STAGE 4: FINAL PROJECT REPORT
Project Name: Long-term Deposit Subscription

1. Problems and Motivations

This problem seeks to predict whether a customer will subscribe to a long-term deposit for a bank after the bank has used telemarketing advertisement methods. Term deposit is defined as a type of investment that allows clients to deposit their money into a financial institution account with the aim of earning higher interest. Generally, during this fixed period, money cannot be withdrawn from the account. The term of long-term deposit ranges from one year to five years[3]. The dataset itself was compiled in the years following the financial crisis of 2008[1], and will provide a look into the effectiveness of reaching financial customers through automated and distance-based methods.

We believe the results could be particularly useful given the current climate of COVID restrictions. Not only would it model how effective telemarketing methods are in general, but also to identify which customers are most likely to subscribe to the long-term deposit.

Based on the results, we may be able to build a potential customer profile using the features that are most strongly correlated with a subscription. Marketing strategy could also be modified and improved in order to enhance the effectiveness and attract more clients to the program (Term Deposit).

2. Dataset

The dataset is made up of data collected from a Portuguese bank in the 5 years following the financial crisis of 2008. The original team compiling the dataset used feature engineering to reduce the total amount of features from 150 to the 22 most relevant.[1]

The dataset contains the following features:
+ *age* - Customer's age, in years.
+ *job* - Customer's job, subcategorized.
+ *marital* - Customer's marital status (married/divorced/single).
+ *education* - Customer's level of education (secondary/tertiary/primary/unknown).
+ *default* - Whether a customer has a default in their financial history (true/false).
+ *balance* - The customer's current bank balance.
+ *housing* - Whether the customer has a housing loan (true/false).
+ *loan* - Whether the customer has a personal loan (true/false).
+ *contact* - What method was used to contact the customer (cellular/ unknown/telephone).
+ *day* - The date of contact (1-31).
+ *month* - The month of contact (January – December)
+ *duration* - The duration of the call in seconds.
+ *campaign* - The number of contacts performed during this campaign and for this client.
+ *pdays* - The number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted).
+ *previous* - The number of contacts performed before this campaign and for this client.
+ *poutcome* - The outcome of the previous marketing campaign (categorical: ['unknown' 'other' 'failure' 'success'])

These features are used to predict the output variable: *deposit* - Whether the customer subscribed to the long-term deposit (yes/no).

There are a total of more than 11,100 instances in this dataset, with each entry containing 17 attributes. We believe that this dataset will be large enough for training a learning model to predict whether a customer is subscribed to the long-term deposit or not.

3. Proposed Solution

3.1. Preprocessing data

- An exploratory data analysis is conducted on the dataset with the aim of understanding the major characteristics of the dataset as well as each feature's behaviors. Based on that, we may eliminate unimportant or irrelevant features that have little contribution to the predicting process.
- Some of the tasks in this step are listed as below:
  + Identifying any missing values and spotting anomalies
  + Using histogram and boxplot to illustrate each feature's distribution
  + Processing categorical features and converting to numerical values if needed
  + Determining dependencies and relationships among different features (e.g. using correlations to check how strong two features are correlated)

- After data is completely preprocessed, the dataset is splitted into training and testing sets with the ratio of 70% and 30%, respectively. The sets of data can be standardized if needed.

3.2. Model Selection and Model Building

- The dataset features are quite varied; therefore, an optimal Machine Learning (ML) approach is not immediately apparent. In an effort of finding and completing the optimal algorithms and parameters, we analyze this dataset with each of nine ML algorithms. Below is the list of algorithms used:
    + Perceptron
    + Adaline
    + Bagging
    + SGD
    + SVM
    + AdaBoost
    + Decision Tree
    + KNN
    + Random Forest
- The parameters for these algorithms will be tuned by hand using an approximate binary search method.
- The models' performances are evaluated based on running time and scores from the confusion matrix.

4. Data Analysis

As shown from *Figure 4.1*, there are 11,162 records and 17 features in total, including the target feature (*deposit*). Also, no missing values are found within this dataset. Two types of data exist including numeric features and categorical features.

- Numeric features: *age, balance, day, duration, campaign, pdays, previous*
- Categorical features: *job, marital, education, default, housing, loan, contact, month, poutcome, deposit* (this is the target variable)

```
RangeIndex: 11162 entries, 0 to 11161
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        11162 non-null  int64
 1   job        11162 non-null  object
 2   marital    11162 non-null  object
 3   education  11162 non-null  object
 4   default    11162 non-null  object
 5   balance    11162 non-null  int64
 6   housing    11162 non-null  object
 7   loan       11162 non-null  object
 8   contact    11162 non-null  object
 9   day        11162 non-null  int64
 10  month      11162 non-null  object
 11  duration   11162 non-null  int64
 12  campaign   11162 non-null  int64
 13  pdays      11162 non-null  int64
 14  previous   11162 non-null  int64
 15  poutcome   11162 non-null  object
 16  deposit    11162 non-null  object
dtypes: int64(7), object(10)
```

Figure 4.1: Data information

4.1. Analyzing numeric features

	age	balance	day	duration	campaign	pdays	previous
count	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000
mean	41.231948	1528.538524	15.658036	371.993818	2.508421	51.330407	0.832557
std	11.913369	3225.413326	8.420740	347.128386	2.722077	108.758282	2.292007
min	18.000000	-6847.000000	1.000000	2.000000	1.000000	-1.000000	0.000000
25%	32.000000	122.000000	8.000000	138.000000	1.000000	-1.000000	0.000000
50%	39.000000	550.000000	15.000000	255.000000	2.000000	-1.000000	0.000000
75%	49.000000	1708.000000	22.000000	496.000000	3.000000	20.750000	1.000000
max	95.000000	81204.000000	31.000000	3881.000000	63.000000	854.000000	58.000000

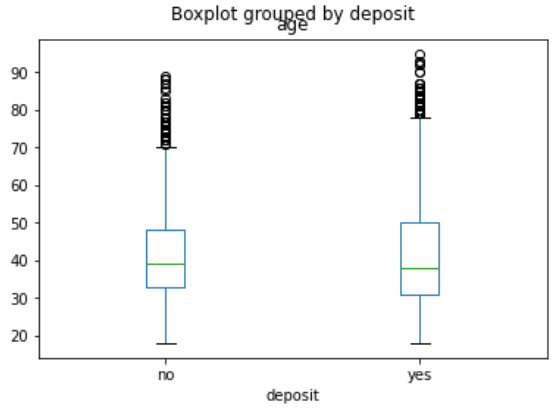
Figure 4.2: Descriptive statistics of numeric features

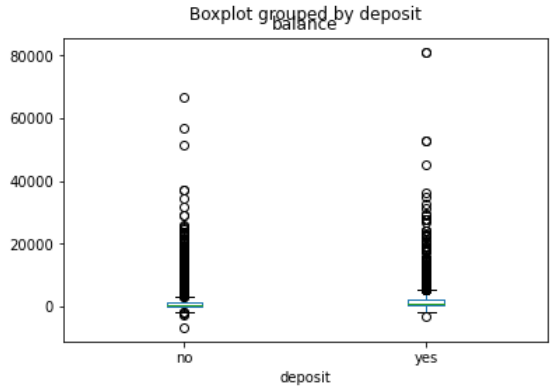
- In *Figure 4.3*, the correlation matrix provides pairwise correlations for any pair of features in the table. It is used to identify any potential linear relationships between different features with the aim of determining the best set of features used for the machine learning model. *Figure 4.3* illustrates that none of these features are highly correlated to each other because their absolute correlation values are extremely low. As a result, these numeric features can be incorporated in the classification model.

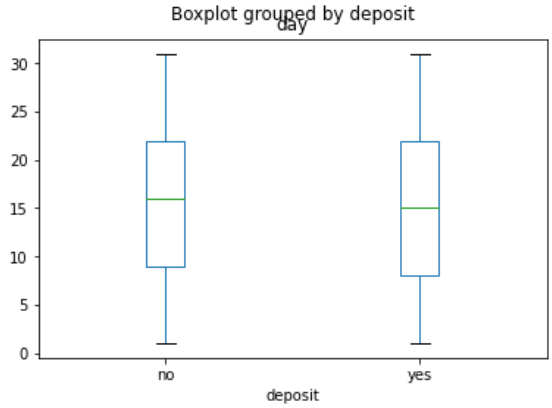
	age	balance	day	duration	campaign	pdays	previous
age	1.000000	0.112300	-0.000762	0.000189	-0.005278	0.002774	0.020169
balance	0.112300	1.000000	0.010467	0.022436	-0.013894	0.017411	0.030805
day	-0.000762	0.010467	1.000000	-0.018511	0.137007	-0.077232	-0.058981
duration	0.000189	0.022436	-0.018511	1.000000	-0.041557	-0.027392	-0.026716
campaign	-0.005278	-0.013894	0.137007	-0.041557	1.000000	-0.102726	-0.049699
pdays	0.002774	0.017411	-0.077232	-0.027392	-0.102726	1.000000	0.507272
previous	0.020169	0.030805	-0.058981	-0.026716	-0.049699	0.507272	1.000000

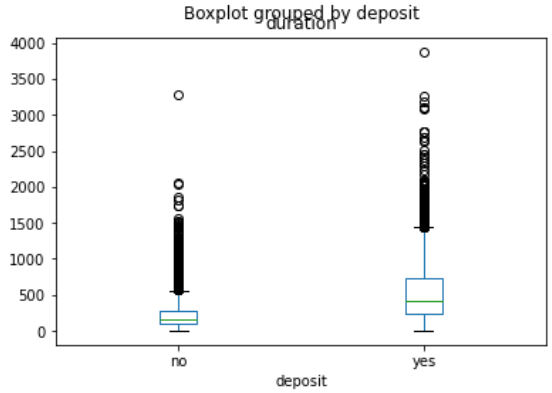
Figure 4.3: Correlation matrix among numeric features

- *Figure 4.4* represents the distribution of each numeric feature grouped by the target variable *deposit* using boxplots.
+ The boxplot of *age* suggests that people in the range of 30 to 50 years old are not likely to have long-term deposits. In real life, people at those ages are mostly the main workforce, who have pretty stable income sources and also have high demand for spending. That probably explains why they do not want to deposit their money in the bank at that time. On the other hand, there are more retirees who age above 60 that agree to subscribe to the long-term deposit program (indicated by "yes" group).
+ In the boxplot of *balance*, it is illustrated that the range of the account balance of people who join the program is slightly higher than that of other people.
+ The boxplot of *day* shows that most clients are reached out in the middle of the month (from the 8th to the 22nd).
+ The boxplot of *duration* reflects that when the duration of the last contact is around zero, which means no calls are made, there are no subscriptions made (indicated by "no" group). The values of *duration* feature are known only after the contacts are made. However, as we attempt to build a model to predict program subscription before the contact is made, this feature should be dropped.
+ The last three box plots relate to the previous marketing campaigns, which shows that there were only a few calls made before this current campaign.









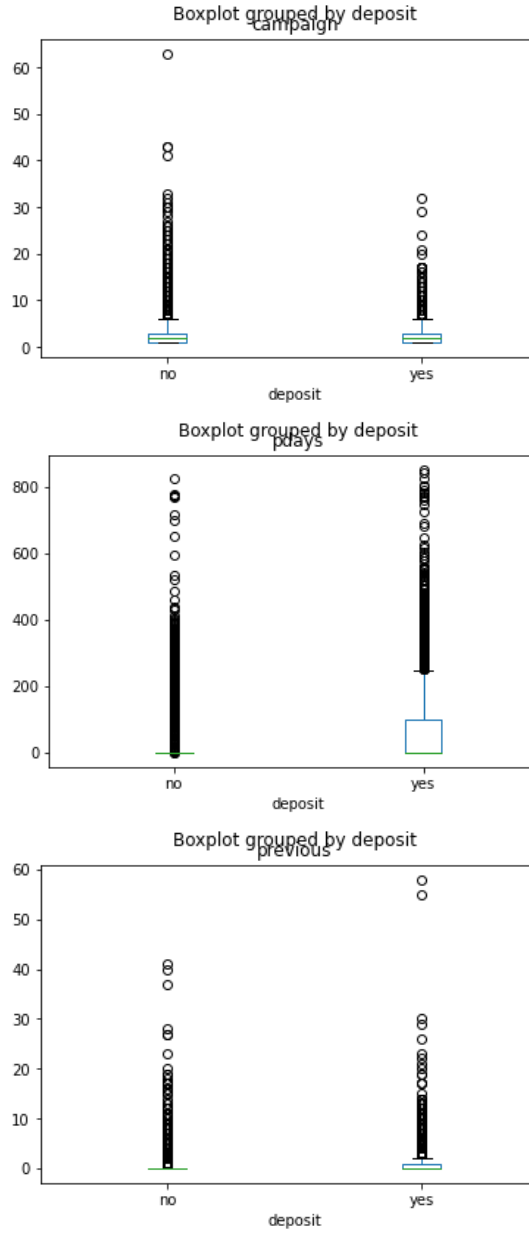


Figure 4.4: Boxplot of each numeric feature grouped by "deposit" variable

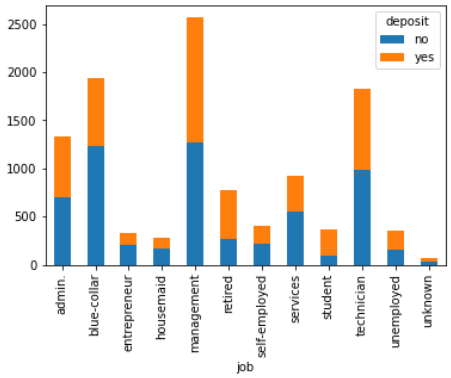
4.2. Analyzing categorical features

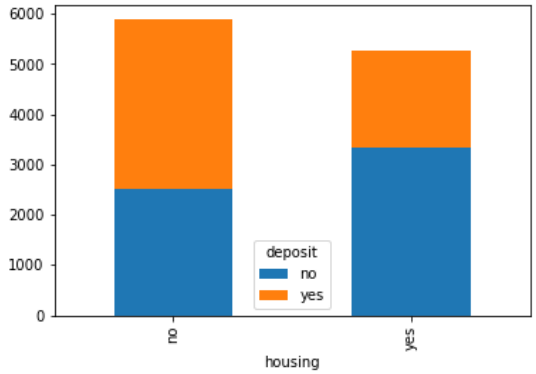
- First off, column *deposit* in *Figure 4.5* indicate that this dataset is a balanced sample because the two labels are distributed uniformly ("no" accounts for 52.62% and "yes" accounts for 47.38%).

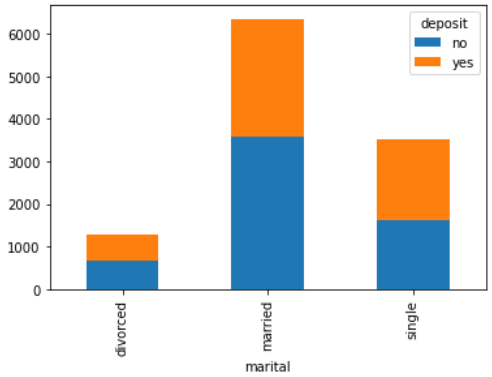
	job	marital	education	default	housing	loan	contact	month	poutcome	deposit
count	11162	11162	11162	11162	11162	11162	11162	11162	11162	11162
unique	12	3	4	2	2	2	3	12	4	2
top	management	married	secondary	no	no	no	cellular	may	unknown	no
freq	2566	6351	5476	10994	5881	9702	8042	2824	8326	5873

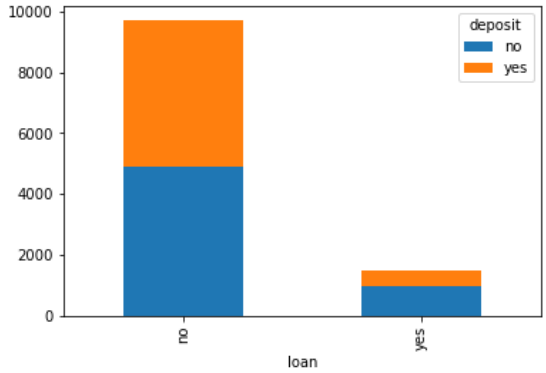
Figure 4.5: Descriptive statistics of categorical features (including the target variable "deposit")

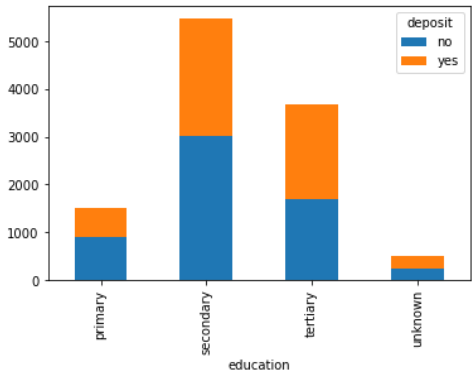
- *Figure 4.6* consists of multiple stacked bar charts of all categorical features grouped by the target variable *deposit*.
  + In the chart of *job*, even though management, blue-collar and technician workers are contacted the most, they do not have the highest rate of subscription. Indeed, students and retirees are the ones who have the highest possibility to join the program.
  + The charts of *martial* and *education* show the same pattern. That is, the subscription rate is distributed almost evenly among different subgroups.
  + The chart of *default* shows that the bank only targets clients who have no default in their history.
  + Looking at the charts of *housing* and *loan*, it can be stated that people who have either a housing loan or personal loan are not likely to put their money into long-term deposits.
  + *month* feature may be dropped because it does not show any specific patterns in determining client subscription.
  + *poutcome* represents the previous marketing campaign results. As shown in the corresponding chart, the probability of client subscription to the current program significantly increases if the previous campaign was successful.

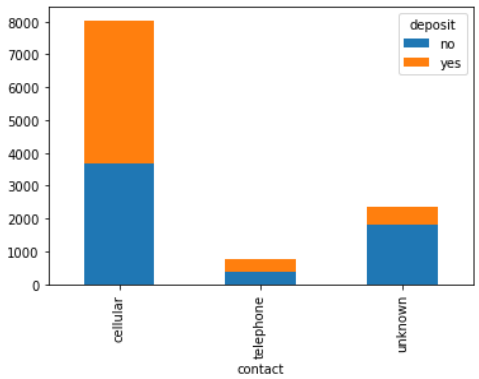

deposit
no
yes
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
unknown
job

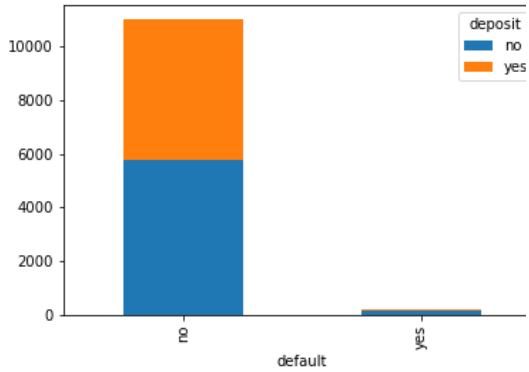

deposit
no
yes
no
yes
housing

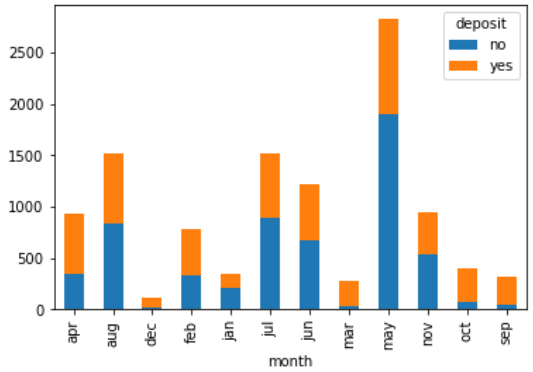

deposit
no
yes
divorced
married
single
marital


deposit
no
yes
no
yes
loan


deposit
no
yes
primary
secondary
tertiary
unknown
education


deposit
no
yes
cellular
telephone
unknown
contact


deposit
no
yes
no
yes
default


deposit
no
yes
apr
aug
dec
feb
jan
jul
jun
mar
may
nov
oct
sep
month

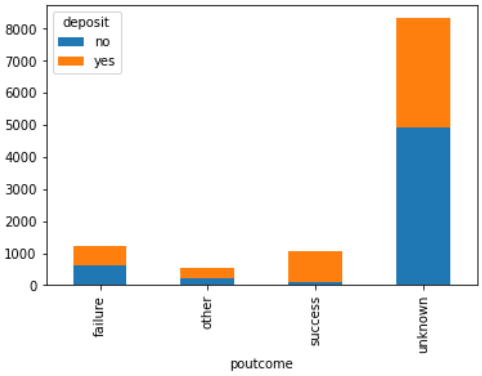


Figure 4.6: Stacked horizontal bar charts of categorical features grouped by "deposit" variable

5. Classification Result Analysis

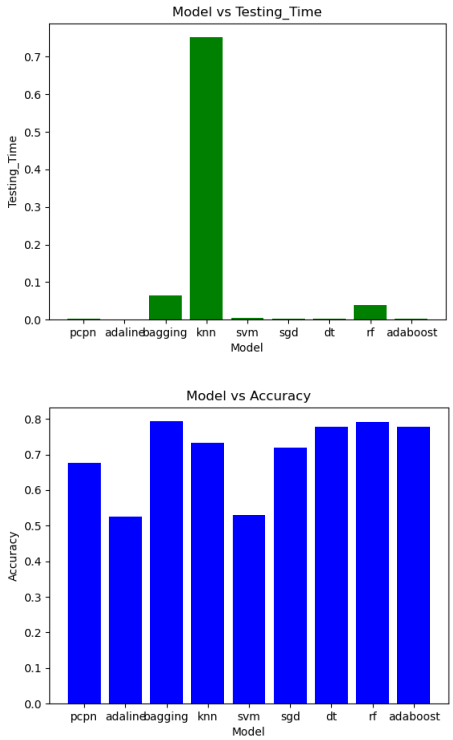


Figure 5.1: The results of running the models over the original dataset with all columns

- The models perform well, but not particularly high. The highest examples of accuracy all tend to come from aggregative methods. The only real anomaly in terms of time is KNN, which takes 0.6 seconds more than any other model to train and run.

contact	-0.249847
housing	-0.203888
campaign	-0.128081
poutcome	-0.122369
loan	-0.110580
day	-0.056326
default	-0.040680
month	-0.037121
age	0.034901
job	0.063395
marital	0.067610
balance	0.081129
education	0.095948
previous	0.139867
pdays	0.151593
duration	0.451919

Figure 5.2: Correlations between all features versus the target feature, "deposit"

- In the interests of gaining an accuracy boost, or potentially optimizing the amount of time it takes to run through our dataset, we can remove anything not well correlated (in either direction) with our target variable. In this case, because the maximum correlation between single variables is ~.45, we drop anything in the range of -.10 to .10 correlation - **education, balance, marital, job, age, month, default, and day**. With a smaller number of features to process and train against, we should be able to achieve similar accuracy in less time.

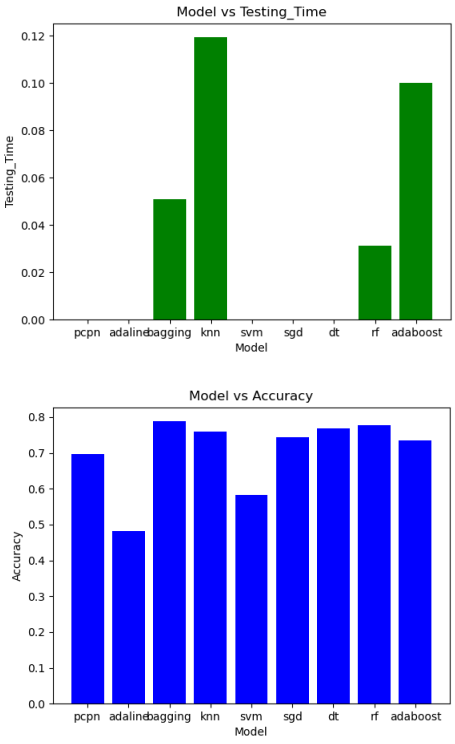


Figure 5.3: Results from running the models over the dataset with dropped columns

Model	*Time*	*Accuracy*	*Time*	*Accuracy*
PCPN	*0.0029*	*0.676620*	*0.0027*	*0.696626*
Adaline	*0.0*	*0.525231*	*0.0010*	*0.482831*
Bagging	*0.0639*	*0.793072*	*0.0377*	*0.788295*
KNN	*0.7518*	*0.733055*	*0.0906*	*0.760525*
SVM	*0.0039*	*0.530009*	*0.0029*	*0.581965*
SGD	*0.0033*	*0.719021*	*0.0019*	*0.520155*
DT	*0.0020*	*0.777545*	*0.0030*	*0.768588*
RF	*0.0379*	*0.790982*	*0.0156*	*0.778143*
AdaBoost	*0.0029*	*0.776650*	*0.0922*	*0.735145*

Figure 5.4: A table with the original dataset results on the left and the dataset with dropped columns on the right.

- In terms of accuracy, the reduced dataset shows little improvement over the full dataset, even coming in noticeably worse for some models such as Adaline (which is almost 5 percent less accurate) and SGD (which is almost 20 percent less accurate!). Most of the models, however, show around the same accuracy. The real draw in reducing the dataset is actually the time taken to train and test the model: with the exception of Adaline, every single model performs better in terms of time. Bagging, Random Forest and Decision Trees all keep their high accuracy while clocking in at extremely low times, particularly Decision Trees, which seems to perform the best of all of the models when considering both accuracy and time.

6. Conclusions

In conclusion, the accuracy of the Adaline, SVM, and SGD methods is far too low to do anything realistic with. The aggregative methods such as Bagging, Decision Trees, and Random Forest have the highest accuracy and runtimes (whether the data is reduced or not) and would be able to accurately predict whether a customer would make a deposit or not for about 4 out of every 5 customers. The majority of the correlated features tend to be related to how the bank handles customer interaction. For example, what method was used to contact the customer and the duration of the call with the customer. This is opposed to customer data like education, age, or whether they have previously had a default in their financial history. Because of this, we are confident that improving the bank's marketing strategy would be the most efficient way to increase deposits, and this work can be successfully done without having to worry about collecting large amounts of private customer data.

7. References

[1] [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[2] Bank Marketing Dataset. Retrieved from https://www.kaggle.com/janiobachmann/bank-marketing-dataset

[3] Chen, J. (March 2020). Term Deposit Definition. Retrieved from https://www.investopedia.com/terms/t/termdeposit.asp

[4] Applied ML Project. https://github.com/psauers/Applied-ML-Project