Long Term Deposit Prediction of Bank Using Data Mining Techniques

Team members:

- Joshna Devi Vadapalli
- Jitendhar Reddy Adulla
- Shrivatson Ramaratnam Giridharan
- Syama Ravi Teja Jerrypothula
- Vamshi Krushna Lakavath

The Bank's Offer Campaign Prediction

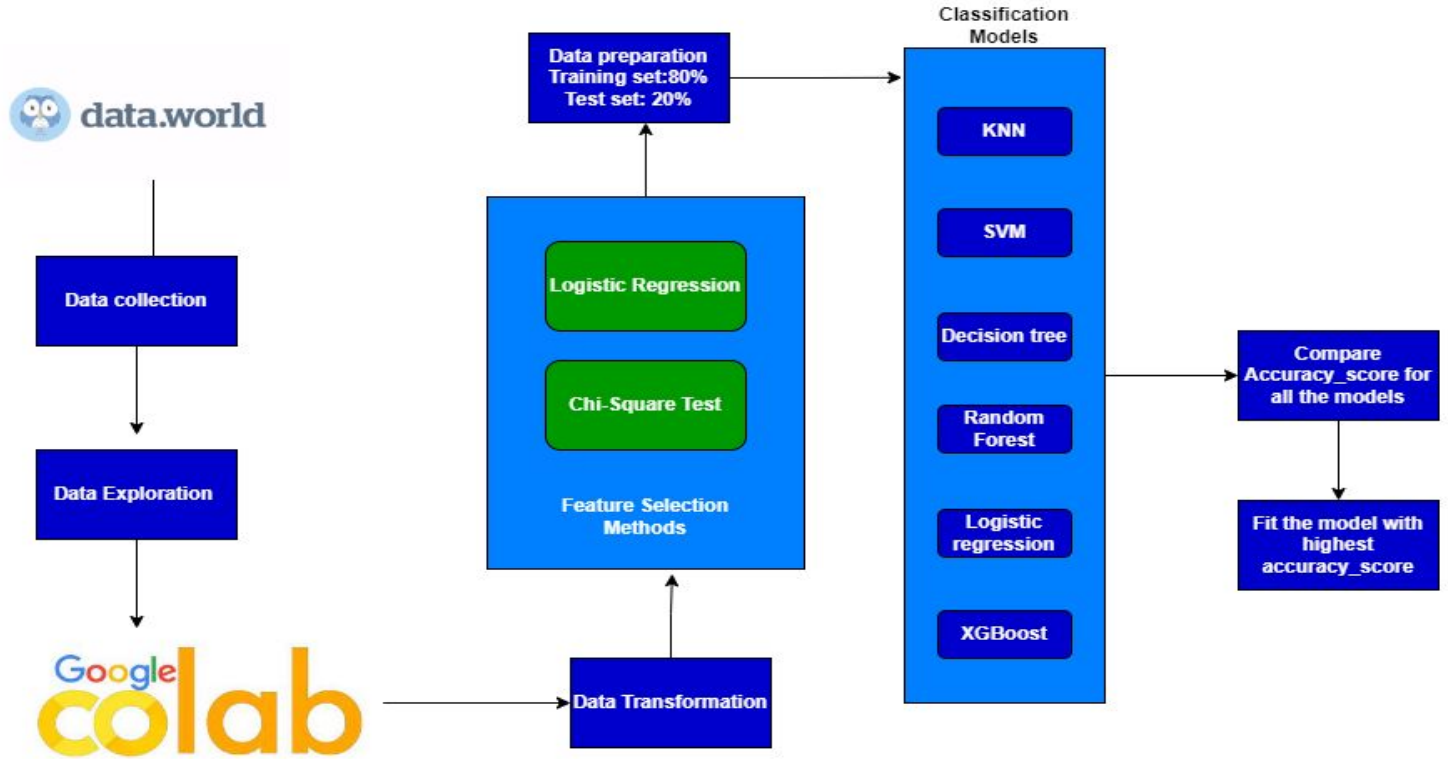
Motivation

- The effective collection and analysis of customer data for better customer experiences is one of the finest methods for an organization to increase its success in the market.
- There are two primary strategies used by marketing sectors, including mass campaigns and direct marketing.
- Compared to direct marketing, mass campaigns receive far less reaction. Because of this, many banking institutions prefer to use direct campaigns that involve phone calls, which have a high success rate.
- Data mining technology is developing and becoming increasingly common in banking institutions and customer-oriented enterprises to forecast target consumer groups for increased sales.

Introduction

- There is a need for a data-driven marketing strategy that uses data analysis to find patterns and trends in consumer behavior to help banks better understand their target market.
- For our study, we have choose to analyze a dataset from the Data World repository that relates to direct phone call banking campaigns in Portuguese banks.
- We are going to provide a model based on bank marketing data to forecast if the consumer has chosen a term deposit or not.
- Machine learning algorithms are employed for statistical analysis in this project, while Python is used as the coding language.
- Since our data is labeled, we intend to classify and analyze it using supervised machine learning models like Random Forest, Support Vector Machine, K-nearest Neighbors, and Naive Bayes.

Flow of the project



Data Summary

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Dataset Download:

https://data.world/xprizeai-ai/bank-marketing

Data File:

Bank_additional_full.csv

Attribute Description

Age: Age of the bank customer.

Job: Customers' job type.

Marital: marital status of the customer

Education: Highest education of a customer

Default: Whether the loan taken by customer is default or not?

Housing: Whether the customer has availed housing loan?

Loan: If the customer has personal loan?

contact : What kind of contact communication type customer has?

month : In which month of the year a customer was last contacted

day_of_week : Last contact day of the week with the customer

duration : last contact duration with customer, in seconds

campaign : number of times contacts performed during this campaign and for this client, including last contact.

pdays : number of days that passed by after the customer was last contacted from a previous campaign

previous : number of contacts performed before this campaign and for each customer

poutcome : result of the previous marketing campaign Like

emp_var_rate : employment variation rate - quarterly indicator

Cons_price_idx: consumer price index - monthly indicator

Cons_conf_idx: consumer confidence index - monthly indicator

Euribor3m: euribor 3 month rate - daily indicator

nr_employed: number of employees - quarterly indicator

y: Whether the client subscribed a term deposit or not?

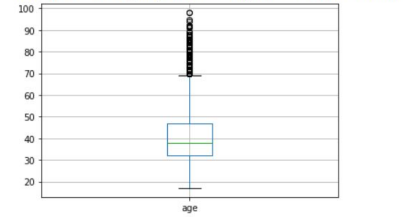
Data Pre-processing

Data Exploration:

- Total Columns: 21
- Columns Int Type: 5
- Columns Float Type: 5
- Columns Categorical Type: 10
- Target Columns Type: bool

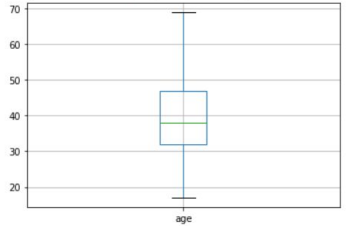
```
data.boxplot('age')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbf45924d90>
```




```
] data.boxplot('age')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbf4535bfd0>
```



Fix Outliers: The columns Age, duration, campaign, Pdays, emp_var_rate has noticeable amount of outliers, which are replaced with median value of the respective columns.

Transform categorical to numerical features : All the categorical columns has been converted to numerical features using Labelencoder() from python library.

Target Class Imbalance

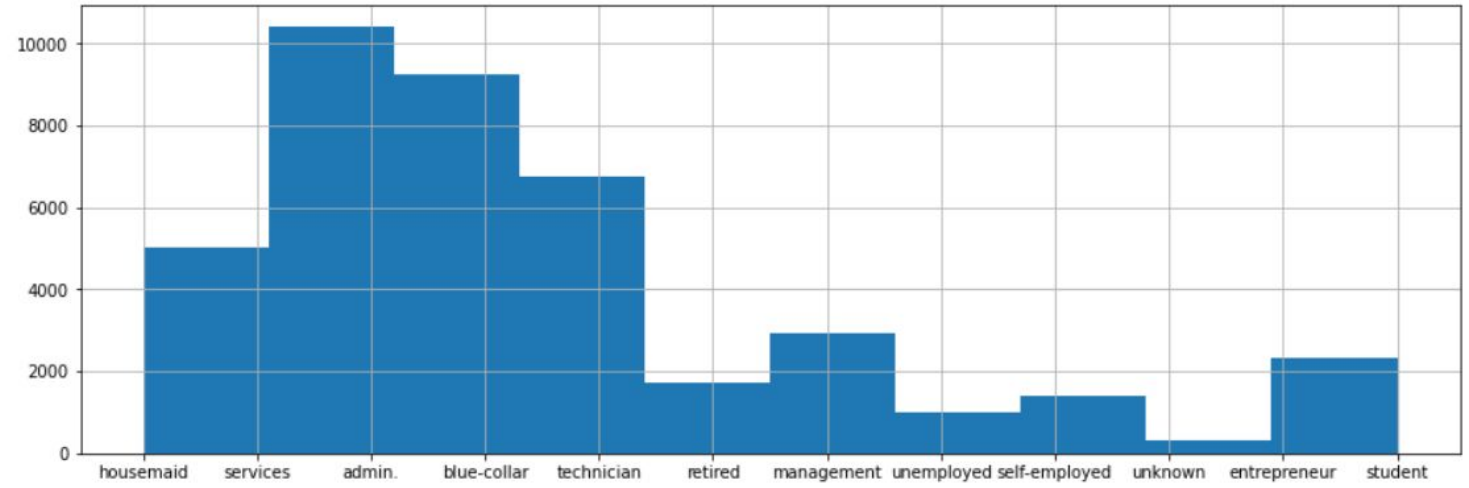
- In the given dataset the target attribute is not balanced.
- i.e., there are 4,640 (11.27%) true values and remaining 36,548 (88.73%) are false values.
- We used oversampling method to balance the training dataset and balanced the target value.

```
over_sampler = RandomOverSampler(random_state=42)
X_train_2, y_train_2 = over_sampler.fit_resample(X_train_2, y_train_2)
```

Data Exploration

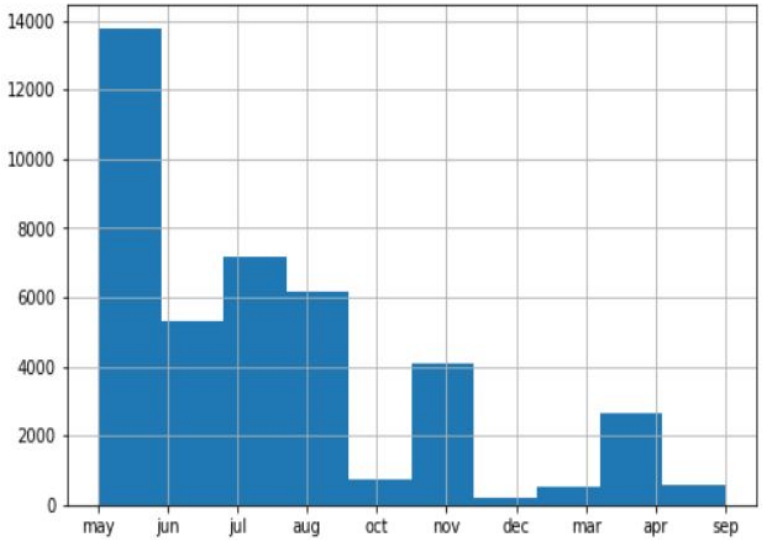
```
[ ]  data['job'].hist(figsize=(15,5))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8b97c29cd0>
```



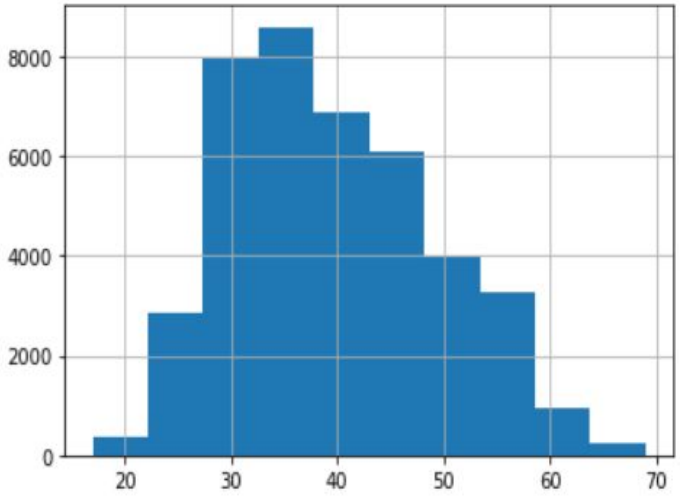
```
[ ]  data['month'].hist(figsize = (8,5))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8b9742b8d0>
```



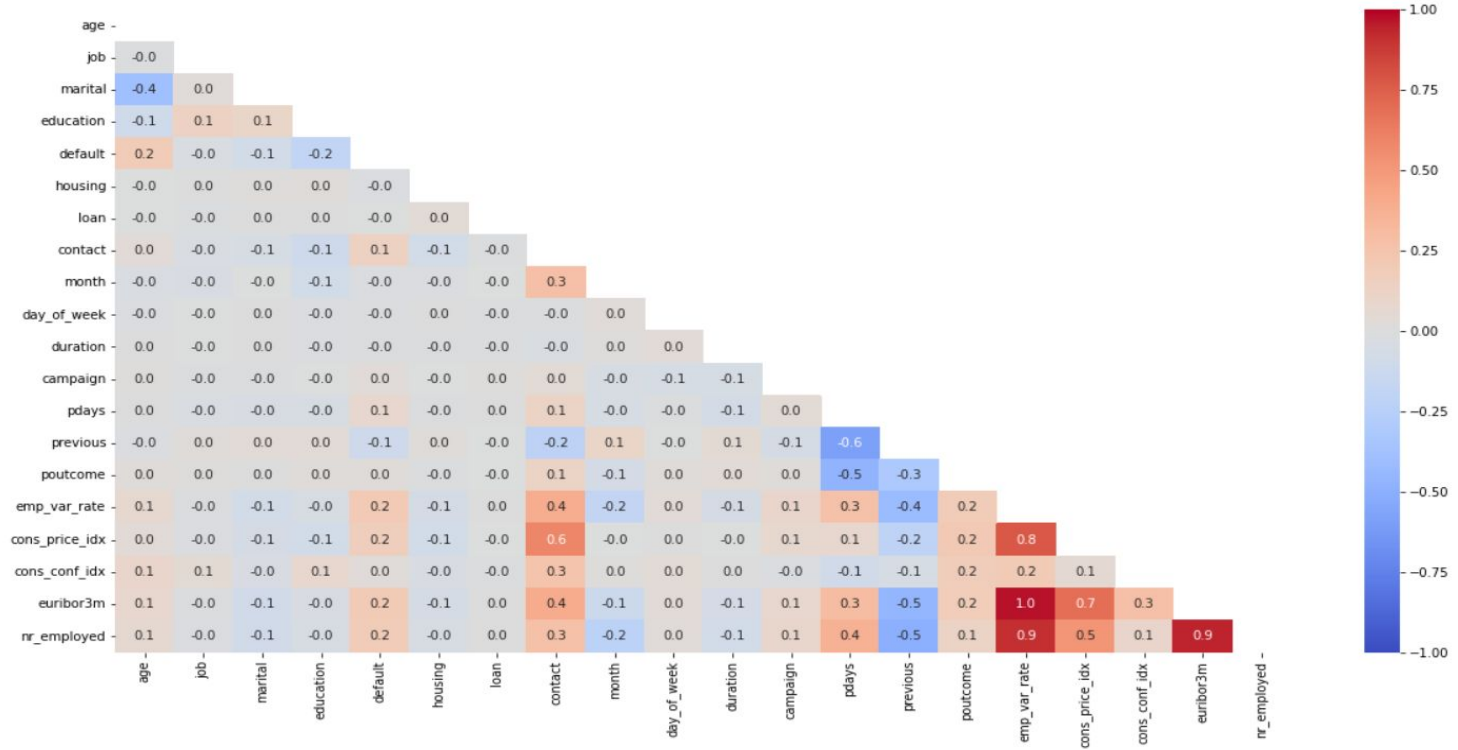

```
data['age'].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0fd8194a00>
```



Correlation Matrix

<matplotlib.axes._subplots.AxesSubplot at 0x7f4906802070>



Feature selection methods

Logistic regression:

- Logistic regression is the type of regression analysis that is used to find the probability of a certain event which is occurring.
- It is the best-suited for cases where we have a categorical dependent variable which can take only discrete values.
- Based on p-value, the features such as age, job, housing, loan, campaign, and preceding are disregarded when employing a 90% confidence level.

```
[ ]  logit_ml = sm.Logit(output_data,input_data).fit()
     logit_ml.summary2()
```

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
age	-0.0013	0.0020	-0.6836	0.4943	-0.0052	0.0025
job	0.0047	0.0050	0.9433	0.3456	-0.0051	0.0145
marital	0.0557	0.0321	1.7372	0.0824	-0.0071	0.1186
education	0.0347	0.0088	3.9408	0.0001	0.0174	0.0519
default	-0.2847	0.0573	-4.9670	0.0000	-0.3970	-0.1723
housing	-0.0089	0.0183	-0.4862	0.6268	-0.0448	0.0270
loan	-0.0071	0.0251	-0.2820	0.7780	-0.0563	0.0421
contact	-0.6857	0.0527	-13.0154	0.0000	-0.7889	-0.5824
month	-0.0988	0.0076	-12.9812	0.0000	-0.1137	-0.0839
day_of_week	0.0512	0.0131	3.9185	0.0001	0.0256	0.0768
duration	0.0043	0.0001	36.1873	0.0000	0.0040	0.0045
campaign	-0.0019	0.0155	-0.1220	0.9029	-0.0323	0.0285
pdays	-0.0010	0.0002	-6.6821	0.0000	-0.0013	-0.0007
previous	-0.0819	0.0536	-1.5273	0.1267	-0.1870	0.0232
poutcome	0.4204	0.0723	5.8123	0.0000	0.2786	0.5621
emp_var_rate	-0.8094	0.0578	-14.0103	0.0000	-0.9226	-0.6961
cons_price_idx	0.6367	0.0281	22.6548	0.0000	0.5816	0.6918
cons_conf_idx	0.0146	0.0041	3.6054	0.0003	0.0067	0.0225
euribor3m	0.6534	0.0671	9.7385	0.0000	0.5219	0.7849
nr_employed	-0.0123	0.0006	-22.3107	0.0000	-0.0134	-0.0112

Chi-Square Test :

- It is a feature selection method that tests the relationship between the features.
- It is also used to test the independence of two events.
- The most significant features, according to the findings of this chi-squared test, are contact, default, education, outcome, job, campaign, age, marital, day_of_week, housing, cons_price_idx, month, loan

```
[ ]  selector = SelectKBest(chi2, k=8)
```

```
[ ]  # dropping negative columns
     final=selector.fit_transform(df_num.drop(['cons_conf_idx','emp_var_rate'],axis=1), data['y'])
```

	Features	P-Value
7	contact	3.500598e-121
4	default	5.521476e-72
3	education	2.464796e-38
14	poutcome	3.722828e-23
1	job	2.179406e-21
11	campaign	1.030246e-16
0	age	3.399709e-10
2	marital	1.348325e-07
9	day_of_week	1.380665e-03
5	housing	2.566075e-02
15	cons_price_idx	9.810078e-02
8	month	1.650559e-01
6	loan	2.077547e-01

Results for Logistic Regression Feature Selection Method

accuracy_score	KNN	SVM	Decision tree classifier	Random Forest	Logistic Regression	XGBoost
With Feature selection	**88%**	**53%**	**89%**	**92%**	**79%**	**86%**
Without feature selection	**88%**	**50%**	**89%**	**91%**	**79%**	**86%**

Even though we reduced 7 features from the total features models such as KNN, Decision tree Random forest etc. are performing with comparable accuracy.

Results for Chi-squared Test Feature Selection Method

accuracy_score	KNN	SVM	Decision tree classifier	Random Forest	Logistic Regression	XGBoost
With Feature selection	**85%**	**43%**	**84%**	**93%**	**70%**	**84%**
Without feature selection	**88%**	**50%**	**89%**	**91%**	**79%**	**86%**

With feature selection we reduced 6 features from the total features. Models such as KNN, Decision tree are performing with comparable accuracy.

Columns importance based on Domain Knowledge

With a small research on banking domain, Among 20 columns in the dataset, we can consider:

- Age is one of the important factor to determine the chance of subscription. People with age of 20 to 50 will have a good income and has high chance of applying for term deposits.
- Job is also one of the important factor, which reveals the capability of the customer to subscribe the term deposit.
- Attributes related to employment ratios are important to know the current financial situation of customers in the market
- Education is also important factor which may have relation with the income that relates with the capacity of customer to subscribe.
- These features are related with each other since income depends on job which internally depends on education level.

Columns importance based on the Analysis:

From the obtained results:

- Discarding the columns age, job, housing, loan, campaign, previous doesn't affect the performance of the model. That implies, somehow these columns are not necessary to predict whether customer subscribes or not.
- As expected SVM is not performing well for this dataset. As we are having many columns it may lead to overlap of the target class. It may be one of the reason for poor performance.
- KNN works well with smaller dataset but we have over 40,000 rows and over 20 features. So it does not perform well.
- Since XGBoost does weight calculation internally there is not much performance difference between feature selected data and data without feature selection.
- Random Forest model with chi square is performing better than all the other models since chi-square can do numerous splits at a single node, resulting in greater precision and accuracy. It also determines the statistical significance between sub-nodes and the parent node.
- It is building multiple trees and ensembles them to get the best performance while being robust to unbalanced data.

Conclusion

- Machine learning will help in this problem to predict about the customer subscription for term deposit.
- This approach will reduce the effort of telephone operators to filter the customers list or atleast to prioritize them to campaign.
- Using machine learning is also inexpensive approach when compared to traditional methods.
- Six machine learning methods applied on the dataset and listed out the best performing method.
- Our future scope would be that we can collect data of different banks and integrate them which may increase the performance of the model and also the quality of the data.