Dental Magic – Predicting Churn

Data mining for business decisions

Table of Contents

1. Introduction

Employee turnover is costly and disruptive for every organization. Dental magic is like many other organizations working towards lowering their employee turnover. Churn of an organization's employee are associated with a lot of issues and expenses. It will not only include costs in terms of money, but also areas such as disruption of team dynamics, loss of productivity, damaging employer branding image and resources to train the new employee. In fact, the cost of churners is calculated to exceed 100% of the annual salary for the vacated position (Bryant & Allen, 2013). The emphasize on this project, will be to predict the voluntary churners, because they are valuable assets to the Dental magic. By decreasing the voluntary churn rate by only a few percentages, the annual savings would compile to potentially a vast amount of cost saved for Dental magic.

1.1 Objectives

The data mining project revolves around increasing focus on voluntary employee attrition, that within the recent years have become more and more important for organizations. Dental magic does poses data on their employees on a decent level, which could hold valuable information about their employees and their behavior. Besides that, improvement in data mining techniques allows for more data to be processed, and the increased organizational attention towards employee attrition have made the area of particular interest. Furthermore, a vast amount of scientific papers is focusing on improving the retention rate of employees and can create a close tie for Dental magic to increase their employee retention rate. ***Hence, the data mining objective for Dental Magic is to classify potentially employee churners that are most inclined to voluntarily churn and those who are least likely to churn.***

The overall objective will in turn, have a data mining objective along with a business objective. The distinction between these two objectives are important for our task at hand. The data mining objective should only be interpreted as a mean to reach the business objective, by utilizing the information drained from the data. This designates, that after achieving the data mining objective, much more effort is still necessary in order to reach the overall business objective. ***Thus, the business objective for Dental magic is to identify employees prone to voluntarily churning, in order to strategically target those employees and improve the turnover rates.***

The overall aim of this project as consultants, is firstly to build a model which can detect employees most prone to voluntary termination. It should be noted, that the data mining problem is a supervised classification learning problem, and the focus is on using techniques related to logistic regression and derive easy interpretable. The assignment also cover aspect such as the most important variables, and other specific stated attributes and their effects, all with regards to churn. Lastly, this information from the model should enable insights, to make data driven decisions regarding the improvement strategies, to improve the employee turnover rate.

Data mining objective	Business objective
Build a logistic regression model capable of identifying employee churners who are most prone to termination	Increase the retention rate of potentially voluntary employee churners with various incitements strategies trough targeted selection

Figure 1

2. Method

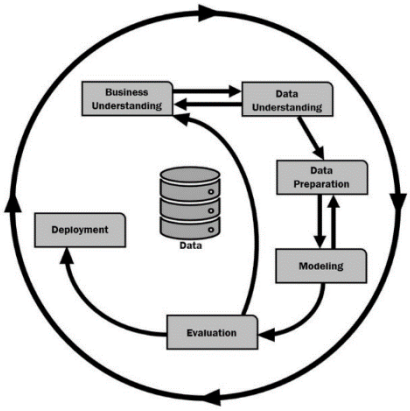


Figure 2

When starting a data mining project, the methodology is of high importance. This report will take its origin from the well acknowledged CRISP-DM framework. CRISP-DM, short for Cross Industry Standard Process for Data Mining is a very convenient tool for end-to-end projects within data mining projects. CRISP-DM as a framework offers immense flexibility and application neutrality. It is also accepting that a project starts with a lot of assumptions, and as the user go through the steps, you acquire a deeper understanding of the phenomena. The empirical knowledge learned from previous cycles can then feed into the following cycles to support continuous development and learning in the process. As depicted in figure 2, the process starts with an inclusive understanding of the business, the data and the phenomena to investigate and hence are the connection between data mining and the business impact for Dental magic. Later it will proceed to the application of the data aspect with data transformation, model building and assessment of the models. It is usually within this loop that most of the iterations takes place. Finally, an evaluation of the final models is conducted, business implications are assessed and a suggestion on deployment is made.

2.1. Data

The data used in this data mining projects consist of 310 observations, holding information about employee's attributes at Dental magic. The dataset contains 10 ID variables, 4 date variables and 21 various variables including numeric-, categorical- and text values. As the case does not provide any information about the sample, it is assumed that the data provided represents the entire population, which makes the derived results transferable and applicable to Dental magic. The transformation of the variables will be elaborated further in section below.

Nb. All text variables have been converted to factors, also included is SpecialProjectsCount and EmpSatisfaction.

Variable transformation

The target variable in the dataset is "vol_term" which is a qualitative factor, consisting of binary output in 2 levels. The variable did not originally exist in the dataset, and was thereby extracted from the EmploymentStatus variable containing the category "voluntarily terminated". (yes) describes if the employee has voluntary been terminated and (no) if the employee has not voluntary been terminated and hence still is employed at the company or fired. In this report the (yes) voluntary terminated employees is of interest.

The date level variables contained information about the date of termination, date of hire, date of last performance review and date of birth. These variables have been transformed, in order to maximize the information from the variables. The date of termination and date of hire has been transformed to a new variable called yearsemp (years employed). It is based on the assumption that seniority has an effect on voluntary termination. Furthermore, the date of birth variable is very inconsistent where dates were denoted 02-05-1956 or 02/05/1956. The dates were corrected to a uniform date, and the age at hiring

point was calculated as a numeric value. The intuition behind the transformation, is that the variable age at hiring point potentially could contain valuable information about whether an employee is about to voluntarily terminate or not. The three original variables where afterwards excluded, despite acknowledging there might be seasonal effects for when voluntarily termination occurs, it is argued that the seasonality will not affect whether voluntarily termination will happen or not.

Please note, if the employee still is employed at Dental magic (NA in date of termination) the years employment cut-off date is 12/6/2020.

Derived from	Transformed variable
Employmentstatus (voluntarily terminated)	Vol_term (target variable)
Date of birth until Date of hire	age (at hiring)
Date of hire until Date of termination	yearsemp
Date of last performance review	daycountpr

Figure 3

Because the extracted target variable only consists of voluntary termination, the observations with "termination with a cause" has been excluded, because they are not of interest. This implies that the dataset only consists of 295 observations.

After transformation, the data consist of 295 observations with 10 ID variables, 4 date variables and 24 various variables numeric and categorical variables.

2.2. Target variable

The target variable is highly unbalanced, where 70.2% of the employees has not churned within the investigated period, while 29.8% of the employees has churned voluntarily. This indicates a heavily skewed target variable, because, as expected most of the employees have not voluntarily churned and are still employed at Dental magic. One could argue, that this would be insufficient to train the model on unbalanced data, and lower performance on predicting voluntarily churners. However, this distribution represents reality and the skewness of the target variable will be kept in mind when assessing the evaluation criterions and thereby avoid over predicting.

2.3. Input variables

As stated earlier, the provided dataset consisted of various variables with different characteristics. As shown in figure 4, the variables were excluded before preprocessing the data. Employee names, ID and manager ID where all excluded, since the algorithm will be confused with the numeric ID's and the name of the employee will not provide any valuable information to the model. Furthermore, all the other ID variables are removed, because they will be encoded manually in the caret package to ensure proper encoding of the categorical variables. Zip code is also excluded from the data, and this is based on having multiple variables, which identifies the employees home address, like zip code and state which carries similar information. One could argue, that the zip code could provide more granular and valuable information to the model, but the vast amount of levels requires thorough transformation and the granularity could potentially harm the algorithm.

It is also argued, that the variable HispanicLatino are inconsistent. This is based on a thorough examination of the variable. The inconsistency is expressed in for instance observation-point number 255 as "not Hispanic or latino", while the variable RaceDesc did state the race the person identifies himself with as Hispanic. It should also be noted, that the race description variable will provide most of the information contained in the variable HispanicLatino.

The termd and Term_reason variable has been removed because the aim of the report is to focus on voluntary termination and ***not*** termination in general. Term_reason is excluded because if you have term_Reason in the models, the algorithm will exactly know if you are terminated or not (accuracy 100%), which intuitively makes sense. This is also the case on daycountpr and yearsemp which is highly correlated with termination, and furthermore implies a crude assumption by taking cut-off as today, because the spread from termination to still active become very wide. Lastly, the dropped EmploymentStatus and time variables was elaborated in section 2.1.

Excluded variables	Reasoning
Employee_Name	No information added to the model
EmpID	Database PK, no information in the variable
ManagerID	Database PK, no information in the variable
MarriedID	Database PK, will be coded manually from MaritalDesc
MartialStatusID	Database PK, will be coded manually from MaritalDesc
GenderID	Database PK, will be coded manually from Sex
EmpStatusID	Database PK, will be coded manually from EmploymentStatus
DeptID	Database PK, will be coded manually from Department
PerfScoreID	Database PK, will be coded manually from PerformanceScore
FromDiversityJobFairID	Database PK, will be coded manually from RecruitmentSource
PositionID	Database PK, will be coded manually from Position
Zip	Overlap of information from state
HispanicLatino	Inconsistent category
Termd	Information needed contained in target variable
Term_reason	Will introduce bias in the model
Yearsemp (uddyb omkring den)	Too much correlation with the target variable, and strong assumption
Daycountpr	Too much correlation with the target variable, and strong assumption
EmploymentStatus	Defined in the target variable - redundant
DaysLatelast30	No information in the variable
LastPerformanceReview	Recalculate to daycountpr (days sine last performance review)
Date of hire	Information contained in yearsemp and age
DOB	Recalculated to age (age at hiring)
Date of termination	Recalculated to yearsemp (seniority)

Figure 4

The final predictors in the data can be seen in appendix 1.

2.4. Data pre-processing

In this section, the necessary preprocessing steps will be outlined. The transformation includes the sampling and partitioning scheme of the data, treatment of outliers and missing values and various relevant feature engineering techniques.

Beforehand it should be noted that the dataset is initially stratified and split into training and test set to avoid data leakage and artificially increasing the test accuracy.

2.4.1 Sampling and data partitioning

In order to split and sample our dataset, certain considerations must be addressed. Several ways of sampling schemes could be applied, in order to obtain a representative subset of the population. The most

common strategy used, is a random sampling strategy which implies that each observation has equal chance to be included in the sample. But random sampling does not ensure that the original data is represented in the sample. In this case, random sampling could become troublesome, because as earlier mentioned, we have a very skewed target variable. So, if we are unfortunate only to have very few voluntary terminations observations in the test set, then our model will have a very poor approximation of out of sample accuracy. Likewise, if we by accident only have very few voluntary termination in the training data, then we will overfit because there will not be enough datapoints to capture the underlying relationship of the target variable.

Stratifying sampling can overcome this limitation, by dividing the observations into homogenous groups (stratas), and from each group (stratum) a sample is drawn to approximate the original structure of the dataset. This ensures higher statistical precision, and also requires much less data to achieve same accuracy as random sampling, and thus also eases the computational requirements (Mahmud, Huang, Salloum, Emara, & Sadatdiynov, 2020).

As formerly described, the partitioning was done prior to pre-processing in order to avoid data leakage. The data is partitioned into training and test set. This is done in order to create an algorithm that fits the past data well and have a high generalizability on future data. The training set serves as data for building and fitting the model, while the test set is for assessing and selecting models and prevent overfitting. You could also argue, that a validation set should be incorporated in order to attain an honest assessment of how well the final chosen model performs and confirming that the result is not being biased.

In the partitioning stage a certain trade-off must be addressed. The division of observations to the respective sets creates this trade-off. The trade-off consists of the value to train models on large amounts of observations, while it is still very important in this case to validate the findings on many observations to ensure high generalizability. In this case, the data mining objective is to classify voluntary terminations, thus the prediction of potentially new churners is of high importance and due to relatively small dataset a three-set partitioning is not considered appropriate. To ensure enough observation to assess and rank the models upon, the data was split into a 70% train and 30% test set which is deemed acceptable in this project.

2.4.2 Treatment of outliers and missing values

In figure 4, the distribution of missing values across all variables are visualized with a missing plot, which is an efficient way to visualize NA's for small- to medium-sized data sets.

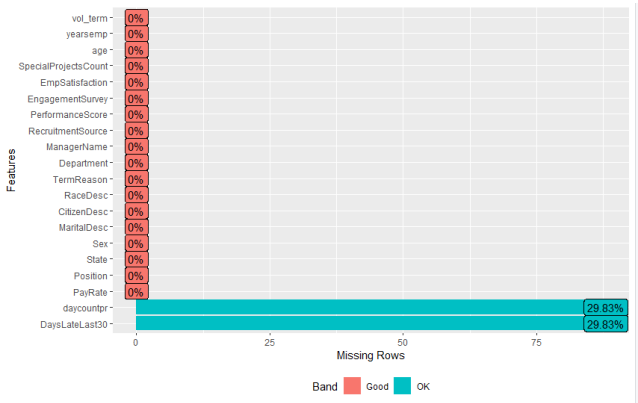


Figure 4

Treating missing values (NA), implies considerations regarding the context and examine the underlying reason for the missing answer. Some of the reason could be information missingness or missingness at random. Information missingness, should be replaced by a category to preserve the information in the missingness. A category “unknown” for instance, would be suitable for such matter. Missingness at random should be either deleted or imputed, because they do not provide any information. If the NA’s are deleted, the dataset will potentially lose some of its signal, due to the lost observation points or variable. But when imputing the NA’s, they have the risk of following sorts of patterns. These patterns will introduce bias in the model, and thereby pick up the replacement technique and not the actual data, which will lead to poorer prediction accuracy on new data. In this case, the missing values corresponds to the terminated employees. Therefore, it is argued to further exclude the 2 variables because they do not contribute with information about the target variable.

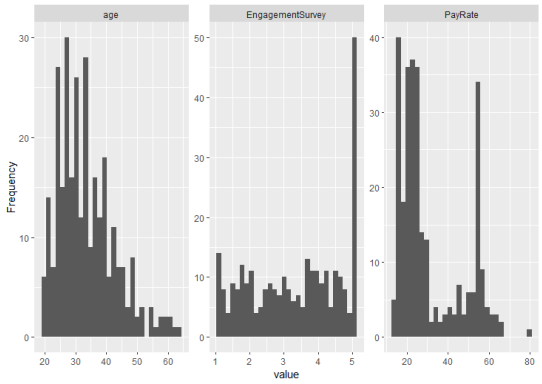


Figure 5

As depicted in figure 5, all the numeric variables after factorizing Empsatisfaction and specialprojectcount in the dataset are visualized. The variables show no sign of outliers. However, the variables PayRate and age are highly left skewed, while EngagementSurvey are highly right skewed. This indicates, that the variables are clearly breaking the assumption of normal distribution, which will be further discussed in the next section.

2.4.3 Variable transformations and derived features

In order to strengthen the models, thorough preprocessing must be applied to achieve the highest accuracy possible. This subsection will focus on transformed and derived features in the feature engineering phase.

Filtering

Feature filtering is performed on the input variables, in order to remove variables with zero or near zero variance (information). Removing non-informative features, has several advantages including that you will end up with a more parsimonious and interpretable model. You will also avoid hurting algorithms such as the lasso method, and the computational power is also significantly increased. In this case Citizen Description, was detected as a non-informative variable and was excluded in the recipe.

Numeric feature engineering

As described in section 2.4.2, the numeric values are heavily skewed. Numeric variables with skewness, outliers or wide ranges of magnitudes can be harmful to certain algorithms counting for instance regularized regression and GLMs. This can be dealt with, by normalizing and standardizing. Parametric model will benefit from minimizing skewness, the Box-Cox (when feature values are strictly positive) or Yeo-Johnson (when feature values are not strictly positive) are valuable in that matter.

So, to even out the distribution, the Yeo-Johnson transformation was carried out, however when it was tested on the model, it yielded no improvement in performance. All the numeric variables were heavily skewed beyond a point, where Yeo-Johnson could normalize them. This has been accounted for, by creating binary variables which are proved less prone to skewness and work well with elastic net and logistic regression. For instance, age was discretized into above or below 35 years, while payrate was discretized into above and below average, and Engagement survey into above or below 3. This further implies, that standardizing or centering and scaling was not necessary.

All discretization's have been carried out in the recipe, ensuring minimal data leakage. However, this is some strong assumptions and could potentially remove some signal from the variables, but due to the scope and time limit of the assignment, this is left for future research to further examine.

Categorical feature engineering
In the dataset, lots of variables had categories contains many unique values but only a few observations in each, these variables have been collapsed into fewer categories, but it was a trade-off between loss in model performance versus more parsimonious model. So various categories were tried lumped, to achieve a more parsimonious model versus the loss in performance. The final categories lumped is stated below:

The state column was transformed into a binary factor with 2 levels, containing state "MA" or "other". This is carried out due to the unbalanced ordinal category.

Position is lumped into categories with a threshold of 5%, of the training data. This is done in order to minimize the number of categories to avoid overfitting. The following categories are included in the position – Production technician I and II, Area Sales manager and other.

Race description is also lumped into categories with a threshold of 5%. The following categories are included in the position – Black or African American, white, two or more races and "other".

Recruitment Source is also lumped into categories with a threshold of 5%. The following categories are included in the position – Billboard, Diversity Job Fair, Employee referral, MBTA, Monster.com, Newpaper Magazine, pay per click Google, professional society, search engine Google/Bing/Yahoo, website banner and "other".

Lastly, PerformanceScore is lumped into categories with a threshold of 8% of the training. The following categories are included in the position – Fully meets, Exceeds and "below average".

One-hot & dummy encoding
In order to transform the nominal variables in the dataset into numeric representations, one-hot or dummy encoding are preferred. One-hot encoding which is the most popular approach, represents the variables with a Boolean value. One-hot encoding will typically result in sparse data, which is efficient for the algorithms to run. But one of the drawbacks of One-hot is that it adds a significantly increase in dimensionality of our data, and in larger dataset with many variables ordinal encoding would be more beneficial. But this is not the case in this context, and therefore position, state, sex, maritaldescription, racedescription, department, managername, recruitmensource and performancescore was one-hot encoded and dummy encoded seperatly, however the one-hot yielded the best results, thus one-hot is incorporated in the recipe.

Label encoding
During the project, ordinal label encoding was carried out in the recipe on the variable Employee satisfaction and Performance score. This could potentially be a strong assumption to assume that the variable has some ordering between the categories, and the conversion would be in levelled numeric order. Slightly counter intuitive, the result yielded worse model performance, and the variables were afterwards one-hot encoded.

Summary
To ensure no data leakage, a blueprint of feature engineering tasks has been carried out with the recipe function in the Caret package in R. This minimizes the risk of data leakage, by isolating each resampling iteration and by doing that, we can use each resample for test and training set because they are designed to act isolated. To sum up, the following feature engineering task has been carried out (appendix 2) on both the training and test set (in sequentially order).

- Filtering out zero or near zero variance features (all predictors)
- Mutate numeric to categorical factors (Payrate, age and engagement survey)
- Lumping (Position, Race Description, Performance Score and Recruitment Source)
- One-hot encoding (categorical features)

Iterations of the initial preprocessing
During the iterations of the recipe, the state variable was transformed into a binary variable that identified if the employee was from MA or not, and thus arguing that all other states does not hold critical new information. Despite, acknowledging that other states than MA, might have more impact on voluntarily termination, it is believed that it still provides valuable information to see if the employees are from other states than MA, which assumingly is where Dental Magic have HQ. If time allowed for more iterations, it is recommended to test whether including multiple states would provide more information to the model without overfitting.

3. Results

In this section, logistic regression models with or without shrinkage will be introduced and compared based on different evaluation criterions. These criterions will measure their ability to predict employees that churned, and the fraction of actual churners it predicts.

3.1 Candidate models

In order to select the right models with regards to the data-mining objective, which is predicting voluntarily termination from employees. Logistic regression with and without shrinkage is employed. Logistic regression are methods, considered having high bias, and thereby is much more inflexible than for instance Support Vector Machines. Furthermore, Logistic regression is a parametric model, and some of the limitations is that it requires observations are independent of each other. Moreover, the logistic regression is not very good at handling many predictors for small samples, which in turn could cause problems in this particular setting with originally 38 variables on 310 observations.

The candidate models and their tuning parameters are illustrated in figure 6. The algorithm and their characteristics are introduced in their respective sections below.

Logistic regression
Logistic regression is known as a powerful tool for classification tasks, with binary response variables, furthermore it uses the method of maximum likelihood. The intuition behind using maximum likelihood estimation, is rather than modeling the response variable directly, logistic regression models the probability that the target variable belongs to a specific category. The threshold for classification is by default 0.5, meaning that estimated values below 0.5 is classified as a non-voluntarily churners and above is classified as voluntarily churners. This threshold could be adjusted according to specific goals of predicting. Furthermore, the objective in the assignment express inference as a high priority, logistic regression is very useful for interpretability and making inference. Logistic regression is widely used by researchers, and usually yields solid results when considering the principle of parsimony.

Shrinkage
A lot of coefficients can make the model unstable, so by picking the coefficients predicting the best, you will end up with a more parsimonious and stable model. As an alternative to the logistic regression, we can fit different models also containing all p predictors but shrinks the coefficient estimates towards zero. By Shrinking the coefficient estimates, its possible to significantly reduce the variance, which is highly appropriate considering the bias-variance trade-off. The two best known techniques for this are ridge regression or the lasso. The lasso enforces a constraint on the sum of the absolute values of the parameters, where the sum has a defined constant (constraint) as an upper limit. This constraint imposes the coefficient for some variables to shrink towards zero. Ridge regression works by includes all the predictor variables in final model. The penalty will shrink all coefficients towards zero, but it will not set any of them to exactly zero unless lambda is infinite. This is theoretically not a problem for prediction accuracy, but it can create challenges for interpretation in settings where the number of variables is large.

Elastic net regularization is a compromise between ridge regression and the lasso. Elastic net will average highly correlated features instead of removing some of them, and yet still encourage a sparse solution with regards to the coefficients and their averaged features. This indicates that we avoid group selection like the lasso tends to do, and still achieving sparsity like ridge regression does not provide.

If the elastic net chooses an alpha value = 0 then ridge regression is applied, and if the alpha value = 1 then a lasso model is applied. The lambda value is the shrinkage applied.

In order to achieve a more parsimonious model the one standard error rule is used; this will be elaborated further in section 3.2.

Models	Tuning parameters
Logistic regression	None
Elastic Net	α, β
Elastic Net (oneSE)	α, β

Figure 6

3.2 Model selection approach
In the model selection approach, it is vital to have the overall objective in mind, to ensure the best fit between algorithms and the problem to be solved. Secondly, it is also of vast importance to not only rely on a single measure for evaluation, but rather have several criterions to get a broader perspective and a more holistic view of the model performance. Because the purpose of the study is to classify voluntarily

terminations, the most central metric is derived from the confusion matrix and consist of precision, recall and F1 score. All of these measures all are metrics that takes skewness of the target variable into account. Precision and recall rate are essential in these matters, since it is the direct link the data mining and business objective, because it explicitly measures how good the model is, at detecting voluntarily terminations and the fraction of truly churners. And thus, it is dominant importance that it can capture the churners and correctly, because otherwise it would induce unwanted cost to Dental Magic. To equal out the goals, the F1 score will balance these two measures.

As introduced earlier, the metrics above will be assessed with cross validation on the training set, in order to build and fit the most suitable model. Lastly the models will be trained on the entire dataset and tested on the test set, to ensure that the model does not only fit its own train data, and after deployment the prediction power will drop significantly.

Bias-variance trade-off

In order to assess the various model, the bias variance trade-off must be addressed. Bias-variance trade-off is of enormously importance within machine learning. The paradigm revolves around having a low or high error rate due to bias or variance. Bias refers to the error that is introduced by approximating a real-life problem and assessing how good the model is at capturing the underlying structure of the data. The variance reflects how much the estimate varies around the average. More flexible methods usually have higher variance which can capture non-linear trends. But too high variance then small changes in the training data would result in large changes in the predicted values, which is undesired. This trade-off is often considered in relation to model selection. In this context, the focus of the project is interpretability and generalization, and it should therefore be reflected in the models having less flexibility to capture non-linear trends but with higher interpretability.

Cross validation

Cross validation is a resampling method, and it was used in order estimate the in-sample goodness of fit, on the train error on the logistic regression models. K-fold CV is a technique that estimate the error rate by dividing the dataset into randomly equally sized subsets k. Next the subset (k-1) are combined into one training set while the remaining fold is used as a holdout set, then it is applying the statistical-method to those holdout observations and repeat the process k times and the estimates from the prediction is averaged out and hence gives a more stable performance measure. It also guided the selection of appropriate flexibility level in the shrinkage method. The latter is known as model selection while the evaluation of performance is known as model assessment. Empirically a 5- or 10-fold cross validation have been shown to yield test error rate estimates that suffer neither from excessively high bias nor from variance. K-fold CV are aiming towards lowering the expected loss as much as possible. This means that it will minimize how much the developed model predicts incorrect. Some of the drawbacks of k-fold is that the training set is not as big as the original training set which implies that the estimates of prediction error will be biased upwards. K=10 with 5 repeats is deemed a good bias variance trade-off in this case.

Please note, that the LOOCV was not considered due to the opportunity cost by taking computational efficiency into consideration. LOOCV was not feasible in this case, even though it might be theoretically the correct thing to do.

1SE RULE

The one standard error rules implies, that the estimate of the errors have an inherent variability, and the performance between minimum error compared to one standard error is not much of a different. Moreover, it is assumed that tuning parameters overfit, and thereby allowing a less complex model within

one standard error to be a more optimal model with regards to overfitting and simplicity in these circumstances.

Precision

The precision measure is the ratio of observations that is correctly classified as voluntarily churners, compared to the total number of voluntarily churners classified. The objective is to maximize this measure, meaning that a precision of 1 is desired. Furthermore, precision assumes that true classification outweighs the cost of false positive (non-churner classified as churner). In this context, precision is highly relevant because it is argued that it is more important to actually capture all churners because they are expensive to lose, than to capture non-churner as churners which only will induce small cost associated with various HR-initiatives.

$$Precision = TP/TP + FP$$

Recall

The recall measure is the ratio of the truly voluntarily churners, compared against the truly positive churners and the false negatives ones. In other words, it classifies the number of true positive churners against the positive churners calculated by the algorithm. As before, the objective here is to achieve a recall near 1.

$$Recall = TP/TP + FN$$

F-measure

The F-measure is reflected in the balanced mean of precision and recall, and it is often used when the classes of predicting are imbalanced. The balanced mean will penalize extreme values.

$$F1 = 2/((1/precision) + (1/recall))$$

3.3 Final model

As seen in appendix 3, the in-sample goodness of fit assessment of the three models is based on the training data. The logistic regression seems to outperform shrinkage with and without one standard error, on the ability to predict churners. However, the model will not be selected based on the training error, but logistic regression still indicates the best solution even if it is a bit more complex. As seen in the tuning parameter and alpha value around 0.4, which indicates a more weight on the lasso technique. This will in turn, makes the selection path a bit more unstable which is also reflected in the error rate. If it was opposite and more weight was put on ridge regression, we would encounter more stability in the model but less feature selection. All three models are held against the test sample, and the results are depicted in Figure 7, which shows the accuracy of out of sample predictions, where the target variable is evaluated on the prediction accuracy on the independent test sample. The models will be evaluated based on the confusion matrix and the accompanied metrics below.

	Confusion matrix	Precision	Recall	F-measure	Accuracy
Logistic regression	**Observed values** Y N Y 17 10 N 9 52	62.96%	65.38%	64.15%	78.41%
Elastic net	**Observed values** Y N Y 8 9 N 18 53	47%	30.8%	37.21%	68.18%
Elastic net one SE	**Observed values** Y N Y 5 7 N 21 55	41.66%	19.3%	26.37%	69.32%

Figure 7

Above the in-sample goodness of fit was assessed, where we now examine the out-of-sample performance.

As depicted in figure 7, the logistic regression is clearly outperforming the other model on all parameters. Due to the importance of the precision and recall measure, the logistic regression is chosen as the best model to proceed with and will be further discussed next.

3.3.1 Overall predictive accuracy

The chosen model - logistic model, do only achieve an accuracy on 78.41% which is not very impressive, when the benchmark of just predicting the most frequent class (non-churners) is 70.2%. But the accuracy is not really of interest in this study. The precision and recall metric are more interesting for the scope of this assignment. The precision are the percent of the employees that the model predicted to churn, against who actually churned, and the model predicted 62.96% of the predicted churners correctly. Furthermore, recall is showing how many of the churners logistic regression actually predicts, and on this data, it catches 65.38%, which seems fairly good. ***Given we have a random sample of 100 employees***, the model would overall predict 30.8% of the sample as churners, and within the churner category around 62.96% of them would be correctly classified as churners– equivalent to approximately 19 employees. And theoretically if we assume the sample consisted of 100 true churners, the algorithm would detect 65.38% of those – equivalent to 65 employees. Given the context of the data and the case, the model must be assessed as a very satisfactory model. It clearly states the intentions to use data mining techniques is possible within employee churning and it potentially could save many resources for Dental magic which can allocate them strategically elsewhere.

3.3.2 Observed versus predicted target values

As seen in the confusion matrix in figure 7, the logistic regression is significantly better at classifying non-churners compared to the churners. This intuitively makes sense, because it's the majority class and the algorithm have much more data to learn the underlying structure of the non-churners. In each business case, an evaluation of the effects of different misclassification within each class should be addressed. When considering Dental magic, the scope is to capture voluntarily churners as primary target. Considering the cost of using resources on non-churners, NOT likely to churn is very difficult to calculate, whereas the cost associated with not targeting the employees prone to churning are expected to be much more damaging to Dental magic. Therefore, it is considered that false negative proposes a greater risk than false positives,

given the before stated consequences of the outcome. Hence the model chosen are not capturing all churners and are not predicting them correctly. However, it still yields better results than random classification.

4. Discussion

In section 3, the logistic regression was chosen as final model, and the model performs satisfactorily. Moreover, the results from the model needs to be further discussion to create business value for Dental magic. The last sections in the report will discuss deployment recommendations and follow-up activities.

4.1 Assessment of model performance

In the context of the variable basis and the skewed dataset, the overall balance between precision and recall of the model is decent and is expected to create some business value for Dental magic. However, the model has yet still to reach its saturation point, which allows for improvement possibilities. Further data mining might also reveal valuable information in excluded variables, or different transformation possibilities. Lastly, a thorough assessment of more flexible models would be desired, in order to see if a more flexible model is better at capturing churners correctly.

Increasing the amount of observations and other predictors about the employees, could also potentially improve the data richness, and thereby help improving the prediction whether an employee is about to voluntarily churn or not. These variables could include variables such as fixing the day late last 30 days etc.

4.2 Contribution to the solution of the business problem

Figure 8

The 10 most important variables are shown in figure 8. These variables are mostly associated with the response variable – voluntarily termination. These variables contain valuable information to the model and could potentially give important insights for Dental magic. They are all indexed, where the recruitment Source – Diversity job fair are most associated with voluntarily termination. When further extracting the coefficients (appendix 5), you can see that when recruiting from job diversity fair, the probability of voluntarily churning decreases. Furthermore, it is noted that when the employee lives in the Massachusetts state, they are more likely to churn. This intuitively makes sense, because over 85% of the observation lives in Massachusetts and therefore the amount of voluntarily churners must be assumed to be higher in this category. Next, ***the manager Peter Monroe has the most positive association with voluntarily churning of***

employees*, *which indicates that he is the manager where employees are most likely to churn. Furthermore, the other managers on the list (Kelley Spirea and Ketsia Liebig) are also positively associated with churn of the employees. ***On the other hand, Amy Dunn are the manager where the employees have the highest tendency to stay based on the coefficient and variable importance*** (appendix 5). More interesting notes derived from the plot, is that the recruitment source search engines such as Google, Yahoo and Monster.com are negatively associated with the response. These recruitment sources have lower probability of associated with churning, while professional society and employee referral are positively associated with churn. Lastly, if the employee has the race "white" will also lower the probability of churning.

It is denoted in the assignment that the effect of performance score, engagement survey, employee satisfaction survey and age of time at hiring are of particular interest for Dental magic, and the variables will be briefly discussed next.

Remarkably, when examining the effects of the engagement survey, it does not contain much information about voluntarily churning. It is seen, that an engagement scores below 3 does not have any impact on voluntarily churn, while an engagement score above 3, are more likely to churn, according to the model. But the relative size of influence on engagement score above 3 is small.

Some further interesting patterns are found in the model, regarding performance score which was encoded as a 3-category dummy variable (exceeds, fully meets and below average). It is derived from the model, that categories "exceeds" and "fully meets" are impacting churn in different ways, while below average has no impact on churn. The fully meets are negatively associated with churn, and thereby increasing the likelihood of the employee to stay. The "exceeds" category are having a positive coefficient, and therefore potentially increases the probability of churning. This could be interesting to dive into for Dental magic, because it could seem like the employees who believe they exceeds their performance, are more likely to churn. However, both categories are indexed around 10% on the variable importance plot, which indicates that they are not heavily associated with churn.

The employee satisfaction survey showed that the category 5 did not have any particular association with churn, but the satisfaction categories 2,3 and 4 were actually fairly important to the model. Category 2,3 and 4 showed less probabilities of churning. The category with employee satisfaction of 1, showed a probability which tends to favor churning, which intuitively makes sense that a low satisfaction rate is associated with voluntarily leaving the company. But it should be noted, that the size of influence on category 1 is minor.

The age at hiring point was encoded as a binary variable (above and below 35), and above 35 years are shown to have negative impact on the probability of churning, with a size of influence below average compared to all variables. According to the model, if the age at hiring point is below 35 are not saying whether you are churning or not.

4.3 Recommendations, deployment and follow-up activities Deployment

Figure 9 demonstrates the flow that will help Dental magic improve their target marketing and thereby their retention rate.

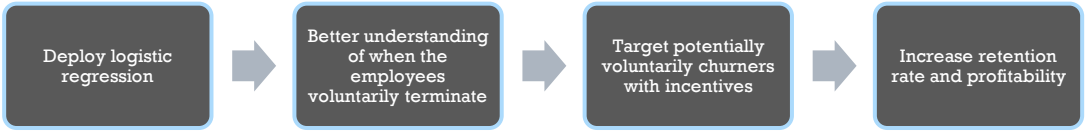


Figure 9

The business objective was to increase the retention rate of Dental magics employees, and it was the main focus for the developed algorithm to correctly classify potentially churners, and thus improve the targeted incentives. By improving churn classification, it helps Dental magic understand some of the underlying reasons for voluntarily termination and moreover, streamline and target their HR incentives effectively. The final result in the chain should increase the retention rate of the employees, which in turn will higher the profitability of Dental magic.

Taking offset from the model built, many possible initiatives can be derived, and it becomes clear that Dental magic has some areas where they possibly can counter employee churn. First and the utter most important recommendations, is to catch the "low hanging fruits" by exploiting the easiest step to improvement in the retention rate. This could be done, by knowledge sharing or creating workshops where Amy Dunn could share her experience, and some potential guidelines to follow. Furthermore, it is recommended, that the management team should consider the managers carrying a high rank, where the employees tends to churn most. This list includes managers such as Peter Monroe, Kelly Spirea and Ketsia Liebig.

Another interesting view is that when the employee has the performance score "exceeds", they have a higher probability of churning. This intuitively makes sense, because if they believe they are exceeding their performance compared to their compensation, then they become more likely to churn. This could indicate that the compensation package should be revised, or the employee should be rewarded in other ways in order to make him/her stay at Dental magic.

Besides that, considerations must also be taken when looking at the recruitment source, which takes up 5 out of 10 in the most important variables. So, this is deemed very important for employee churn. Dental magic should reflect on their recruitment sources. The model is suggesting some recruitment sources to be more potentially giving than others, with regards to less probability of the employee churning. These sources include diversity job fair, and online platforms such as Google and Monster.com which are all associated with a higher probability not voluntarily churning. This implies that Dental magic should be skeptical or more aware when using professional societies and employee referrals as recruitment outlets, because they have a higher tendency of churning afterwards.

As mentioned in the case description, some of the criteria for the final solution was to be easy intelligible. This is supported by the logistic regression being easily understood and easy to implement. Furthermore, this model provides great insights for the CEO and HR department, which could use this model to potentially target employees, more prone to churning instead of targeting them based on intuition. The model also allows for cross organizational employee retention strategy, that can both address individual employees tendency to churn, but it could also address entire departments and see if some departments are more exposed than others.

For the model to work accordingly it should be deployed properly, operate efficiently and keep extracting data from the source. This could help Dental magic in the future to capture if certain employees have characteristics of churning or capture major shifts in trends of voluntarily churning. All this needs continuous maintenance, but according to Dental magics employee list, they already have a business intelligence unit, which would be natural to incorporate in this solution. And as a final remark, by connecting the findings with one or more colleagues at the HR department, it will further enable Dental magic to pursue the business objective of improved employee turnover rate.

References

Bryant, P. C., & Allen, D. G. (2013). Compensation, Benefits and Employee Turnover: HR Strategies for Retaining Top Talent. *Compensation and benefits review*, 171-175.

Mahmud, S. M., Huang, J. Z., Salloum, S., Emara, T., & Sadatdiynov, K. (2020). A Survey of Data Partitioning and Sampling Methods to Support Big Data Analysis. *Big Data Mining and Analytics*, 85-101.

Appendix

Appendix 1

Input variables	Original state	Transformation performed	Result
Payrate	Numeric	Discretized to adjust for skewness	Binary variables
EngagementSurvey	Numeric	Discretized to adjust for skewness	Binary variables
Age	Numeric	Discretized to adjust for skewness	Binary variables
Position	Fact(32 levels)	Lumped to reduce levels and create a more parsimonious model + One hot encoding	4 Binary variables
State	Fact(28 levels)	Lumped to reduce levels and create a more parsimonious model + One hot encoding	2 Binary variables
Sex	Fact(2 levels)	One-hot encoded	2 Binary variables
MaritalDesc	Fact(5 levels)	One-hot encoded	5 Binary variables
CitizenDesc	Fact(3 levels)	Near-zero variance - omitted	N/A
RaceDesc	Fact(6 levels)	Lumped to reduce levels and create a more parsimonious model + One hot encoding	5 Binary variables
Department	Fact(6 levels)	One-hot encoded	6 Binary variables
ManagerName	Fact(21 levels)	One-hot encoded	21 Binary variables
RecruitmentScore	Fact(23 levels)	Lumped to reduce levels, thus a more parsimonious model. Creating a "other" category. (+ one-hot encoded)	11 Binary variables
PerformanceScore	Fact(4 levels)	Lumped to reduce levels and create a more parsimonious model + One hot encoding	3 Binary variables
EmpSatisfaction	Numeric	Factorized and One-hot encoded	5 Binary variables
SpecialProjectsCount	Numeric	Discretized to adjust for skewness	2 Binary variables

Appendix 2

```
Inputs:

      role #variables
   outcome          1
 predictor         15

Training data contained 207 data points and no missing data.

Operations:

Zero variance filter removed no terms [trained]
Sparse, unbalanced variable filter removed CitizenDesc [trained]
Variable mutation for PayRate [trained]
Variable mutation for State [trained]
Variable mutation for age [trained]
Variable mutation for EngagementSurvey [trained]
Factor variables from EngagementSurvey [trained]
Factor variables from age [trained]
Factor variables from PayRate [trained]
Factor variables from State [trained]
Collapsing factor levels for Position [trained]
Collapsing factor levels for RaceDesc [trained]
Collapsing factor levels for PerformanceScore [trained]
Collapsing factor levels for RecruitmentSource [trained]
Collapsing factor levels for SpecialProjectsCount [trained]
Dummy variables from PayRate, Position, State, Sex, MaritalDesc, RaceDesc, Department, ManagerName, RecruitmentSource, PerformanceS
core, EngagementSurvey, EmpSatisfaction, SpecialProjectsCount, age [trained]
> |
```

Appendix 3

Logistic regression

```
          Reference
Prediction yes  no
       yes  40  12
       no   22 133

               Accuracy : 0.8357
                 95% CI : (0.7781, 0.8835)
    No Information Rate : 0.7005
    P-Value [Acc > NIR] : 5.424e-06

                  Kappa : 0.5896

 Mcnemar's Test P-Value : 0.1227

            Sensitivity : 0.6452
            Specificity : 0.9172
         Pos Pred Value : 0.7692
         Neg Pred Value : 0.8581
             Prevalence : 0.2995
         Detection Rate : 0.1932
   Detection Prevalence : 0.2512
      Balanced Accuracy : 0.7812

       'Positive' Class : yes
```

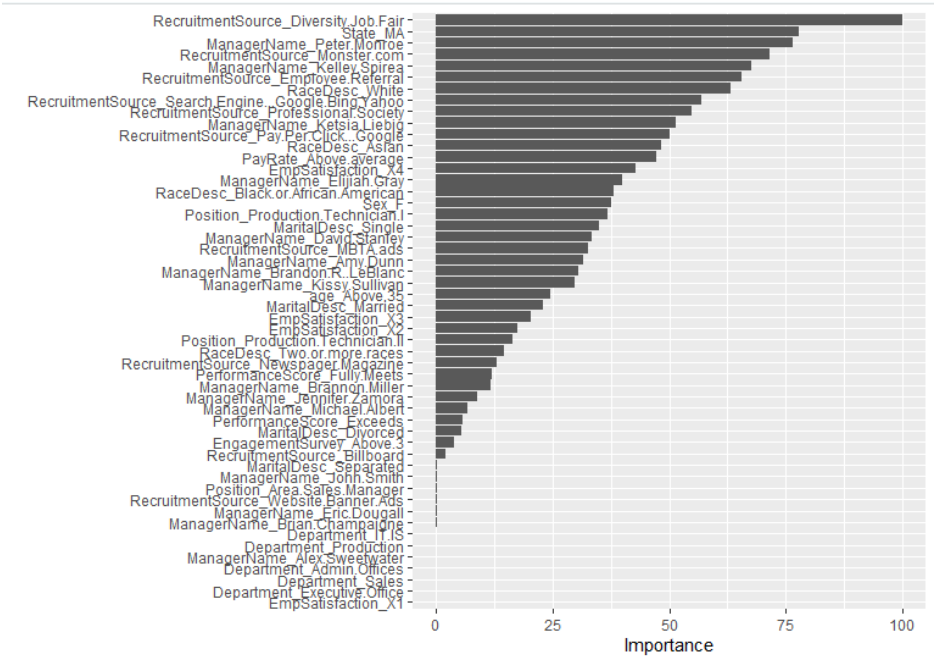
Elastic net

```
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes  31  10
       no   31 135

               Accuracy : 0.8019
                 95% CI : (0.741, 0.854)
    No Information Rate : 0.7005
    P-Value [Acc > NIR] : 0.0006391

                  Kappa : 0.4773

 Mcnemar's Test P-Value : 0.0017873

            Sensitivity : 0.5000
            Specificity : 0.9310
         Pos Pred Value : 0.7561
         Neg Pred Value : 0.8133
             Prevalence : 0.2995
         Detection Rate : 0.1498
   Detection Prevalence : 0.1981
      Balanced Accuracy : 0.7155

       'Positive' Class : yes
```

Elastic net (oneSE)

```
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes  25   5
       no   37 140

               Accuracy : 0.7971
                 95% CI : (0.7358, 0.8497)
    No Information Rate : 0.7005
    P-Value [Acc > NIR] : 0.00112

                  Kappa : 0.4327

 Mcnemar's Test P-Value : 1.724e-06

            Sensitivity : 0.4032
            Specificity : 0.9655
         Pos Pred Value : 0.8333
         Neg Pred Value : 0.7910
             Prevalence : 0.2995
         Detection Rate : 0.1208
   Detection Prevalence : 0.1449
      Balanced Accuracy : 0.6844

       'Positive' Class : yes
```

Appendix 4



Appendix 5

	x
Department_IT.IS	-21.19796433
Department_Production	-19.99250649
ManagerName_John.Smith	-19.08793705
ManagerName_Alex.Sweetwater	-18.64324759
Department_Admin.Offices	-16.00829647
Department_Sales	-12.97837699
RecruitmentSource_Diversity.Job.Fair	-4.82706904
Department_Executive.Office	-3.83945160
RaceDesc_White	-3.78621967
PayRate_Above.average	-3.65414756
RaceDesc_Asian	-3.09117923
RecruitmentSource_Monster.com	-2.78652083
RaceDesc_Black.or.African.American	-2.25730628
RecruitmentSource_Search.Engine...Google.Bing.Yahoo	-1.79467448
ManagerName_Amy.Dunn	-1.27316375
EmpSatisfaction_X4	-1.02636925
RaceDesc_Two.or.more.races	-0.92123594
EmpSatisfaction_X2	-0.89851070
age_Above.35	-0.58394222
EmpSatisfaction_X3	-0.48577698
MaritalDesc_Divorced	-0.45333146
PerformanceScore_Exceeds	-0.27185761
RecruitmentSource_Billboard	-0.07516844
EngagementSurvey_Above.3	0.07884031
ManagerName_Michael.Albert	0.26694417
PerformanceScore_Fully.Meets	0.44735641
RecruitmentSource_Newspager.Magazine	0.47430400
ManagerName_Brannon.Miller	0.50458439
Sex_F	0.81716694
ManagerName_Kissy.Sullivan	1.23658952
RecruitmentSource_MBTA.ads	1.29170950
ManagerName_David.Stanley	1.32722804
EmpSatisfaction_X1	1.39772472
Position_Production.Technician.II	1.50918835
ManagerName_Elijiah.Gray	1.65379903
MaritalDesc_Married	1.90719969
ManagerName_Ketsia.Liebig	2.24588151
RecruitmentSource_Pay.Per.Click...Google	2.45260413
RecruitmentSource_Employee.Referral	2.81678849
RecruitmentSource_Professional.Society	2.88902867
MaritalDesc_Single	2.93171607
ManagerName_Kelley.Spirea	3.18674922
Position_Production.Technician.I	3.38235491
ManagerName_Brandon.R..LeBlanc	3.44580523
ManagerName_Jennifer.Zamora	3.85462543
ManagerName_Peter.Monroe	8.13063252
State_MA	8.44456679
(Intercept)	13.60862280
Position_Area.Sales.Manager	18.25737265
RecruitmentSource_Website.Banner.Ads	18.91110094
ManagerName_Brian.Champaigne	19.23072954
MaritalDesc_Separated	22.48590925
ManagerName_Eric.Dougall	33.32936126
PayRate_Below.average	NA
Position_other	NA
State_other	NA
Sex_M	NA
MaritalDesc_Widowed	NA
RaceDesc_other	NA
Department_Software.Engineering	NA
ManagerName_Board.of.Directors	NA
ManagerName_Debra.Houlihan	NA
ManagerName_Janet.King	NA
ManagerName_Lynn.Daneault	NA
ManagerName_Simon.Roup	NA
ManagerName_Webster.Butler	NA
RecruitmentSource_other	NA
PerformanceScore_Below.average	NA
EngagementSurvey_Below.3	NA
EmpSatisfaction_X5	NA
SpecialProjectsCount_X1	NA
SpecialProjectsCount_X1.or.more	NA
age_Below.35	NA