Employee churn prediction

June 2020

By Simon Stausholm Rasmussen

As part of a masters exam in the course "Data mining for business decisions"

Contents

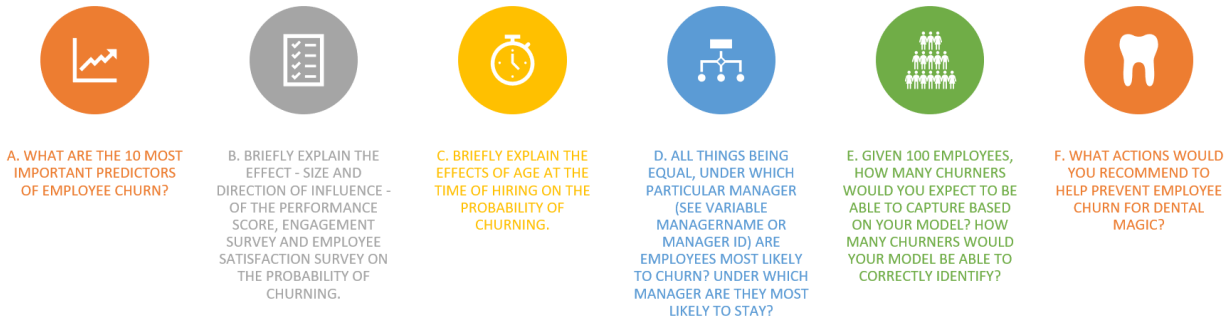
1. Introduction

Employee churn(attrition) is an infamous problem across businesses, as the loss of employees has a measurable effect on a numerous amount of business critical areas. These areas includes disruption of team dynamics, stall out of employee development, loss of team productivity, loss of valuable relationships and knowledge and finally damage to employer branding (Novakovic, 2020). Additionally is the process and cost of acquiring new employees substantial. Senior economist and former professor at Duke university, Bill Conerly did an "off-the-shelf" estimation for forbes.com which set the cost of an entry-level position turning over at 50% of their annual salary; mid-level positions at 125% and senior-level positions at 200% of their salary (Conerly, 2018). Even though these are just rough estimates, it can clearly be derived how much of a crucial task this is for organisations to manage. For Dental Magic the management team has taken the first step in trying to address these issues by collecting various data on their former and current employees. With this data in hand Dental Magic has reached out to external business intelligence consultants from Aarhus University in the hopes that they can help them turn data from information to knowledge and potentially improving their competitiveness.

1.1. Objectives & overview

The main objective as a consultant working for Dental Magic is trying to predict employee churn, by pre-processing the data, training and testing variations of a logistic regression, evaluating performance and presenting tangible business applications. While the overarching objective in the eyes of Dental Magic is reducing employee churn, thus reducing their overall HR-related costs across the organisation.

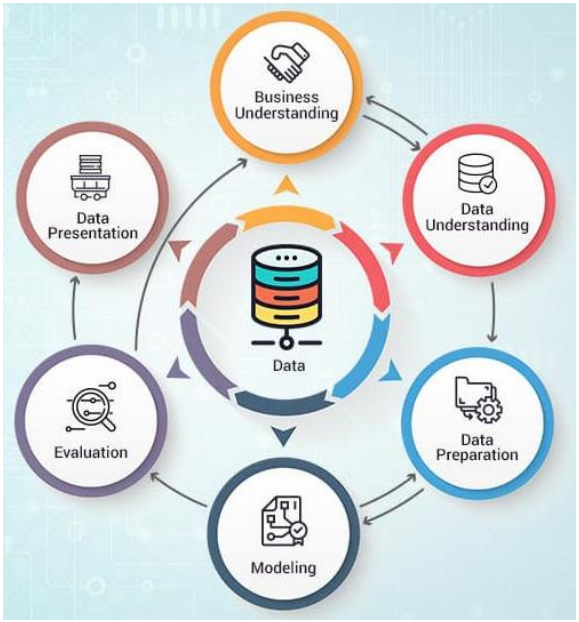
More specifically Dental Magic has scoped 6 concrete business implications that they are looking to gain knowledge based on the analysis. The scope of the project is presented below and will be the benchmark of the project.



Before proceeding with the next sections it is worth noting that this is a supervised classification problem and the main focus is using techniques related to logistic regression to derive and provide tangible business actions for Dental Magic. It should be noted that the datamining techniques are just a "mean" to reach the business objective. Furthermore since the scope presented by Dental magic is so geared towards interpretability, the principle of parsimony will have a major impact on the model selection approach.

2. Method

The structure of this report follows the well-recognized Cross Industry Standard Process for Data Mining (CRISP-DM). This framework provides a unified end-to-end guideline for practitioners within the field of data mining and data science. As shown on the figure, it is an iterative process that originate in a comprehensive business and data understanding. Then the process expands to the more technical implications such as data preparation, modelling and evaluation. It is in these steps where most of the iterations is usually taken place. Finally the framework addresses the aspect of implementation and deployment. The key takeaway of this approach in relation to this report is the element of combining datamining and business implications.



2.1. Data

As defined in the CRISP-DM framework the 2nd and 3rd step is to understand the given data, convey it to business understanding and pre-process it. Consequently the following paragraphs are going to address these steps.

First and foremost, the presented data from Dental Magic consist of 310 observations spread across 35 variables containing various employee information such as gender, payrate, race and much more. These variables will be further elaborated on in the following sections.

As an extension to the original dataset, 4 new variables has been created based on other existing variables. Resulting in 39 total variables. These new variables include:

Created variables	**Procedure**
target	Based on Employmentstatus = voluntarily terminated Binary (yes(1), no(0))
age	Based on DOB, calculated age based on time of hiring. Numeric (19-63 years)
yearsemp(seniority)	Initially based on Datehired and Dateterminated, but with iterations seniority was calculated for active employees as of 13-06-20. Numeric(0-14 years)
daycountpr	Based on LastPerformanceReviewDate - count of number of days since last performance review as of 13-06-20. Numeric (470-527 days)

Moreover it is assumed that this dataset is representative of the entire population and thus the derived analysis and applications are transferable to the actual organisation.

2.2. Target variable

The target variable isn't initially provided, thus having to be derived from the "EmploymentStatus" variable. Since Dental Magic primarily scoped the project to only involve the employees who left the company voluntarily, this has to be incorporated in the new target variable. In order to create a target variable that encompasses this requirement, an if-else-statement is expressed based on the "EmploymentStatus" level

"voluntarily terminated" to mutate a new variable called "target". This process results in a binary variable with the levels "yes" for voluntary termination and "no" for being active(still employed).

This new binary target variable has a distribution of 70,2% no(0) and 29,8% yes(1), thus being relatively skewed, however realistic in relation to "real-life" scenarios where more people tend to stay rather than leave. This will definitely imply that the final model will be a lot more accurate in predicting no's and one should be aware of this in regards to evaluating the models to have a good balance between predicting yes's and no's.

NB: In the process of creating the target variable, 15 observations related to "terminated by cause" has been omitted, since these datapoints has nothing to do with voluntary churn. Thus changing the percentage of yes's from 28% to 29%.

2.3. Excluded variables

In this section each of the individual variables has been assessed to gain an initial understanding of how they could affect the model. In this process some variables has been deemed as redundant and these are therefore excluded. These excluded variables are visualized in the table below with clarification.

Excluded Variables	**Reason**
Employee name	A person's name doesn't infer churn, therefore dropped.
Termd	A target variable has been created instead called "target", thus this is redundant.
EmploymentStatus	This has been used to create the new target variable and is redundant.
ZIP	Is represented by State – Though admitting that ZIP contains more granular information, the aspect of encoding such a continuous variable is quite a computational task.
Hispanic/Latino	Is very inconsistently encoded and should be included in the RaceDesc variable. By inconsistent it is meant that the interpretability isn't clear compared to RaceDesc.
LastPerformanceReview_Date	Used to calculate daycountPR (days since last performance review)
DOB	Used to calculate age at hiring
Date of termination	Used to create the Yearsemp(seniority) variable
Date of hiring	Used to calculate age at hiring
TermReason	Introduces information about employees which is too correlated with the target variable.
Yearsemp(senority)	Was deemed to introduce too much signal with basis in a too crude/bias assumption.
daycountpr	Was deemed to introduce too much signal with basis in a too crude/bias assumption.
DaysLatelast30	Didn't carry any information
EmpID	Database generated primary key and is represented by another factor variable
MarriedID	Database generated primary key and is represented by another factor variable
MaritalstatusID	Database generated primary key and is represented by another factor variable
GenderID	Database generated primary key and is represented by another factor variable
EmpstatusID	Database generated primary key and is represented by another factor variable
DeptID	Database generated primary key and is represented by another factor variable
PerfScoreID	Database generated primary key and is represented by another factor variable
FromDiversityJobFairID	Database generated primary key and is represented by another factor variable
PositionID	Database generated primary key and is represented by another factor variable
ManagerID	Database generated primary key and is represented by another factor variable

2.4. Data pre-processing

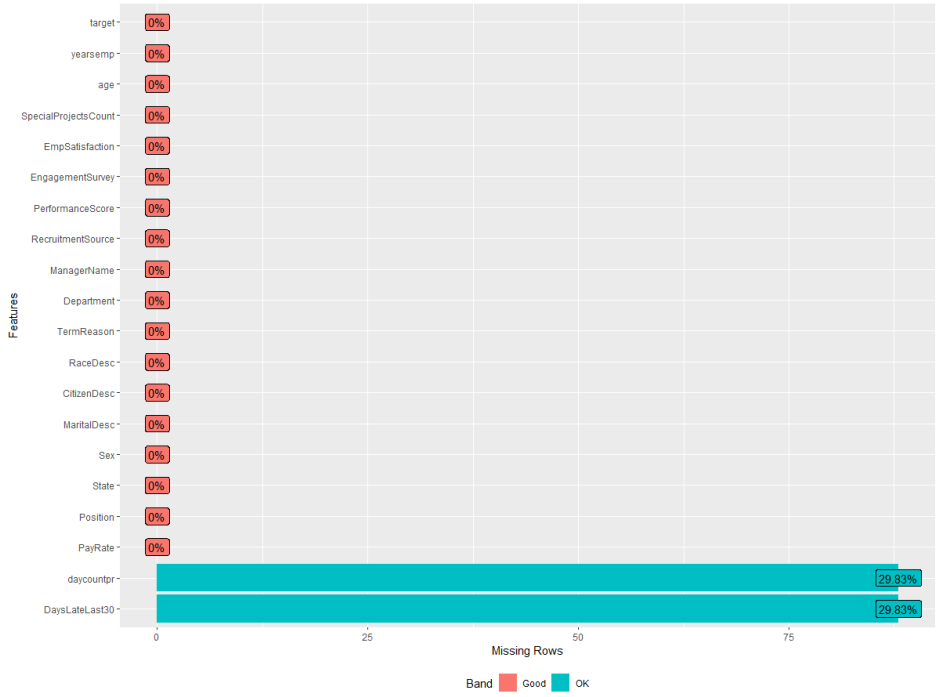
This section will focus the pre-processing steps performed including; treatment of missing, outliers, variable transformations, sampling, partitioning and presentation of the final input variables.

NB: Before doing any feature engineering, transformations or treatment of missing values the dataset is partitioned(split). This is done to address and minimize data leakage. This will be further clarified in the "sampling and partitioning" section of the report.

2.4.1. Treatment of outliers and missing values

Missing values

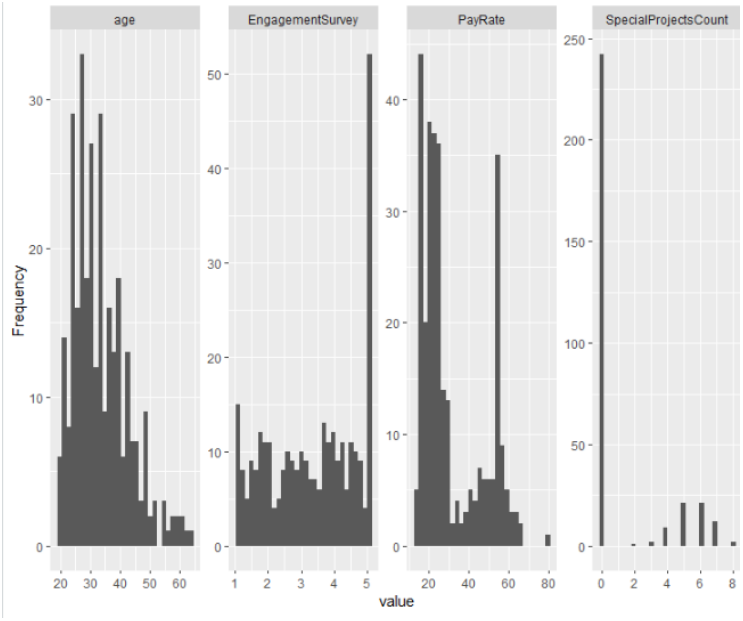
First of all it is important to assess the kind of missing's that is present in the data. There are 2 overarching types of missing's – The random and the not random. This is important to assess since it can have significance in the context of the data and how one choses to handle them. The random missing's doesn't require much consideration, and can follow standard procedures such as omitting and imputing. The missing's that isn't random needs to be handled a bit more carefully, these considerations is important so that one doesn't introduce unnecessary bias by treating them under the wrong assumptions.



Based on the plotted missing 2 variables is found to have 29,83% N/A's, these missing values are not random and are all related to employees being terminated. This is an indication that the database has deleted this information on the termination day. The "daycountpr" was initially created in hopes of it containing valuable information about the frequency of performance reviews, but unfortunately this is not the case. Based on the above mentioned assumptions about the missing, these variables will be omitted.

Outliers

The numerics that are considered in this project is plotted below and it can be seen that the variables PayRate, age, SpecialProjectsCount and EngagementSurvey are quite skewed. This skewness will be further adressed in the variables transformation section later. Furthermore, the variables shows no sign of outliers, thus nothing will be adressed further on this matter.



2.4.2. Variable transformations and derived features

In this section each of the selected input variables have been through various transformations. These transformations can be seen in the overview below. Every transformation is then argued for in the following paragraph.

Input variables	Original state	Transformation performed	Result
Payrate	Numeric	Discretized since it was skewed	2 Binary predictors
EngagementSurvey	Numeric	Discretized since it was skewed	2 Binary predictors
Age	Numeric	Discretized since it was skewed	2 Binary predictors
Position	Fact(32 levels)	Lumped to reduce levels, thus a more parsimonious model. Creating a “other” category. (+ one-hot encoded)	4 Binary predictors
State	Fact(28 levels)	Lumped to reduce levels, thus a more parsimonious model. Creating a “other” category. (+ one-hot encoded)	2 Binary predictors
Sex	Fact(2 levels)	One-hot encoded	2 Binary predictors
MaritalDesc	Fact(5 levels)	One-hot encoded	5 Binary predictors
CitizenDesc	Fact(3 levels)	Near-zero variance, thus omitted	N/A
RaceDesc	Fact(6 levels)	Lumped to reduce levels, thus a more parsimonious model. Creating a “other” category. (+ one-hot encoded)	5 Binary predictors

Department	Fact(6 levels)	One-hot encoded	6 Binary predictors
ManagerName	Fact(21 levels)	One-hot encoded	21 Binary predictors
RecruitmentScore	Fact(23 levels)	Lumped to reduce levels, thus a more parsimonious model. Creating a "other" category. (+ one-hot encoded)	11 Binary predictors
PerformanceScore	Fact(4 levels)	Lumped to reduce levels, thus a more parsimonious model. Creating a "below avg." category. (+ one-hot encoded)	3 Binary predictors
EmpSatisfaction	Numeric	Factorized then One-hot encoded	5 Binary predictors
SpecialProjectsCount	Numeric	Discretized since it was skewed	2 Binary predictors

Categorical transformations:

Encoding**:** As introduced in the overview above, all categorical has been encoded with a one-hot approach. This approach yielded better results compared to dummy encoding – this is due to the fact that dummy encoding uses a n-1 approach, thus reducing the number of predictors. The trade-off with dummy encoding vs one-hot where dummy encoding reduces variance was simply not worth it compared to the bias penalty the model suffered. There was no real argument for assuming an equal-length distance between categories, so ordinal encoding was not applied.

Lumping**:** Since some of the factors provided, included an excessive amount of levels, lumping was considered in order to create a more parsimonious model. Lumping was not applied ManagerName since this is needed for interpretability in the project scope by Dental Magic. The PerformanceScore levels "pip" and "needs improvement" was lumped together as "below avg." even though this predictor also was scope for interpretability. This is done since it improved the overall model while still maintaining sensible interpretation. The lumping resulted in a significant reduction in the number of predictors, thus reducing variance and improving interpretability. That being said lumping has a quite a accuracy trade-off, but the variance reduction was considered more important given the objectives of the project.

Numeric transformations:

All the included numeric predictors were heavily skewed beyond a point where logging or using YeoJohnson as a transformation technique wouldn't normalize them. Therefore, all the numeric variables was discretized into simple categories, e.g. payrate was discretized into "above average" and "below average"; special projects into "no special projects" and "1 or more special projects" etc.

All of the above mentioned transformation is performed trough the caret package within the resampling boundary, thus ensuring minimal data-leakage. Moreover different transformation configurations for future iterations are addressed in the model performance discussion.

2.4.3. Sampling and data partitioning

Due to the relatively sizeable amount of observations, the dataset is initially partitioned into a training and a validation set of 70 % for training and 30% for the validation. This split is based on a the trade-off between having enough data to train on, while still having a considerable test set for model performance assessment. Since the target variable is heavily skewed, random sampling consequently run the risk of introducing a

significant sampling bias. If the sampling isn't representative of the distribution(population), then one can define restrictions that ensures that the sampling is distributed into homogeneous subgroups called strata. Thus to satisfy a representative distribution, the data has been stratified on the target variable leaving around 29% churners in both the test and training set.

As briefly mentioned in the introduction to the pre-processing section, it has been chosen to do the partitioning as the first step in the data preparation process. This is done to minimize data leakage, which would have been generated if e.g. transformations, calculations or imputations were performed on the combined dataset beforehand. Moreover on the subject of sampling and data leakage, then it is just as important to address data leakage in the resampling process as well. The resampling process refers to the model fitting on drawn samples from the training set e.g. cross-validation. This entire process is handled in caret to ensure a consistent resampling process.

3. Results

This section will stepwise go through the process of training a number of candidate models, selection criteria, performance assessment and presentation of a final model. What is really important in this section is keeping an eye out for the business objectives and project scope. As stated in the introduction the main purpose is developing a model that can predict and capture as many churners as possible. Additionally to meet the scoped requirements the emphasis should be on model interpretation and simplicity.

3.1. Candidate models

Following the project scope, the candidate models considered is limited to logistic regression with or without inclusion of shrinkage techniques. Moreover the process follows the "no free lunch theorem", ensuring that the basis of the model training is following a iterative approach and that no specific model fits certain problems better until proven through tests.

Proposed models:

- Logistic Regression
- Elastic net
- Elastic net *(with oneSE)*

Logistic Regression

Logistic regression is a very powerful tool for binary classification tasks such as this one presented by Dental Magic. Logistic regression is based around using a maximum likelihood method for model fitting and parameter estimation. The intuition behind the maximum likelihood method is that based on a defined threshold the model will assign a value based on its probability of belonging to either of the binary categories. The major advantage of the logistic regression is that is very useful in the sense of interpretability and making inference whilst being comparatively parsimonious. One could argue that the disadvantage of such a model is that it isn't very flexible and other more complex model is able to capture more signal in the data compared.

As introduced, when applying a binary logistic regression one must define a classification threshold, this threshold determines when a value with a given probability should be assigned as being either yes(1) or no(0). Moreover, the threshold specification is problem dependent, which can make for a useful tuning parameter, when targeting type 1 or type 2 errors.

***Elastic net** with/without one standard error*

Within the field of shrinkage there are two main approaches; ridge regression and the lasso. These techniques tries to regularize the coefficient estimates by shrinking the coefficients towards zero. The technique of shrinkage can significantly reduce model variance, which is highly appropriate considering the bias-variance trade-off. The shrinkage penalty for both ridge regression and the lasso is called lambda. The larger the penalty lambda, the closer to zero the algorithm shrinks the coefficients. The lasso can further be used as a feature selection method since it can truly shrink coefficients to zero.

The elastic net is an algorithm that allows for incorporation of "best of both worlds". This algorithm takes the same tuning parameter lambda as ridge regression and lasso but also an additional parameter alpha. The alpha parameter is used as a scale representing the distribution-effect between ridge regression and the lasso. If the elastic net selects an α value of 0 then pure ridge regression is applied, and if α has value of 1 then a pure lasso model is applied. The lambda value is the amount of shrinkage/penalty applied.

The main advantage of an elastic net with or without oneSE is that it is a more parsimonious model and that it can stabilize models. The main disadvantage is comparable to the logistic regression that it is very inflexible, thus being penalized with a higher error rate relative to other more flexible models such as random forest and support vector machines.

Models	**Tuning parameters**
Logistic regression	Threshold
Elastic Net	α, λ
Elastic Net (oneSE)	α, λ

NB: The oneSE specification is further elaborated on in the resampling section below.

3.2. Model selection approach

This paragraph addresses 4 overarching topics related to model selection. These 4 subjects include resampling methods and the argument of selection hereby; a discussion of evaluation metrics and the argumentation for choosing one over the other; a debate covering the bias-variance trade-off paradigm; and finally an argumentation for error-cost in the context of Magic Dental.

Resampling

Following the "no free lunch theorem", then there is no such thing as a "best" model selection approach, it is very dependent on the objectives and the available data. Since the preferred model is indicated based on training predictions, it is important to address coherent resampling method that fits the objective and data. The resampling approaches considered in this project consisted of different cross-validation configurations such as k=5/10 and repeats=10/5 with the inclusion of one standard error in the elastic net. The most fitting approach was deemed to be K=10 and repeats=5, this approach is more catered towards having a smaller "test" set when model training. This compared to the k=5 equates to a smaller error rate since the algorithm has more datapoints to train on and less test points to validate on.

OneSE

The one standard error approach is based on the notion that error estimations have an inherent variability. The disparity in performance between minimum error approach compared to an oneSE approach is though

often minimal. OneSE has the advantage that it is further based on an assumption that tuning parameters tend to overfit, thus the oneSE combined with an elastic net is the most parsimonious model considered in this project.

Evaluation metrics

When selecting a final model to present Dental Magic, one must consider which terms the model selection should be based on. This aspect is by far the most detrimental model selection aspect, since presenting a model based on inappropriate metrics can end up in critical problems when trying to address the business objective.

The default metric for classification(using caret) is accuracy. The problem with accuracy is that it doesn't take the imbalances of the target variable into account. Since the objective is geared towards both predicting and catching churners the accuracy metric isn't sufficient, thus all the models are trained using an "area-under-the-curve" (AUC) approach. Training the models using an AUC approach shifts the focus from getting the highest accuracy to getting a more balanced model evaluation metric. There are a few relevant performance evaluation metrics that is relevant to look at when comparing candidate models in the Dental Magic context.

Precision*:* The precision metric illustrates the ratio of churners correctly labelled as churned compared to the total number of churners predicted and is defined as: $TP/TP+FP$ where FP is a representation of "false alarms".

Recall*:* The recall metric illustrates the ratio of true churners compared to the true churners and false negatives. Recall is defined as $TP/TP+FN$ where FN is a representation of "misses".

F1 score/F-measure*:* The F1 score is a representation of the geometric mean between recall and precision. The metric measures the accuracy of the model, by balancing the precision and recall, and is commonly used when the target variable is imbalanced. F1 is defined as: $F1 = 2/((1/precision) + (1/recall))$

As a final remark to the metric section, one should also consider the misclassification cost. For example, is one error type worse than the other? In the Magic Dental case, one might argue that classifying someone as a churner who isn't is better than not catching a churner. The cost of HR initiatives on a false positives should be significantly lower than not executing HR initiatives which result in a churner.

Bias-variance trade-off

As hinted at throughout the prior sections of the project, the bias-variance trade-off is a topic which is extremely important in all kinds of data mining projects. This paradigm is a discern between having a low/high error rate due to bias or variance. Bias is a measurement of how far in general the predictions are off correct values, which is gives a sense of how well the model conforms to the underlying structure of the data. Variance is on the other hand is a reference to the variability of the model prediction aka. how well does the model generalize on "unseen" data. This paradigm is often discussed in relation to model selection, moreover in the aspect of one should choose a more flexible or parsimonious model. In the context of Dental Magic, it is assumed that the core of the project is on interpretability and generalisation, since the models that are predefined are relatively simple and not too flexible.

3.3. Final model

With basis in the presented candidate models and selection approach, this segment pinpoints the preferred model that is going to be used to make inference and be presented to Dental Magic for implementation.

Based on the metrics derived from the training predictions in appendix 4 it can be found that the logistic regression outperforms both the elastic net on all the relevant metrics. In the table below each of the 3 models are tested against the validation set with the accompanied metrics of relevance. This will be the basis of the final model selection.

Model	**Specification**	**Precision**	**Recall**	**F1 score**
Logistic regression	Logit(0,5 threshold)	62,9	65,3	64,1
Elastic net	alpha = 0,7 \ lambda = 0,017733	47,0	30,8	37,2
Elastic net (oneSE)	alpha = 0,1 \ lambda =0,09463	41,6	19,3	26,3

As seen on the performance evaluation schemes above it can be discovered that the logistic regression outperforms the elastic net(with and without oneSE) on all of the metrics again. The models perform very differently on the metrics and it is worth noting that the logistic regression has a F1 score of 64,1 compared to the elastic nets with an F1 score of 37,2 and 26,3. Based on this performance assessment, the logistic regression is selected for further analysis.

Before moving on with the logistic regression, the specification of the elastic nets are worth quickly discussing. It can be found that the elastic net has a alpha of 0,7 which indicates a larger proportion of ridge regression, compared to the elastic net - oneSE with an alpha of 0,1 which indicates an almost fully lasso effect. Finally both of the elastic nets are more parsimonious models due to their penalizing tuning parameter lambda, which is probably the reason why it suffers so much in the error rate.

3.3.1. Overall predictive accuracy

Since the scope if the project doesn't evolve around accuracy per say, but rather the precision and recall one should assess how well the model is at capturing churners. The precision is the percent of employees that the algorithm predicts as churners who actually did churn. The logistic regression resulted in a precision of 62,9% which compared to all the configurations tried seem reasonably well performing. Moreover, the recall is an indication of how many employee churners the algorithm predicts in the dataset and in this case the model catches 65,3%. These measures will be put into context of Dental Magic in the discussion point 4.2.5, showing how these metrics equate to useful and understandable business information.

3.3.2. Observed versus predicted target values

To compare, assess and elaborate on the observed versus predicted variables one should investigate the relevant confusion matrix below.

Model	**Confusion matrix**		
Logistic regression	**Observed values**		
		Y	**N**
	Y	17	10
	N	9	52

The objective with the confusion matrix here is to highlight the correct, but more importantly the incorrect classifications. As illustrated, out of the test sample of 88 employees, the model predicts 17 churners and 52 non-churners correct. Looking at the false positives, the model predicts 10 employees as churners, without them actually being churners. In the context of Dental Magic, then this type of "false alarm" is not the worst kind of error. The only issue with this is that Dental Magic spends unnecessary resources on trying to target these employees, but this cost should be minimal. On the other hand is the false negatives located which is a more detrimental type of error. This error refers to the blatant misses, where the model isn't able to catch the churners, which would mean that Dental Magic would have 9 people churning without them being aware of the situation.

***NB*:** How Dental Magic should handle these issues is covered in section 4.2.6.

4. Discussion

This final part of the project will first of all discuss the model performance and how different approaches/ techniques could have altered the direction of the model development. Following this assessment, the scoped business problems will be discussed and coherently elaborated on.

4.1. Assessment of model performance

With reference to the already outlined model performance in the prior section, this section will include some final thoughts related to the overall performance and how it can be established in the context of Dental Magic including future iterations.

Given the circumstances of a skewed dataset without countless datapoints, the model finds a decent balance between precision and recall, which was the objective of the analysis. That being said, as mentioned in the predicted versus observed section, the model does come with certain errors that can have significant influence on Dental Magic, thus having room for improvement.

It is worth mentioning that all the ID variables was dropped primarily since they didn't carry any "real-world" information. That being said, it was found that e.g. the positionID carried fewer levels than "position". By figuring out how these were related, one could potentially present a even more parsimonious model by using some of the ID's. The introduction of e.g. the seniority and DaysSincePerformanceReview variables did improve model accuracy significantly, but was deemed too bias, since the calculations were based on very crude assumptions about the data. If these variables were provided initially, then the model performance would probably be improved. Finally the inclusion of feature extraction techniques could have further contributed to the overall result by deriving relevant variables.

4.2 Discuss how the final model contributes to the solution of the business problem

When presenting the results of how the model contributes to the solution of the business problem, it is important to lead the attention to both the data mining and business aspect. In the visualization below the path from datamining objective to business solution is mapped out.

As shown on the chart, the data mining project can be used as instrument to gain insights on potential churners, this information was not available to the management team and HR prior to the project. This new knowledge about the employees gives Dental Magic the opportunity to create a cross-organizational employee retention strategy, that can encompass individuals. If this strategy is executed correctly, then Dental Magic has the opportunity to reduce their HR related costs, thus solving their initial business problem of spending too many resources on voluntary terminated employees.

NB: The more specific strategies and focus points are elaborated on in the following sections

4.2.1 What are the 10 most important predictors of employee churn?

The top 10 most important predictors when predicting employee churn is listed below and seen in appendix 1. The direction of influence is also included since this could lead to valuable insights for Dental Magic and can be seen in appendix 2.

Variable	Direction of influence on churn
1) Recruitment from the diversity job fair	Coefficient (-) = Less likely to churn
2) Location Massachusetts	Coefficient (+) = More likely to churn
3) Employment under Peter Monroe	Coefficient (+) = More likely to churn
4) Recruitment from Monster.com	Coefficient (-) = Less likely to churn
5) Employment under Kelly Spirea	Coefficient (+) = More likely to churn
6) Recruitment from an employee referral	Coefficient (+) = More likely to churn
7) The race white	Coefficient (-) = Less likely to churn
8) Recruitment from search engines such as google, bing and yahoo	Coefficient (-) = Less likely to churn
9) Recruitment from a professional society	Coefficient (+) = More likely to churn
10) Employment under Ketsia Liebig	Coefficient (+) = More likely to churn

4.2.2 Briefly explain the effect - size and direction of influence - of the performance score, engagement survey and employee satisfaction survey on the probability of churning.

With basis in appendix 2 and 3 each of the variables; performance score, engagement survey and employee satisfaction is briefly addressed in relation to size and direction.

Performance score:

All the performance scores are relatively low on the overall importance rank.

“Fully meets” has the largest weighting with a positive coefficient meaning that employees in this category are more likely to churn.

“Exceeds” comes next with a negative coefficient meaning that employees in this category are less likely to churn.

“Below average” has no significant influence/important, thus no coefficient and employees in the category has no influence on churn.

Engagement survey:

“Engagement survey above 3” is very near the bottom of the rankings with a positive coefficient, thus employees within this category is more likely to churn.

“Engagement survey below 3” has no significant influence/important, thus no coefficient and the category has no influence on churn.

Employee satisfaction:

Score = 4, is quite high on the rankings of importance with a negative coefficient, thus employees with this employee satisfaction is less likely to churn

Score = 3 & 2, is approximately in the middle of the rankings of importance with negative coefficients, meaning employees with this score is less likely to churn.

Score = 1, is near the very bottom of the rankings and has a positive coefficient, meaning this very low score equates to employees being more likely to churn.

Score = 5, has no significant influence/important, thus no coefficient and the category has no influence on churn.

4.2.3 Briefly explain the effects of age at the time of hiring on the probability of churning.

The age variable was initially split into 2 binary predictors(above 35 and below 35), with this in mind the employees with an age above 35 at their hiring date are less likely to churn.

Hiring employees below the age of 35 has no influence on their probability of churning. Thus implying that with an increase in age decreases the likelihood of churning.

4.2.4 All things being equal, under which particular manager are employees most likely to churn? Under which manager are they most likely to stay?

With reference to appendix 2 and 3 again and following an approach of locating the highest ranked manager with a positive and negative coefficient it was found that employees are most likely to churn under Peter Monroe and less likely to churn under Amy Dunn.

4.2.5 Given 100 employees, how many churners would you expect to be able to capture based on your model? How many churners would your model be able to correctly identify?

In an example of 100 random selected employees, the model would identify(predict) 30.8% of the sample as churners and within the churner-category about 62,9%(precision) of them would be correctly classified as churners, equal to around 19 employees. If on the other hand one were given a sample of 100 churners, then the model would detect 65,3%(recall) of these, equal to 65 employees.

4.2.6 What actions would you recommend to help prevent employee churn for Dental Magic?

There are numerous actions that Dental Magic can take to counter employee churn. By starting out with looking at the insights in the abovementioned paragraphs, many potential initiatives can be derived. First and foremost it is recommended that Dental Magic starts with the "lowest hanging fruits". This could for example be creating a workshop where Amy Dunn could share her insights on how she manages to maintain her employees. Similarly the management team should look into the managers that carries a high rank and positive coefficient, such as Peter Monroe, Kelly Spirea, Ketsia Liebig amongst other. Magic Dental should moreover take a closer look at the way they recruit people. For example, the diversity job fair seems like a very fitting platform to recruit new employees, while using employee referrals and professional societies as an outlet for recruitment is more likely to fail. Moreover, could Dental Magic have more focus on employing individuals older than 35, since these are less likely to churn.

Deployment recommendations and recommended follow-up activities

As for deployment recommendations, then one would propose the c-level management team, to use this tool in corporation with the HR team on a monthly basis, to make an assessment of the current employees and their well-being. The tool should be used as a point of reference for discussion internally and additionally used as an instrument that can contribute to the overall organisational HR strategy.

As a consultant one should also give Dental Magic a point of orientation in regards to how they further can develop the model to improve it. In this case Dental Magic should first and foremost start tracking seniority consistently, since this predictor seemed to contain a lot of signal. Furthermore it would be advised, that Dental Magic could try developing more flexible models to get a lower misclassification rate.

It is additionally **incredibly important** to address the fact that the model should **not** be followed blindly since as discussed the model includes false negatives and isn't perfect.

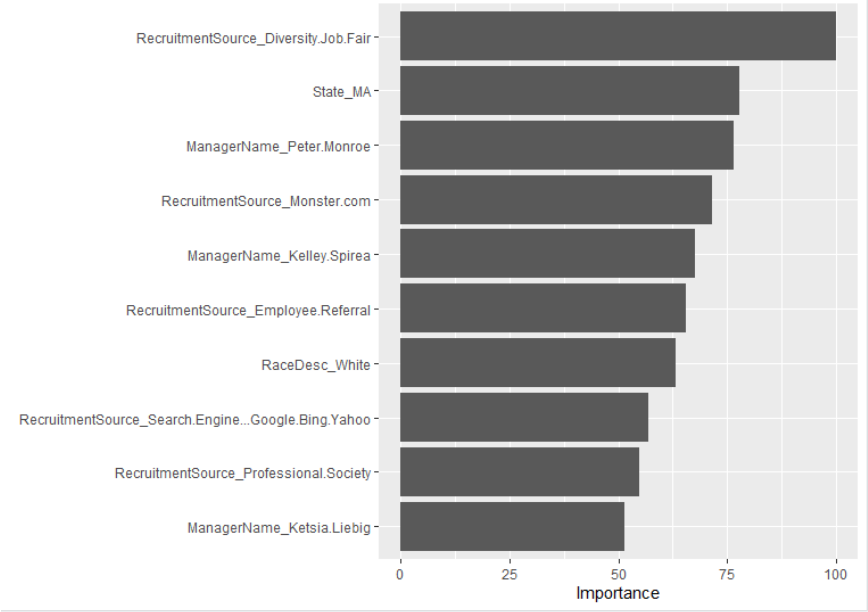
5. Bibliography

Conerly, B. (2018, 01 01). *Forbes*. Retrieved from Forbes: https://www.forbes.com/sites/billconerly/2018/08/12/companies-need-to-know-the-dollar-cost-of-employee-turnover/#e913dfd590ac

Novakovic, A. (2020, 1 1). *Insperity*. Retrieved from Insperity: https://www.insperity.com/blog/cost-of-employee-turnover/

6. Appendix

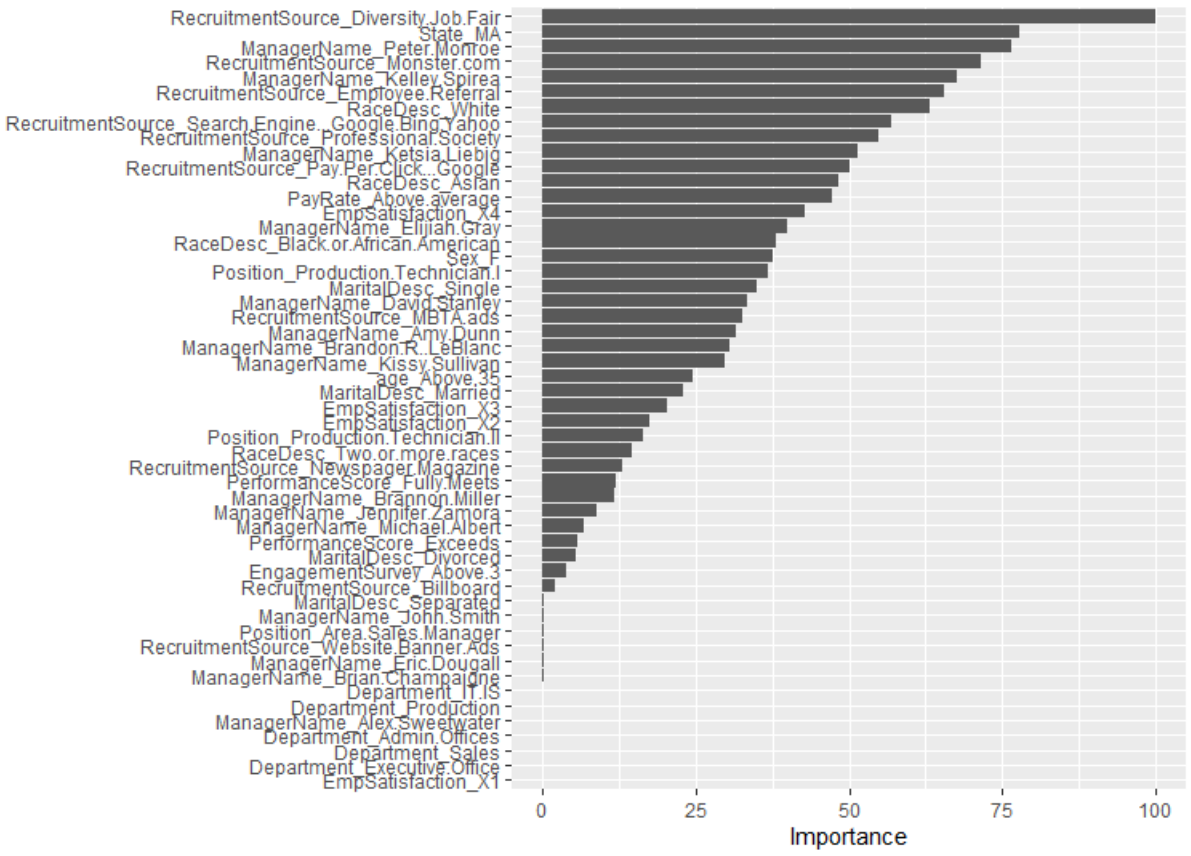
Appendix 1: Top 10 predictors



Appendix 2: Direction of influence (coefficients)

	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

Appendix 3: All predictors ranking



Appendix 4: Training errors for LR, EN and EN(oneSE)

LR training errors

```
          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
```

EN training errors

```
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
```

EN(oneSE) training errors

```
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
```