Data-Driven Innovation Final Project CRISP-DM *

Aitore Issadykova
Data Science
Nazarbayev University
Nur-Sultan, Kazakhstan
aitore.issadykova@nu.edu.kz

Assem Kussainova
Data Science
Nazarbayev University
Nur-Sultan, Kazakhstan
assem.kussainova@nu.edu.kz

Zhaniya Koishybayeva
Data Science
Nazarbayev University
Nur-Sultan, Kazakhstan
zhaniya.koishybayeva@nu.edu.kz

***Abstract*—CRISP-DM is most used and effective methodology for all Data related projects. CRISP-DM methodology consists of 6 steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. Every process of the CRISP-DM methodology helps to figure out how to add a value for the business processes. This paper provided all stages of the CRISP-DM process for the marketing campaign new product line of the Bee Bank which is a term deposits. The resulted model can be used for the future marketing campaign of the bank with adding additional data**

***Index Terms*—CRISP-DM, Data Science, Data Mining, Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment, Marketing Campaign, Bank, classification**

I. Introduction

CRISP-DM process was proved as one of the successful methods for managing, running and deploying the Data Science projects.CRISP-DM consists of 6 stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. Every stage is crucial for businesses, since it should provide detailed cost-benefit analysis, as well as the timeline of the whole project, the all processes that were done during the project, its results and future maintenance and support issues.

This paper will perform CRISP-DM process for the domestic commercial bank ”The Bee Bank”, which would know to predict the results of the marketing campaign for their customers with new term deposit product. Every single stage of the CRISP-DM will be covered explicitly.

II. Business Understanding

A. Business Objectives Determination

1) Business Background: The commercial bank ”The Bee Bank” is domestically founded Tier 2 bank which have several offices in each city of the country. ”The Bee Bank” known in the country as the most reliable bank and have a great loyalty program for its customers. The major bank issues as well as bank development is solved on the Board of Directors meeting. The daily operational issues under consideration of the bank CEO, which was selected by Board of Directors of the bank. The organizational structure of ”The Bee Bank” can be seen on the Fig.1, where participants in the current project were highlighted by turquoise color.

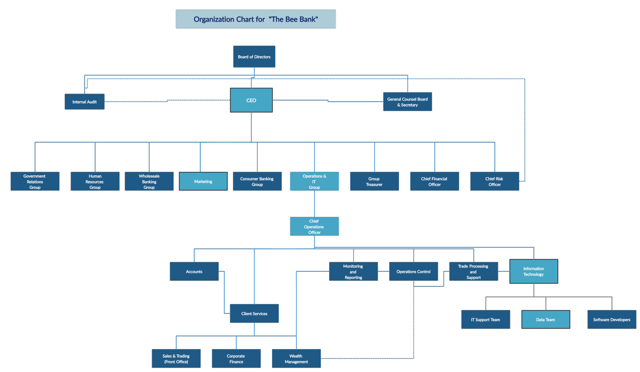

Fig. 1. Organizational Structure

2) Business Objectives: ”The Bee Bank” introduced new bank product - term deposits. As part of their loyalty program, term deposits are currently available only for their current customers. The main goal of this project is to predict the certain marketing campaign result on the particular client. Results of this process will directly affect Marketing and Client Services Departments. By obtaining the prediction model of the campaign based on the general customer data, it will decrease costs of marketing campaign because it will be targeted on the particular group of clients, and Marketing and Client Services Departments will reach only this group of clients; maintain current bank reputation by not disturbing all clients and will obtain general model for future campaigns, since the data of customers may just be added to the particular model. Moreover, the successful marketing campaign will be awarded by the Board of Directors and will increase communication on the experience sharing with National Bank and other banks.

3) Success criteria: The successful prediction model of the marketing campaign results will rise funds available for the banks, since targeted customer will open term deposits, save costs on the introducing campaign for not interested customer and will not lose customers due to enormous activity from marketing campaign and in the long run it will attract new customers. It will resulted in rising of the operational turnover, customer loyalty and bank reputation. Moreover, successful model will attract other banks for additional agreements as well as new communication with National Bank.

B. Assessing Situation

1) Resource Inventory

Personnel:: The project will be mostly lead by Data Team Department of "The Bee Bank". Other participants in this project are Marketing Department and CEO. Marketing Department employees will communicate with Client Services Department and will help with data collection from their customers. CEO will help with communication with the Board of the Directors as well as perform as auditor on the current project.

Data: The Marketing Department provided the data of the bank current customers, which included the age, job, phone number, martial status, education level, housing availability and loan existence. The Marketing team will expand the data by introduce to the some part of the clients the new term deposits product and then will record whether customer is interested to get term deposit in the bank. In addition, data analysts in the Data Team will perform some each consumer data analysis and provide some data about consumer price index, consumer confusion index and possible outcome.

Hardware resources: The bank have their own Data Team Department, which includes data engineers, data scientists and data analysts. The Data Team have sufficient hardware resources such as new multicore computers with several monitors and was granted partial access to the customer database.

2) Requirements, Assumptions, and Constraints: There are several requirements for this project that should be assessed first.

Legal Requirements:

According to the Law, Bank employee should not threat the bank customers and use the personal information of the customers for their personal goals. The customer should be informed that his records will be included in the bank database and customer should be called by Marketing team only within working hours.

Security Requirements:

Since the customers data is sensitive, the Marketing and Data Team Departments should complete and sign the non-disclosure information form. Moreover, there should be some security actions performed for storage of the data and passwords to the databases.

Assumptions:

There no costs related to the data collection, since the data of the bank customers was collected before project initiation. The data assumed to be sufficient for the Data Science Project. CEO and Board of the Directors expected to receive the resulted model with some explanation of model selection and results. In addition, bank may sell the model of the marketing campaign company for other banks.

Constraints:

The legal and security constraints will be completed as it was mentioned in the requirements part. The financial constraint of this project is 2 million budget. This budget should include the payments of the whole team during project and costs related to launch the result of this project.

3) Risks and Contingencies: Data:

We already assumed that data quality will be sufficient for model analysis, however it can be resulted that data amount is not enough or there are some required missing data. The bank addressed this risk by asking Marketing team to conduct the introduction of campaign to the part of the customers.All data collection processes conducted by phone by addressing legal requirements.

Scheduling:

The risk related with data quality can affect the scheduling, but the contingency plan of data collection will reduce the risk of scheduling issue. Moreover, the Bank is actively recruit new Data Team employees, which will speed up the time of the whole project.

Financial:

One of the main risk of this project is over budget. This risk will be assessed by CEO of the bank, which will deliver the additional raised costs for the Board of the Directors.

Results:

This project aims to increase operational turnover of the bank and rise its funds for the investments. These factors will increase the profit of the bank as well as reputation of the bank among other banks.

4) Terminology: The terminology will include only some data science terms and banking related terms.

Model Selection - choose the best model between the proposed ones.

Classification - the systematic grouping of the data points by their common features

Regression Analysis - the statistical analysis which shows the relations between input and output data

Term deposit - fixed-term investment that includes the deposit of money into an account at a bank. [1]

Loan - type of credit vehicle in which a sum of money is lent to another party in exchange for future repayment of the value or principal amount. [2]

5) Costs/benefits: The potential costs of this project are additional data collection, additional staff employed and the costs related for creation prediction model, i.e., service server payments for data storage memory increasing, payments for providers of the data modeling software and new equipment. The benefit of this project is to increase indirectly the bank profit without losing in the bank reputation and customer loyalty scores.

C. Data Mining Goals Determination

1) Goals: The main goal of this Data Science project is the prediction of the results on the marketing campaign. For this project two type of the models will be considered. First is classification models such as Decision Tree, Random Forest and others. Second model is logistic regression which will calculate the probability of customer to get term deposit. Both models will get as the input the customer data and will predict label "yes" or "no" with some additional threshold inputs.

These models will assess the business goal of effective marketing campaign on the bank products which will help bank to target only specific auditory and do not disturb other clients. It is expected that effective target marketing campaign will increase the operational profit of the bank.

Fig.2 is an example of effective target marketing by other companies. It shows the difference between traditional marketing campaign and targeted campaign for different sectors of the company [3].

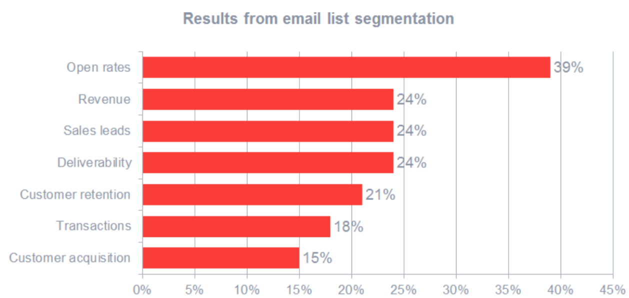

Fig. 2. Increase in outcomes for targeted marketing campaign [3]

2) Success Criteria: All models will be assessed with accuracy for classification by splitting initial available data on test and train. The benchmark of the classification model is accuracy and model will be selected if the accuracy will be higher than 90%, since the main goal is correct customer segmentation for the giving campaign. In case of more than two models will achieve the threshold, the additional analysis on the classification result will be conducted.

If results will outperform of expected results or will give same results as models provides after deployment of targeting marketing campaign, it will be considered as data mining project success and the data science team will be awarded by the Board of Directors and will be collaborate with other banks Data Science Teams.

D. Project Plan

1) Plan: The initial project plan with phases, time completion, resources used and risks are described in Table 1. Risks of each phase as well as risk addressing techniques were covered in II-B3 section.

TABLE I
INITIAL PROJECT PLAN

Phase	Time	Resources	Risks
Business Understanding	2 weeks	Data Team, Marketing and CEO	Financial
Data Understanding	2 weeks	Data Team and Marketing	Data and Scheduling
Data Preparation	4 weeks	Data Team	Data and Scheduling
Modeling	4 weeks	Data Team	Scheduling
Evaluation	2 weeks	Data Team and CEO	Results
Deployment	2 weeks	Data Team, Marketing and CEO	Financial and results

2) Initial Assessment of Tools and Techniques: For this project purposes SQL databases will be used for data extraction and Python supporting software such as PyCharm will be used for Data Preparation, Modeling and Evaluation phases. All licences were bought by bank before project launching.

III. DATA UNDERSTANDING

A. Data Collection

The data about marketing campaigns of a Portuguese banking institution used in this paper belongs to UCI Machine Learning Repository, which is a collection of databases that are used by the machine learning community. This data includes four data files, among which were chosen two datasets for which bank information was collected for 2 years from 2008 to 2010. The phone calls with the customers of bank formed a basis for marketing campaigns of new term deposits [4]. There are could be several calls made to the same customer in order to obtain desired information and get the answer whether the client will subscribe to the new product or not.

Desired data was written into two .csv files:

- bank-full.csv
- bank-additional-full.csv

B. Data Description

Datasets can be freely analyzed by any tool as they are in an open access. The sizes of data files bank-full.csv and bank-additional-full.csv is 3.57Mb and 4.7Mb respectively, and they are represented in the format of table with the total number of records of 45211 for first dataset and 41188 for second [4]. They share 15 attributes which include bank client data, information related with the last contact of the current campaign and other attributes with output variable. The input variables are:

- age: customer age written in numerical format;
- job: type of job of the customer represented as categorical variable in string format;
- marital: values represent marital status of customer in string format;
- education: provides the completed education level of the customer with categorical values in string format;
- default: shows whether the customer has credit represented with binary values 'yes' or 'no';
- housing: values 'yes' and 'no' are used to indicated if the customer has housing loan or not;
- loan: similar to the attribute 'housing' shows information about loan, but of personal type using binary variable in string format;
- contact: variable points out the contact communication type of the customer, for ex. 'telephone' or 'cellular';
- month: represents last month when the customer was contacted in particular year in numerical format;
- duration: indicates how long was the last call with the customer in seconds;
- campaign: shows how many times the customer was contacted during the current marketing campaign in numerical format;
- pdays: variable is used to display the number of days passed after the contact with customer during previous campaign in numeric format;

- previous: number of calls made with customer before current campaign;
- poutcome: whether the previous campaign was successful or not is shown with categorical variable;
- y: the target variable, which denotes whether the client subscribed a term deposit using binary string format.

In addition to the main columns, each dataset also has its own unique attributes, which include the average annual balance of the client in euros, the last day in the month of contact with the client, which is replaced by the day of the week of the last contact with the client, and employment variation rate, consumer price index, consumer confidence index, euribor 3 month rate and number of employees all in numeric format from bank-additional-full.csv dataset.

C. Data Exploration

The most meaningful attributes for exploring current dataset trends are those which represent personal information of customer, the duration of last call with customer and the number of previous contacts, while the least important are the employment variation rate, consumer price index, consumer confidence index, euribor 3 month rate and number of employees as they are not available for first dataset bank-full.csv and they cannot be averaged for all customers. Following is some useful information gained from dataset.

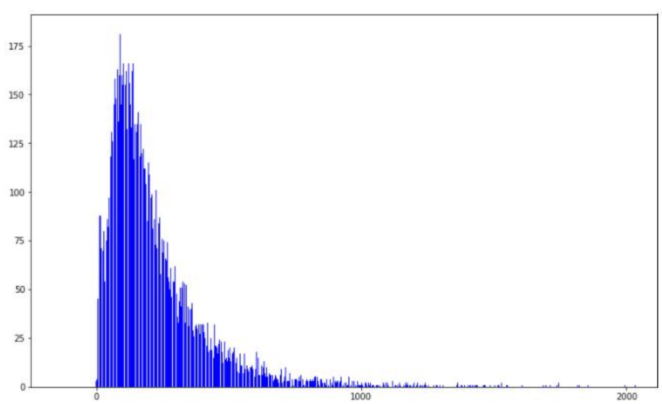

Fig. 3. Refusal from term deposits represented by call duration

The main information gained from this dataset is that the number of customers' negative answers tends to decrease as the duration of call increases. Fig.3 illustrates the histogram where can be seen the distribution of unsuccessful offers for subscription.

The amount of money in the client's account also affects the refusal to subscribe to a new product. According to the constructed histogram on Fig.4 of "no" response frequencies with different amounts of funds on the client's balance, the most negative answers are given by clients with a balance closer to 0, and then this number decreases with an increase in the client's balance. This can be useful for identifying potential customers who might consider a new product proposal.

With regard to the personal data of the client, for example, the distribution of clients who refuse to sign up for a deposit by

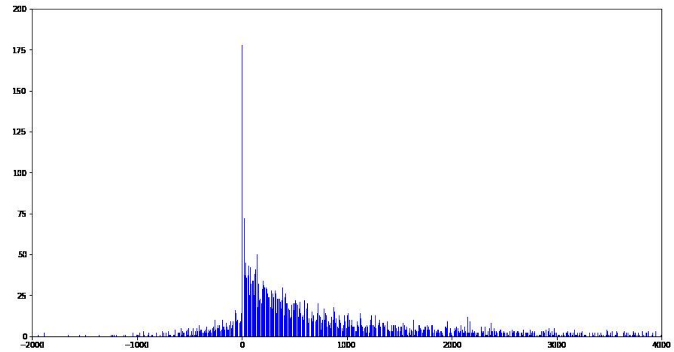

Fig. 4. Refusal from term deposits represented by customer balance

their type of work can be observed in Fig.5, where the largest percentage taken by management, blue-collar and technician workers. Moreover, individual attributes of the client's personal data cannot be good predictors of refusal or acceptance of an offer to open a deposit, but their effect can be studied in aggregate on the desired target.

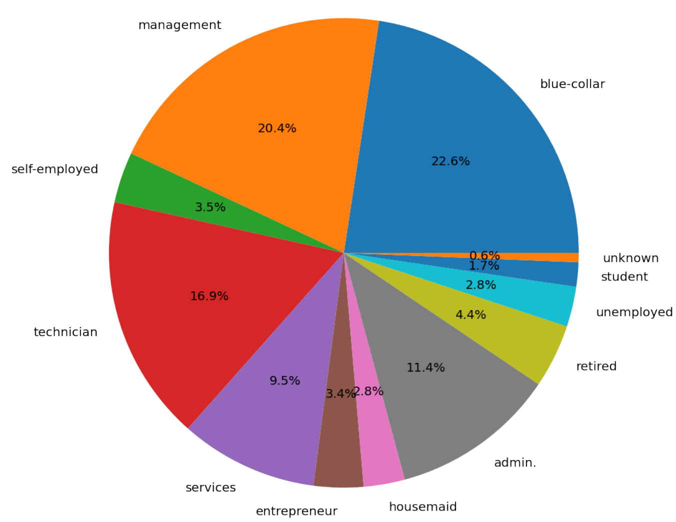

Fig. 5. Clients job type chart

Overall, this datasets can be further used for classification purposes with the help of Data Mining and Machine Learning techniques including Random Forest, Naïve Bayes, SVM, Decision Tree and etc. The regression models might be also useful for this datasets after changing data structure.

D. Data Quality

This step requires an assessment of the quality of the data, as any inconsistencies can affect the progress of the project. After detailed consideration of chosen datasets, it was identified that all the data contain complete information about each customer with some missing values denoted as unknown. In addition, not all necessary information related to the attributes for further data analysis is present, and it would be better to merge two datasets to ensure data completeness and to reduce the number of input attributes which don't affect much on the target variable. Missing data presence could be resolved with deletion of whole row about customer or 'unknowns' can be used as a separate class of attribute. The representation of data full of integrity and is user friendly, so it doesn't need to be restructured. It will take a small time to preprocess data and understand the meaning of fields for other researchers. These datasets have a variety of categorical data, whose values should be transferred to numerical format, and numerical data should be normalized according to specified range of attributes to get rid of outliers before further usage in prediction models. Moreover, repeating customers' data should be removed from both datasets to avoid the data redundancy.

In general, datasets follow the principles of accuracy, completeness, consistency, relevance and reliability, as well as sufficient coherency, definition and time characteristics of good quality data.

IV. Data Preparation

A. Data Selection

To begin with, it was required to select the data that will be used for training in the Modeling phase. At this stage, both attributes and cases are selected. In our project, a strategy is being developed to increase demand for a new bank product among existing customers, therefore, our work has no limitations on cases as all the collected data is directly related to the bank's clients. When choosing the data, the issues of the potential relevance of the attribute in the problem being solved, the quality of the attribute for use in the model and restrictions on the use of attributes were considered.

In each dataset, a selection was made of the most useful attributes, the data of which will be combined into one dataset in the future. For example, for this project the type of connection with customer does not affect the final output, therefore it seems useless as a predictor for forecasting and was removed. Similarly, all mismatched attributes, which were not carrying useful information (for example, up to two unique values in the column) including employment variation rate, consumer price index, consumer confidence index, euribor 3 month rate, and number of employees were removed from the list of final attributes. However, since the customer's balance affects the success of the marketing campaign, this attribute has been retained. The dataset have no attributes that are limited in use due to privacy issues, but some of them contain 'unknown' values, which were chosen to be preserved for future use, since unknown values make up a minority of all values, which does not exclude the attribute's usefulness. An example of such attribute is a job type of customer on Fig.6, where it can be seen that only 288 values out of total are 'unknown'.

B. Data Cleaning

Before this stage, potentially interesting data were selected, and now it is necessary to check their quality. The output is 3 types of attributes - quality attributes, corrected attributes and rejected attributes.

```
{'admin.': 5171,
 'blue-collar': 9732,
 'entrepreneur': 1487,
 'housemaid': 1240,
 'management': 9458,
 'retired': 2264,
 'self-employed': 1579,
 'services': 4154,
 'student': 938,
 'technician': 7597,
 'unemployed': 1303,
 'unknown': 288}
```

Fig. 6. Clients job type values distribution over an attribute

Missing values must be either filled in or removed from consideration. Among the selected attributes, 6 out of 7 associated with the personal information of a bank customer contain unknown values. As mentioned earlier, in all attributes, these values are a small fraction of the total number of all values of each attribute. In addition, some Data Mining models work well with missing data. For this reason, it was decided to consider unknown values for further investigation of the effect on the target variable in the model.

During the Data Understanding phase, no incorrect data was found in any of the attributes that require correction or removal from consideration.

Both datasets have different encoding of values for the same attributes. One such attribute reflects the level of education of the client. Figure 7 shows the values of the education attribute for both datasets.

Fig. 7. Education attribute values for two bank marketing datasets

To convert values to one similar type, a new encoding was created, where categorical values are immediately converted to numeric values for ease of use. A new scale of interval values was created, which was based on the number of years spent on studying. The following Table 2 shows the initial and new value encoding for the two datasets used.

C. Data Construction

In most projects, feature generation is the most important step in Data Preparation, since a well-composed feature can significantly improve the quality of the model. Data construction includes aggregation of attributes, generation of

TABLE II
VALUES ENCODING

Bank-additional-full.csv encoding	**Bank-full.csv encoding**	**New encoding**
illiterate		0
basic.4y		4
basic.6y	primary	6
basic.9y	secondary	9
high.school		11
professional.course		13
university.degree	tertiary	15
unknown	unknown	NaN

cases, conversion of data types for use in different models, normalization of attributes, and filling in missing data. Most of the adjustments in the selected datasets were made due to the type of data that the regression and classification models accept.

At this stage, additional features were created, and the values of several attributes were reconstructed. As a new column, information about the client's age group was added. This column includes divisions on age groups with an interval of 10 years each. In addition, personal information, including the type of work and the marital status, was divided into separate attributes of each value, which are filled with the binary numerical values 0 and 1 for each client. In the attributes of loan, default and housing, it was also required to change the values from categorical to numerical in the form of 0 and 1 indicating the absence or presence of criteria for potential marketing of the bank.

The attributes related to the work of the bank's marketing campaign with the client have also undergone changes. The month of the last customer contact was changed to an interval variable type instead of a categorical type in the range from 1 to 12. In addition, the results of the outcome of the bank's previous marketing campaign were divided into separate columns with binary numerical values. The target dataset variable has also been changed by data type to numeric format as 0 and 1.

Finally, attribute representing number of days that passed by after the client was last contacted from a previous campaign was adjusted. Initially, in bank-additional-full.csv existed value of 999, meaning that client was not previously contacted. Since same meaning was carried by value -1 in bank-full.csv, it was decided to replace current value with -1 in order to be consistent for both data files. As for the balance of customer, the missing values for some of the customers were filled using linear interpolation method.

D. Data Integration

The final data for this project must be loaded from several files and their integration is required to prepare the training sample. Integration refers to both "horizontal" joining (merging) and "vertical" joining (appending), as well as data aggregation. The output is a unified analytical table suitable for delivery to analytical software as a training sample.

Before data was integrated from two source files, the common attributes were constructed for data integrity and all customer information was completed before the final connection of the datasets. Since the datasets duplicated some cases, we selected only unique clients and their information for further use, thereby adding about 1300 useful recordings from the additional file to the main one. After finishing data manipulation, the size of the final data reached to 46532 cases and the final number of attributes is 40 with 1 target variable.

E. Data Formatting

At the last stage of Data Preparation, it is required to bring the data to a format suitable for modeling. Data formatting is mainly carried out only for those algorithms that work with a specific data format.

Since this project touches upon the tasks of classifying the client base for a marketing campaign, data formatting is not required, because this type of machine learning does not require a specific predefined format. All client strings are in the same random order as they were originally written from the data file, and the order of the attributes is also preserved in their original form.

V. MODELING

A. Modeling Technique

After the data was cleared and preprocessed, it was decided to try several methods for classification, namely, Logistic Regression, Decision Tree, Random Forest and SVM. These are a classical Machine Learning techniques, often used for classification.

B. Test Design

The data was split to training and testing sets with 80:20 ratio. Training data was used for training a classifier and testing data was used to assess model's ability to predict the label of the data it have not seen before.

Since the objective of the modeling step is a classification problem, the performance of each model was measured by calculating accuracy of the training data and testing data. As it was specified during determination of data mining goals, the main target was to exceed the threshold of 90% accuracy, while also monitoring the difference between the performance of the classifier on training and testing data in order to avoid overfitting or underfitting.

TABLE III
PERFORMANCE OF THE MODELS

Method	Parameters / Specifications	Training Accuracy	Testing Accuracy
Logistic Regression	dropped age and month	88.30%	87.60%
Decision Tree	max_depth: 5, criterion: entropy	88.36%	88.37%
Random Forest	n_estimators: 67, max_depth: 15, criterion: entropy	95.18$	90.06%
SVM	Kernel: linear, C: 0.1	88.93%	88.85%

C. Build Model

While working on Logistic Regression model, all insignificant attributes were dropped one by one (using significance level 5%). Alternating parameters for Decision Tree included max_depth (1-50) and criterion (entropy or gini). Random Forest model was tested under different number of trees (2-100), depths (1-50) and criteria (entropy or gini. Finally, SVM ran under a range of parameter C: 0.0001 to 100000000 (multiplying by 10 each time) and three kernels: linear, polynomial and RBF.

D. Assess Model

Table 3 lists the four used methods with their best-working parameters.

VI. EVALUATION

A. Evaluate Results

Out of the four tested methods, only one came up with the bare minimum accuracy, determined during Business Understanding phase (90% accuracy) - Random forest. The technique gives 95.18% accuracy on the training set and 90.06% on the testing set. Even though the accuracy on the training set is higher, the difference is not significant, and therefore the result is not considered to be overfitting. Therefore, it was decided to choose Random Forest as a main model for project deployment.

B. Review Process

Overall, after modeling the data by testing the four Machine Learning classification algorithms: Logistic Regression, Decision Tree, Random Forest and SVM, the conditions maximizing each model's performance was identified. For Logistic Regression the variables age and month variables were considered as unimportant for the fit, for Decision Tree and random Forest the best-performing parameters were chosen, and SVM model returned the highest accuracy with linear kernel and parameter C equal to 0.1. As the most successful model, Random Forest was chosen to be the main classifier in this project.

The limitations of the modeling stage include the constraints in the capabilities of the machine, which was running the experiments. For example, it was not optimal to perform grid search for the SVM model, because the method itself a very operationally rich process which takes a lot of time to finish. Additionally, due to time constraints, it was not possible to test more advanced methods such as artificial neural networks.

C. Determine Next Steps

As the primary goal of the Data Mining step was achieved, the modeling phase can be terminated. Taking into account the results from the models, models complexity and their runtime - the Random Forest was selected as final model for the marketing campaign. The accuracy of the model met the expectation of threshold that was set at the beginning of the project. The deployment of the project follows.

VII. DEPLOYMENT

A. Plan Deployment

From the Evaluation stage, it was decided to deploy the less costly in time and more efficient and stable solution for client segmentation- Random Forest Model. All team members of the project were communicating efficiently and achieved all goals that were set at the initialization of project stage. The project deployment date will be announces after meeting with the representative of the Board of Directors of the Bee Bank and then will move to monitoring and maintenance part.

B. Plan Monitoring and Maintenance

Maintenance of the segmentation process will be provided by Data Science Team. The Data Science Team will continue to analyze the current customer data in batches and provide the results of modeling to the Marketing Team. The efficiency of the marketing campaign with segmentation will be monitored by both Marketing and Customer Relationship Team. The whole process will also be under CEO monitoring.

C. Produce Final Report and Review Project

This document is the Final Report of the whole CRISP-DM process of the project as well as the final documentation of the marketing campaign by bank. The review project will be completed on the meeting with Board of Directors, Marketing Team, Data Science Team and CEO of the Bee Bank. The completed project business goals to apply new marketing strategy for the new bank product - term deposits by using customer segmentation; while data goals was to perform classification of the customer acceptance of the offer based on the collected data.

VIII. CONCLUSION

The CRISP-DM methodology is considered as the most effective strategy for leading Data Science projects. This paper provides all executed stages of CRISP-DM for the local bank. " The Bee Bank" is added new bank product "term-deposits" and will introduce it to its customers. The Board of Directors of the Bank and CEO would like to introduce the targeting marketing campaign to maintain current bank reputation and customer loyalty as well as increase its operational profit.All six stages of CRISP-DM: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment- were discussed in the separated sections. The last step of the CRISP-DM process is to get the review from the Board of Directors.

REFERENCES

[1] Chen J. ”Term Deposit Definition.”, Investopedia September 2020. Available: investopedia.com

[2] Kagan J., “ Loan Definition.“, Investopedia, October 2020. Available: investopedia.com

[3] ”89% of Businesses Make the Same Email Marketing Mistake”, SuperOffice, October 2020. Available:

[4] 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 superoffice.com