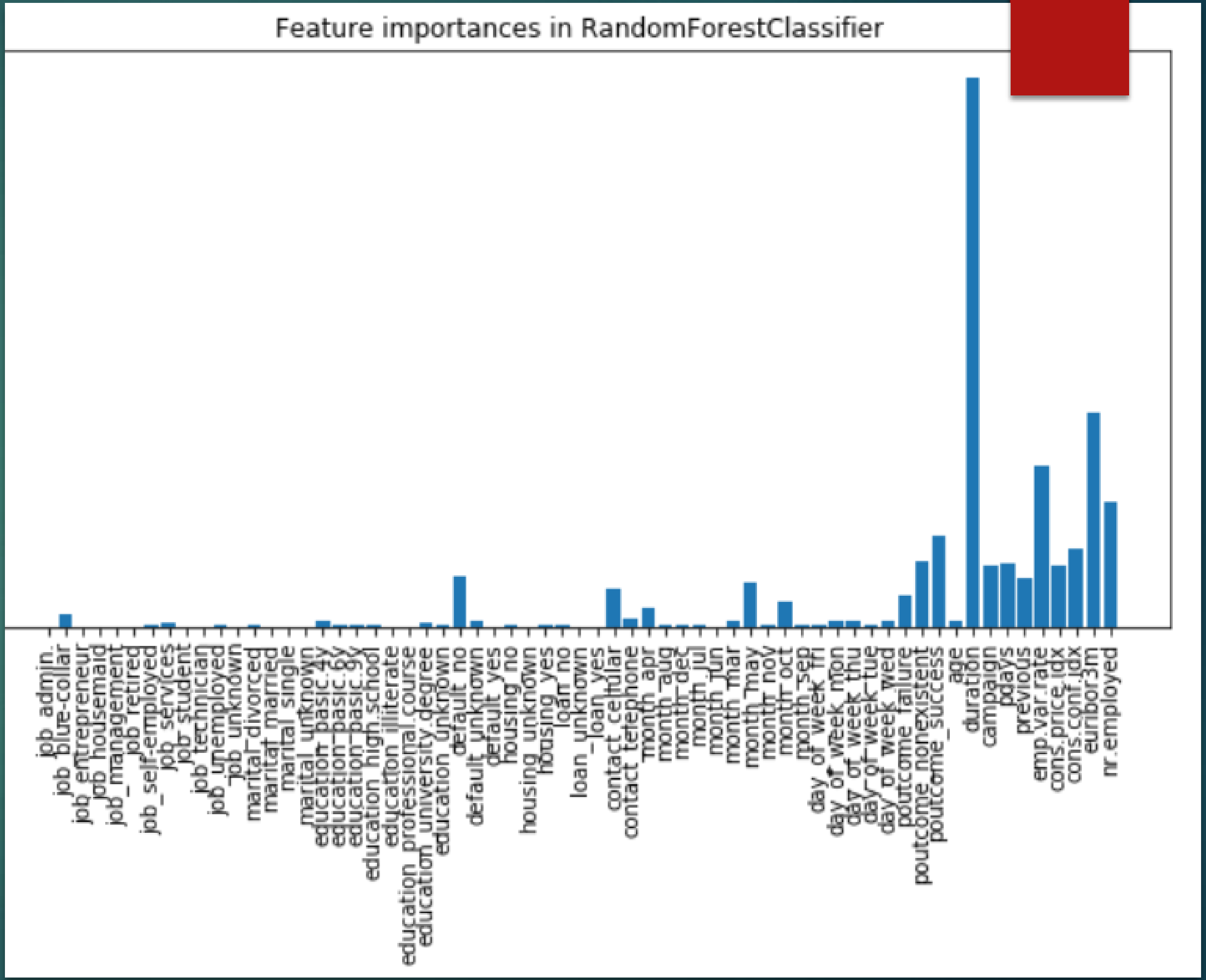
```
from sklearn import metrics
from sklearn.model_selection import GridSearchCV
from sklearn import svm
Cs = [0.001, 0.01, 0.1, 1, 10]
penalty = ['l1','l2']
param_grid = {'C': Cs, 'penalty' : penalty, 'dual' : [False]}
grid_search = GridSearchCV(svm.LinearSVC(), param_grid,  return_train_score=True)
grid_search.fit(X, Y)
```

```
{'C': [0.0001, 0.001, 0.01, 0.03, 0.1, 0.3, 0.6, 1, 1.3, 1.6, 2, 5, 10, 15, 20, 50, 100]}
```

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.model_selection import GridSearchCV
lr = LogisticRegression()

grid = GridSearchCV(estimator=lr, param_grid=param_grid, scoring='accuracy', verbose=3, n_jobs=-1, return_train_score=True)
```

Feature engineering

- Features shortlisted based on RFC importance diagram
- Duration feature removed since this is not known until after the call (when the outcome is known)
- Conversion rate feature added (# of calls divided by outcome)



Feature engineering)

- The reduced dataset exhibited overfitting
- Only after removal of the ‘duration’ column did results become more realistic
- 22 features other than ‘duration’ column

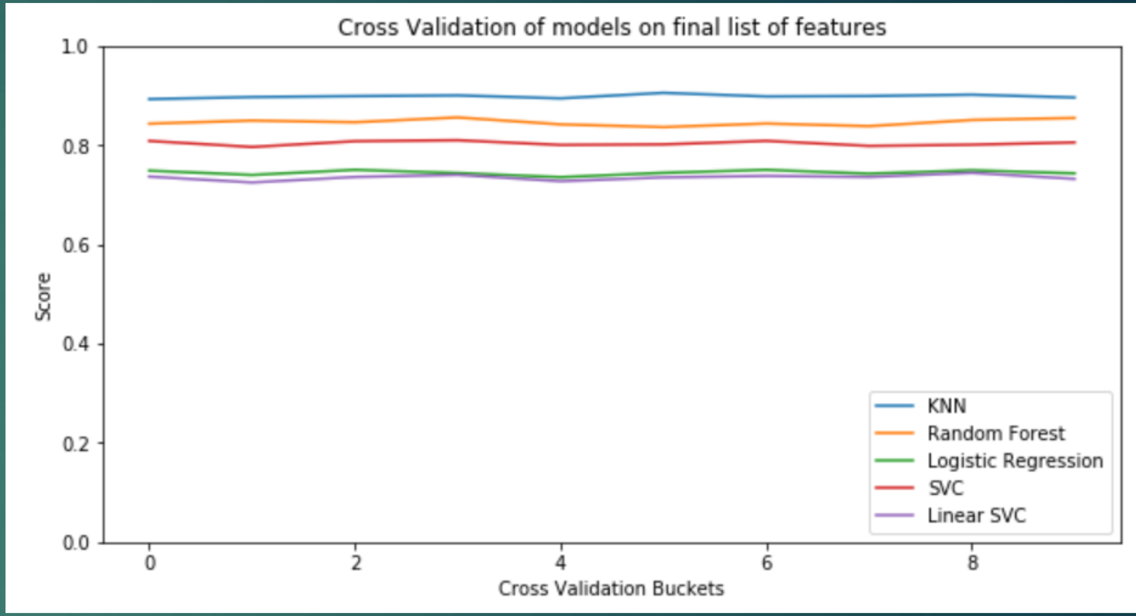
```
X = X.loc[:,X.columns != 'duration']
X_test = X_test.loc[:,X_test.columns != 'duration']

print(X.columns)
```

```
Index(['job_retired', 'job_self-employed', 'job_student', 'marital_unknown',
       'default_unknown', 'contact_telephone', 'month_aug', 'month_jun',
       'month_mar', 'month_may', 'month_nov', 'poutcome_failure',
       'poutcome_success', 'campaign', 'pdays', 'previous', 'emp.var.rate',
       'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed',
       'conversion_rate'],
      dtype='object')
```

	RFC	LogR
Accuracy (%)	98.6	99.9
Type 1 (%)	1.36	0.02
Type 2 (%)	0.05	0.02
Precision (%)	89.0	99.78
Recall (%)	99.6	99.78

Results – without 'duration' column

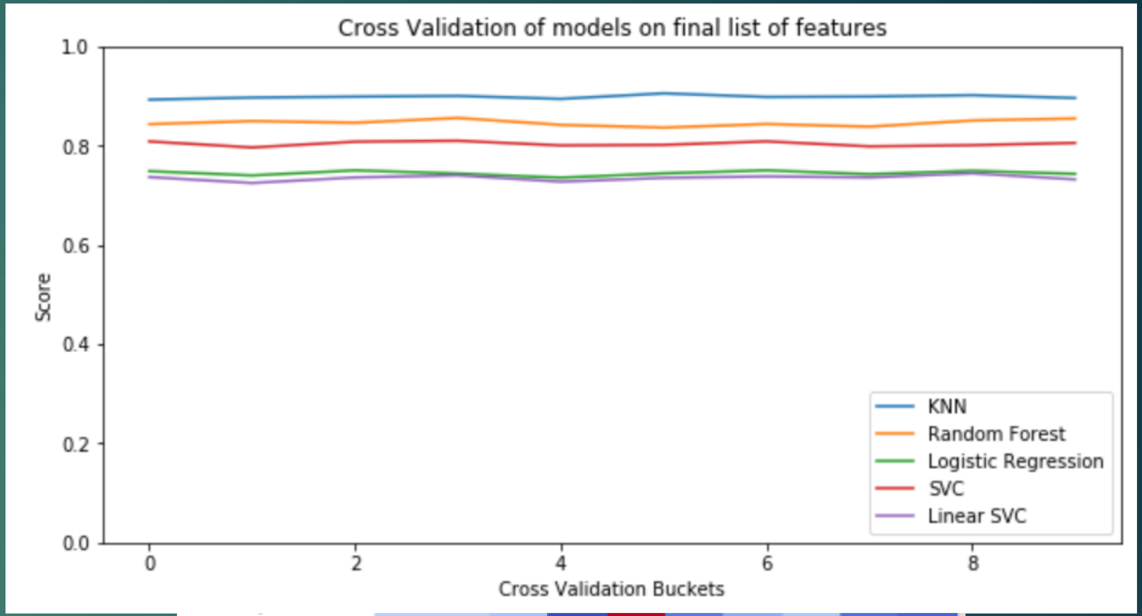
- Consistent cross validation scores
- RFC is the best model despite KNN having better overall results
  - SMOTE method utilizes the same methodology (Euclidean distance) as KNN and therefore KNN is more prone to overfitting on SMOTE-oversampled dataset

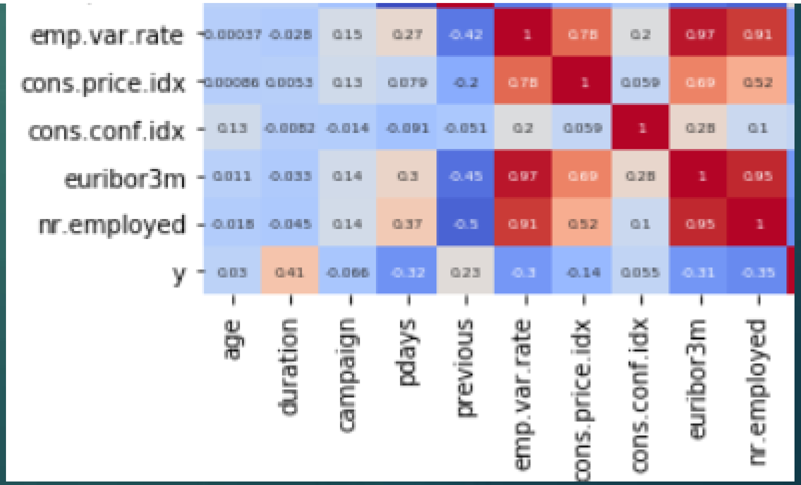


	RFC	LogR	KNN	lSVC
Accuracy (%)	99.2	99.9	97.0	57.0
Type 1 (%)	0.63	0.01	2.25	42.7
Type 2 (%)	0.16	0.07	0.76	0.23
Precision (%)	94.5	99.9	82.1	20.2
Recall (%)	98.6	99.3	93.2	97.9

Final Recommendation - Observations

- Despite good CV scores and model performance, it is unlikely that these model will perform well in real world situations due to high correlations between European socio-economic features.
- Logistic Regression performed the best overall
  - No client should be called more than 20 times
  - Depending on industry standard hitrate, the number of calls should be optimized by putting a maximum limit.





Future considerations

- Data treatment can be altered
  - PCA on highly correlated columns to reduce multicollinearity
  - Undersampling via numpy's random choice function
- Unsupervised techniques should be explored (Nueral networks)
- More features should be added to model
  - Gross income
  - Registered residence location
  - Size of family

Next steps

- Improve on current model performance
  - PCA of features
  - More iterations of tune, feature engineer, test process
  - Undersampling using functions other than np.random.choice
  - Try other ensemble models such a Xgboost, gradient boost
  - Map a neural network