Capstone Project – Milestone Report

Table of Contents

Problem

The first capstone project will focus on mortgage bank loan analytics. Forecasting revenue, year-over-year performance, predictive modeling to identify Types of Loans (FHA or Conventional, Fix or ARM), expected loan amount based on the other features input, expected property types ect..

Client

The client could be any one who mortgage loan application, employee performance analysis, forecasting revenue, including the Federal government, retail and mortgage banks, mortgage brokers, real estate professional, who want to sell buy and sell properties with a mortgage, and others.

Dataset

I will get the data from the pipeline data from mortgage bank. Data sets contain data from October 2014 to December 2018. DataSets contains many columns but we will be using on 12 important features for out project. The Data Preparation stage contains three elements: data pre-processing, feature engineering and feature selection. In the data pre-processing we pre-process the data so that it contains the right format for our models. In the feature engineering we create a list of features, using the predictors. In the feature selection we select a subset of features that are useful for our model.

There are different techniques to encode the categorical features to numeric quantities.
The techniques are as following:

- Replacing values
- Encoding labels
- One-Hot encoding
- Miscellaneous features

Outlines

- Import the required libraries
- Download the data from the Loan Application Pipeline website
- Analyze the dataset and check if there are missing values
- Check the Data types if there is a non-numerical value.
- Check the skewness of the data distribution and scale the data if its skewed.
- Explore the target with different features to determine which feature is more important and decisive.
- The scatter plots and heat map could give as a good hint how the features are related.
- We evaluate the dataset with some baseline algorithms with the non-scaled data.
- After checking the scores of the algorithms, we will proceed to evaluate the same algorithms with a scaled data.

- We tune the algorithms we chose based on the scores we get.
- We prepare a pipeline to show the reproducibility of the above process.
- Present the model.

Business Understanding

The background information is collected on the domain. A theoretical framework is developed by conducting a literature review which contains an overview of the related research in this research area, and an overview of the concepts in predictive analytics and the different models that are feasible for our Capstone project. For the domain analysis, background information on the mortgage application process and the mortgage domain is collected in order to get a better understanding of the different topics.

Data Understanding

In the Data Understanding stage the raw data is collected from the database using an SQL query and its characteristics and distributions are explored. Event logs are kept in the database and can be used for predictive modeling. Once the data is collected and saved as **csv** format, the data is explored and visualizations are made of the different variables to get an understanding of the data. With these visualizations we can already see some of the relationships in the data and identify possible features. The data exploration activity is important for becoming familiar with the data and identifying data quality problems.

Data Preparation

The goal of the Data Preparation stage is to transform and enrich the dataset so that it can be fed into the models. After the data is collected and explored, it can be pre-processed so that it can be used directly in our predictive model. With the pre-processed data one can perform feature engineering. Using historical data and external data, different features can be generated. After the feature engineering activity, a subset of features will be selected that provide predictive value for our models.

Predicting mortgage demand using machine learning techniques

It is difficult for the financial institutions to determine the amount of personnel needed to handle the mortgage applications coming in. There are multiple factors influencing the amount of mortgage applications, such as the mortgage interest rates, which cause the amount of mortgage applications to differ day by day. In this research we aim to provide more insight in the amount of personnel needed by developing a machine learning model that predicts the amount of mortgage applications coming in per day for the next week, using the CRISP-DM framework. After conducting a literature study and interviews, multiple features are generated using historical data from a Dutch financial institution and

external data. A number of machine learning models are developed and validated using cross-validation. The predictions of our best model differ on average mortgage applications per day compared to the actual amount of mortgage applications. A dynamic dashboard solution is proposed to visualize the predictions, in which mortgage interest rate changes can be manually entered in the dashboard, and recommendations have been given for the deployment of the model at the financial institutions.
Methodology

Historical data and publicly available data were used as input for our predictive model, and five machine learning techniques (Decision Tree, Random Forest, Support Vector Machines, Support Vector Regression and KNN) were applied to create the predictions. The models are validated using repeated cross-validation, and evaluated using several evaluation criteria. We have also used ARIMA model, Linear Regression and Logistic Regression for predictive modeling.

The Random Forest model gave the best result on each of the four evaluation criteria used to evaluate the models. The Random Forest model is mortgage applications per day perform best, then Decision Tree model. The SVR model scored is the worse, SVM on a second place. The percent error of the Random Forest model is around of the actual amount of mortgage applications per day.

Linear Regression,

Logistic Regression

Random Forests (RF)

Support Vector Regression (SVR)

Support Vector Machine (SVM)

k-Nearest Neighbors

Decision Tree Classifier

Using Scikit-learn, optimization of decision tree classifier performed by only pre-pruning. Maximum depth of the tree can be used as a control variable for pre-pruning. In the following the example, we can plot a decision tree on the same data with max_depth=4. Other than pre-pruning parameters, We have also tried other attribute selection measure such as entropy This pruned model is less complex, explainable, and easy to understand than the previous decision tree model plot.

Mortgage interest rates Analysis

Mortgage interest rates have a significant impact on the amount of mortgage applications. If the interest rates are low, the mortgages are relatively cheaper for the borrower as they have to pay less interest, which leads to an increased amount of mortgage applications. A high mortgage interest rate means the mortgage borrower pays a high amount of interest to the lender, which makes the mortgage less attractive for the borrower.

In general, there are two types of mortgage interest rate: variable rates (**ARM**) and fixed rates. Variable interest rates are generally lower than fixed interest rates, but can change every month. Fixed interest rates are slightly higher, but are fixed for a certain period of time. A fixed interest rate is generally preferred when the mortgage interest rates are expected to rise, or when the borrower wants to know its monthly expenses upfront. A variable interest rate (**ARM**) is preferred when interest rates are expected to decrease. If a financial institution has a significantly higher interest rate than its competitors, it will generally receive fewer mortgage applications as the independent mortgage advisors will forward its customers to a different mortgage lender.

PREDICTIVE ANALYTICS

Predictive analytics is a field in data mining that encompasses different statistical and machine learning techniques that are aimed at making empirical predictions. These predictions are based on empirical data, rather than predictions that are based on theory only. In predictive analytics, several statistical and machine learning techniques can be used to create predictive models. These models are used to exploit patterns in historical data, and make use of these patterns in order to predict future events. These models can be validated using different methods to determine the quality of such a model, in order to see which model performs best.

There are generally two types of problems predictive analytics is used for: classification problems and regression problems. The main difference between these two problems is the dependent variable, the target variable that is being predicted. In classification problems, the dependent variable is categorical (e.g. credit status). In regression problems, the dependent variable is continuous (e.g. pricing).
The techniques that are used in predictive analytics to create a model depend heavily on the type of problem. For classification problems, classification techniques are used such **Random Forest and decision trees**. These techniques often consist out of one or multiple algorithms that can be used to construct a model. For decision trees, some of the algorithms are Classification and Regression Trees.

For regression problems, **regression techniques** such as multiple **linear regression, support vector machines or time series** are used. These techniques focus on providing a mathematical equation in order to represent the interdependencies between the independent variables and the dependent variable, and use these to make predictions. One of the most popular regression techniques is linear regression. When applied correctly, regression is a powerful technique to show the relationships between the independent and the dependent variables. However, linear regression requires some assumptions in the dataset. One of these assumptions is that there has to be a linear interdependency between the independent variables and the dependent variable. A pitfall of linear regression is that the regression line contains no information about the distribution of the data. It needs to be combined with a visualization of the regression line in order to draw conclusions.

It can be seen that different datasets that have the same **means, variances, correlation** and linear fit, still have a completely different distribution, even though their regression lines are the same. Hence, a regression line always needs to be combined with a visualization in order to draw conclusions about the distribution of the data.

DATA UNDERSTANDING

The Data Understanding stage has been split up in two parts: data collection and data exploration. In the data collection we will discuss characteristics of the event log data, and how the data has been extracted from the database. In the data exploration, the data and its characteristics are explored to extract useful information for our models.

DATA COLLECTION

Since we are only interested in the event log data we will only be using one of the tables. This table contains data about every mortgage application. Every action performed by the system or by a user on a mortgage application is logged, and the status before and after that specific action is logged. For our analysis we are mainly interested in the date and time at which each of the mortgage applications have entered the system.

Besides Mortgage Application DataSet, we have joined two separate (**10 Years US Treasury Rate, Home Supply Index)** with our existing Mortgage Application DataSet to enhance predictive power of our model.

DATA EXPLORATION

Since our dataset can be grouped per day to create meaningful visualizations. The dataset contains data from October 2014 until December 2018. In order to get a feel of the amount of mortgage applications per day and the distribution of the mortgage applications, different visualizations can be made using Python. Increase in mortgage applications during the last few months of each year. The amount of applications per day during these months is higher compared to the other months. This can have multiple explanations so this will have to be accounted for in the model.

The density plot shows the distribution of the amount of mortgage applications. It seems the distribution of the amount of mortgage applications is normally distributed, slightly skewed to the right with a long tail. This is due to the outliers mentioned before. The median seems to be at around mortgage applications per day.

DATA PREPARATION

The Data Preparation stage contains three elements: data pre-processing, feature engineering and feature selection. In the data pre-processing we pre-process the data so that it contains the right format for our models. In the feature engineering we create a list of features, using the predictors. In the feature selection we select a subset of features that are useful for our model.

MODELING

In the Modeling stage we discuss the activities related to the model building part of our project. A selection of five modeling techniques is made that are applicable to our capstone project. From each of these five modeling techniques, a model is built with the feature set provided earlier, and the models are validated using repeated cross-validation.

SELECTION OF MODELING TECHNIQUES

For our modeling we use a combination of predictive techniques. Multiple techniques are selected and applied on the data. For the non-linear regression techniques, we use Support Vector Regression (SVR) and Neural Networks (NN). SVR has shown to obtain excellent performances in regression and time series applications. Neural Networks are a widely used method for time series data that generally gives mixed results.

Another technique we use is Classification and Regression Trees which is a simple technique that is easy to visualize. Also two ensemble techniques are included, in order to improve the performance of the Classification and Regression Trees. These ensemble techniques are Gradient Boosting Machines (GBM) and Random Forests (RF). These techniques create a multitude of regression trees and select a combination of them in order to maximize the performance.

MODEL BUILDING

Using these five techniques (Linear Regression, Logistic Regression, SVM, SVR, Decision Tree, RF, and KNN) we can create five models. We use the list of features mentioned in section 6.3 as input for our models. A total of 12 features are included, the remaining features were excluded after performing feature selection. For each of the five models hyperparameters were tuned, using grid search. Hyperparameters are the model-specific parameters that are used for optimizing the model. They generally have to be tuned in order to optimize the model's performance, and reduce the variance and bias of the model. By training the model with different values of the size and the decay, and evaluating its performance, we can select the hyperparameters that result in the best performing model in terms of predictive power.

ARMA Model will be used for forecasting

Deliverables

Codes will be provided via Jupyter Notebook

Project power point presentation will be provided

Final Project report will be provided

All documents and codes will be shared via GitHub