Implementing a Churn Model on simulated employee data

Claudine Chen

Introduction

In our vibrant entrepreneurial world, companies and organizations need to constantly keep their competitive edge, whether it's to retain customers or the best employees. A way to monitor and react to what is happening in these contexts is to collect and analyze data on human behavior and demographics, looking for which factors have the greatest impact on the desired result. With a model of customer behavior, a company can make informed decisions on how to provide a better experience and to optimize income. Machine learning and statistics offer tools to do exactly that, especially on a large scale.

The specific study of customers cancelling subscriptions or not returning using data is call churn modeling; churn is the overturning of something, and in this case means when a customer is lost. Alternatively this is also called customer lifetime value modeling.

Companies with subscription models, like Netflix, will want to keep members subscribed for as long as possible, and will look for key factors for why subscribers cancel their memberships, so that they can actively prevent this. Online retailers like Amazon don't have monthly subscription fees, but they similarly will be interested in keeping customers returning to Amazon to buy things. Companies that don't have contractual agreements with customers will define churn in other ways, such as by time between visits. Regardless of the definition, when looking for measures that can optimize income, companies will want to implement changes that have the highest probability of making an impact.

Making changes comes with an inherent cost, and proposals that are not based in evidence could potentially be cost inefficient or wasteful of resources. Relying on data for insights and inspiration for improvements is likely to be the most cost effective and efficient way to grow a company.

Data Description

The data I will use to explore churn modeling is employee information, offered by Medium.com on Kaggle (https://www.kaggle.com/ludobenistant/hr-analytics), with an eye on understanding employee churn - why are they leaving when they do. It contains 14999 simulated employees, and features include employee satisfaction level, their last evaluation, the number of projects, their average monthly hours, time spent at the company, whether they have had a work

accident, whether they have had a promotion in the last 5 years, their department, and salary level.

The data provided is rich in information that might impact an employee's attitudes. There are direct measures of an employee's satisfaction level within the company, as well as the company's satisfaction in the form of evaluations. A measure of workload can be gauged from average monthly hours and number of projects. A measure of rewards from the employer can be found in salary and number of promotions in the last 5 years. One thing that is not reflected is their seniority - this could potentially be amalgamated from department and salary level, assuming that the top earning people in each department might be more senior staff, and potentially time at the company.

The data is limited in that it doesn't reflect the full compensation package that employees would receive - it's possible that factors that are key in employees deciding to leave are not provided in this data set. Also, demographic information is not provided, which could be a proxy for an employee's motivations - whether they are looking to advance their career by moving to another company, or if they are comfortable in their position - age could be a potentially useful indicator. A way to probe the employees' motivation is to query their satisfaction level on specific topics, like job advancement, work responsibility, work hours, However, the data does reflect things that the company has some control over, and can improve, such as keeping an eye on workloads or rewards. The data cannot answer if there are personal reasons people are leaving - perhaps there is a toxic person on staff. Another thing that can impact people leaving are hiring practices - are you hiring people who are good matches for the company? It is possible that criteria for hiring people are in conflict to the real traits of good and long-lasting employees; this data will not shed any light on that. It will also not reveal anything about how well the company is working.

The data itself was very clean, and the read in process into Pandas was uneventful. There were no missing values, and all values that had numbers were read in as numeric types. Ratings are normalized to be between 0 and 1.

Data Exploration

A measure of the relationship between each feature is correlation. A correlation matrix of the data will give an indication of any linear relationships at a glance. Note that the categorical features and their N dummy variables have N-1 degrees of freedom, and have an internal correlation.

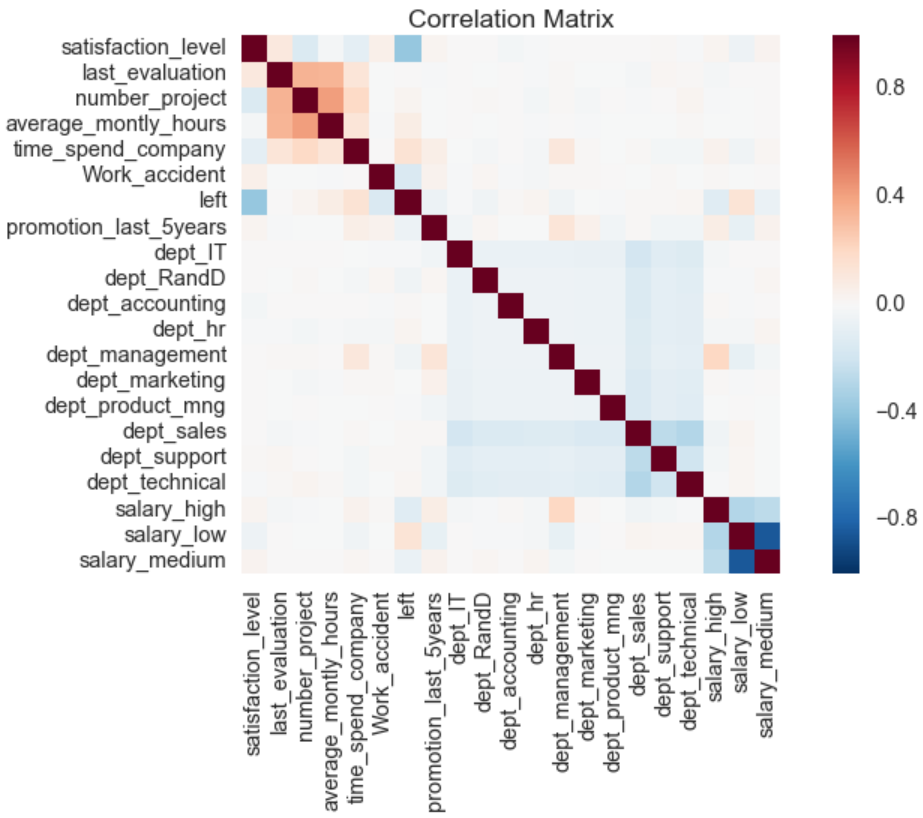


Figure 1. Correlation matrix of all features (categorical features have been expanded into dummy variables).

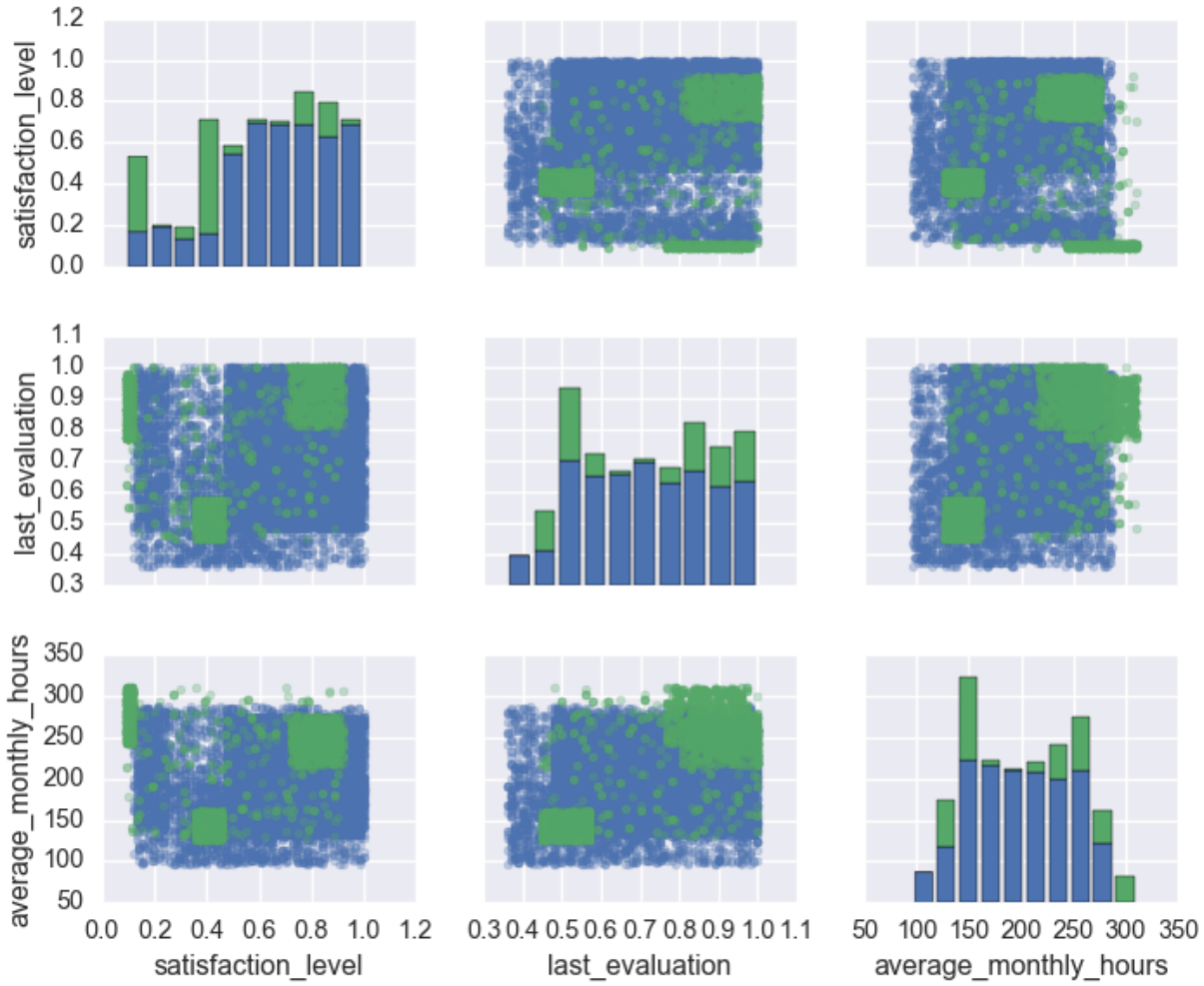
We can see that the features that have some direct correlation with leaving is satisfaction level, along with at a lesser level time spent at the company and a low salary. Between the features, there is positive correlation among last_evaluation, number_project, and average_monthly_hours, and to a lesser degree time_spend_company and satisfaction_level. There are also correlations between being in management and a high salary, getting a promotion in the last 5 years, and the time spent in the company. There is also some correlation between categories within the same features, as highlighted above.

To make sure a correlation is significant, the p-value is also calculated. The calculation for the p-values are shown at the end of this section; All p-values were less than 0.01, so all correlations are significant.

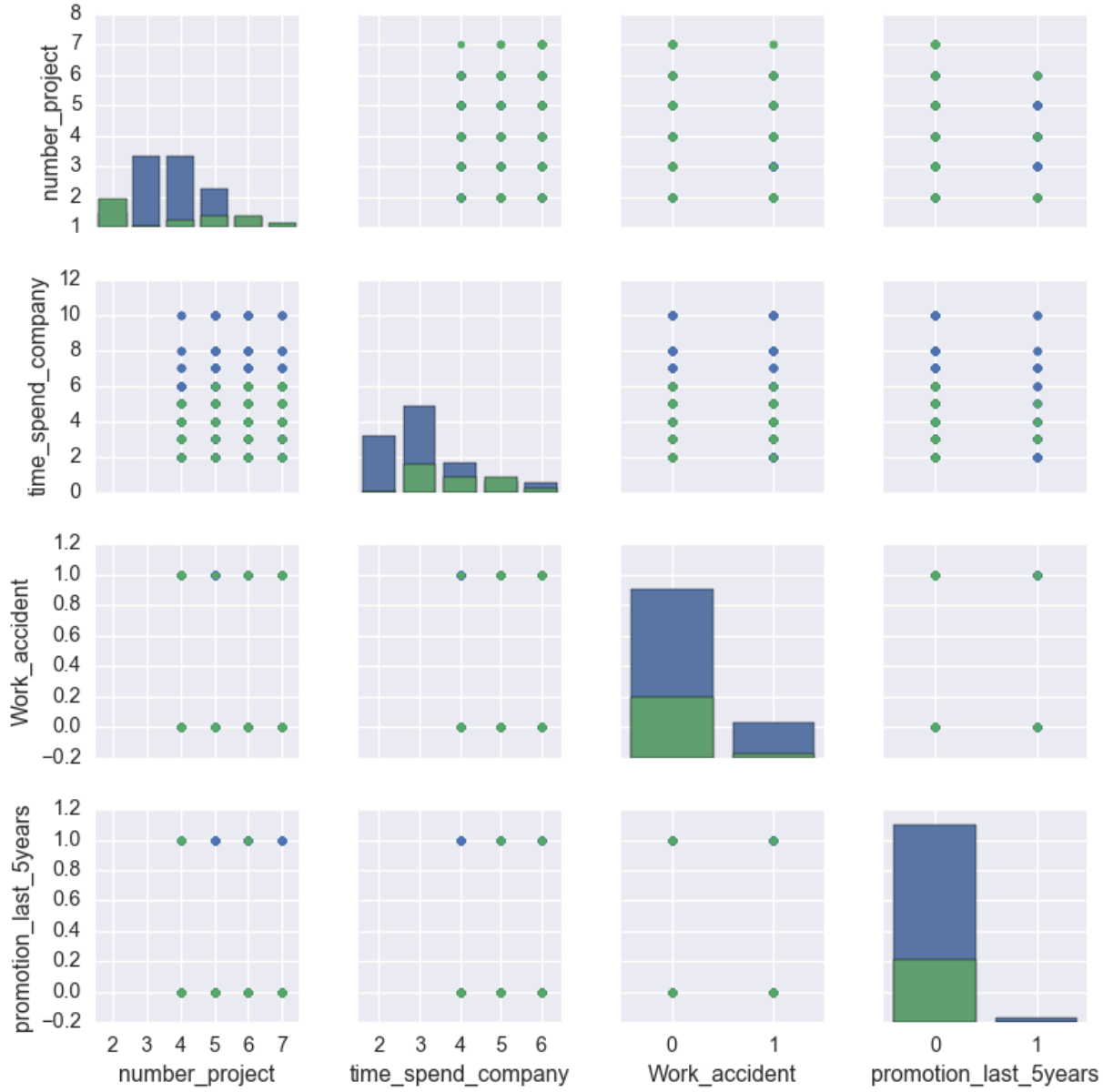
We can look closer at the relationships between features in pairwise plots. Coloring by the churn variable, we can see how the different cases might behave differently.

.

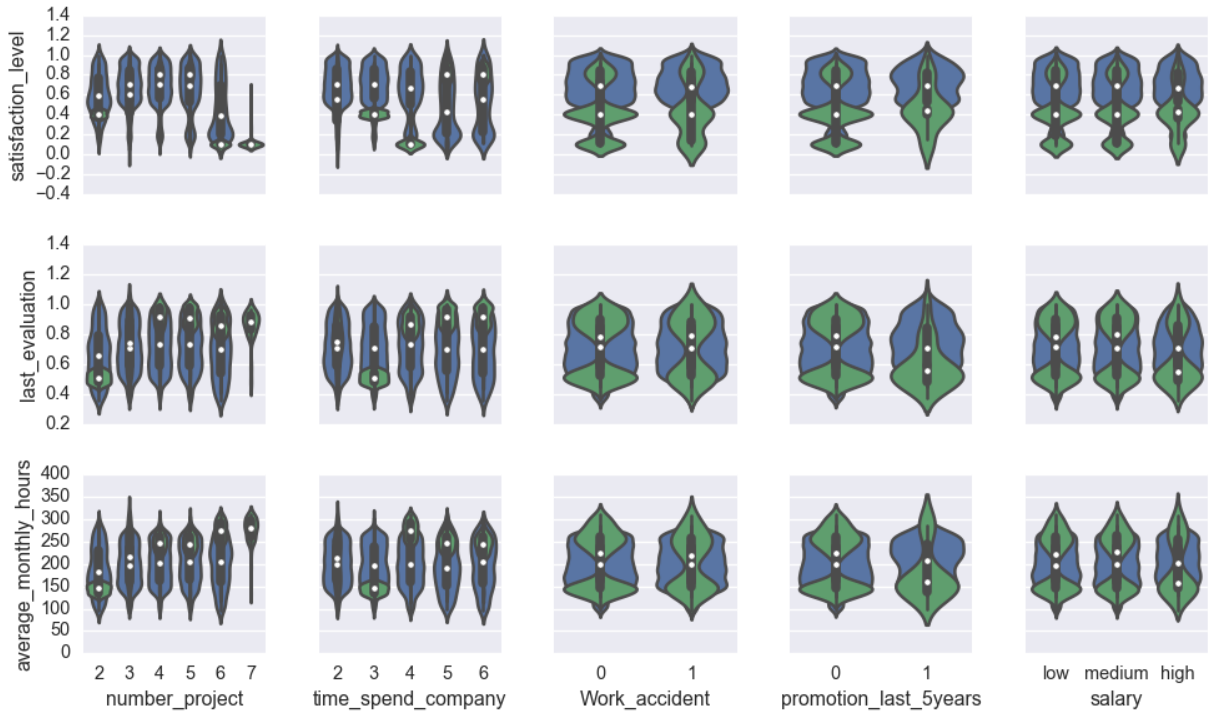
Figure 2. Pairwise comparison of (a) continuous features (satisfaction level, last evaluation, average monthly hours), (b) ordinal or categorical features (number of projects, time spent at the company, if the employee had a work accident, if the employee had a promotion in the last 5 years, salary), and (c) continuous vs ordinal or categorical features. These plots are colored by if the employee left (green) or is still employed (blue).

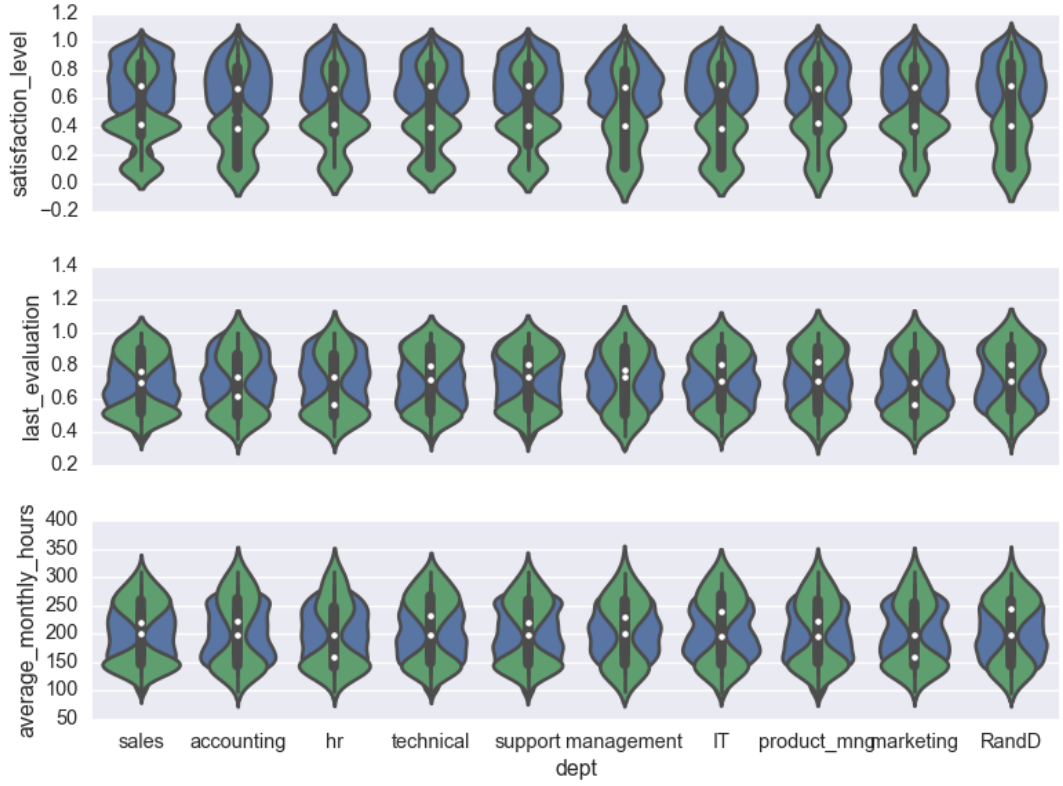


(a)


number_project
time_spend_company
Work_accident
promotion_last_5years
number_project
time_spend_company
Work_accident
promotion_last_5years

(b)


satisfaction_level
last_evaluation
average_monthly_hours
number_project
time_spend_company
Work_accident
promotion_last_5years
salary
low
medium
high


satisfaction_level
last_evaluation
average_monthly_hours
sales
accounting
hr
technical
support
management
IT
product_mng
marketing
RandD
dept

(c)

What we can see from these plots are:

- there is a bimodal distribution for leaving - seems there are two peaks in most plots for leavers - those who have high or low satisfaction, high or low evaluations, and those who work part-time or overtime.
- there are clearly three clusters of leavers when monthly hours and satisfaction level are compared: people who work part-time and have medium satisfaction levels and evaluations, people who work overtime and have high satisfaction levels and evaluations, and people who work overtime and have very low satisfaction levels but high evaluations.
- there are no leavers if the employee has stayed over 6 years. Leavers will leave within 5 years.

The three cluster of leavers are most apparent in the satisfaction vs work hours plots. They partition further by number of projects.

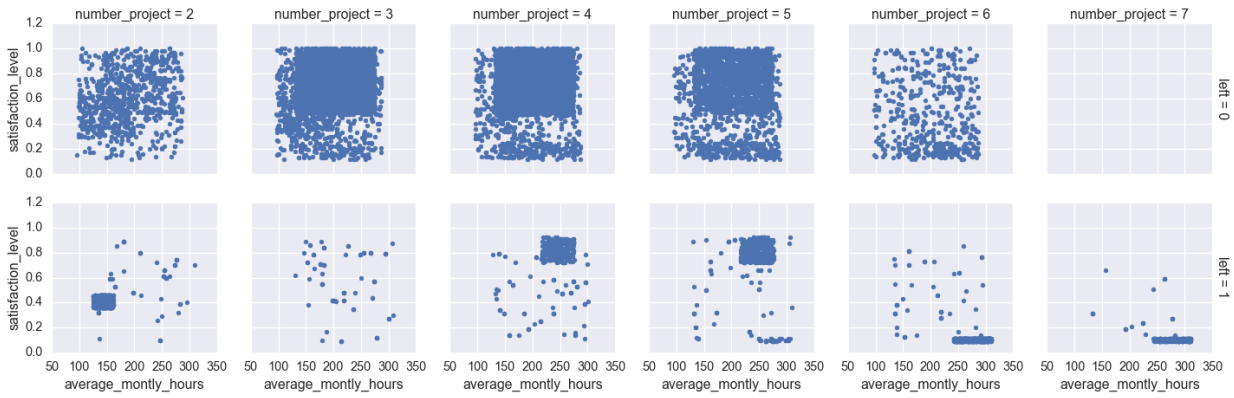


Figure 3. Satisfaction vs. average monthly hours, subsetted by number of projects and leaving.

Those with less than full time hours (less than 176 hours per month) were not vastly unhappy, so a possible interpretation is that people in this group desire more responsibility and a full time job, and therefore leave, or perhaps they are bored and therefore leave. This is supported by lukewarm evaluations. While the distribution of evaluations span from 0.3 to 1.0, the majority of those who leave have evaluations of 0.4-0.6. Helpful knowledge might be to parse satisfaction level into different aspects like sense of responsibility, number of hours, or work topic. If they are hoping to be turned full time and aren't, there might be a threshold under which people leave. This can be seen in that the majority of those who leave with 2 projects leave at 3 years.

Those who worked overtime and had 4-5 projects were very satisfied (rating of >0.7). A possible interpretation is that the number projects felt manageable, hence the high satisfaction level, they worked overtime because they wanted to, and they left because they simply got better offers.

Such people would be expected to get good evaluations, which is the case evaluations were greater than 0.8.

Those who worked overtime and had 6-7 projects and are very unsatisfied (rating of 0.1). In fact, all the people with 7 projects leave. A possible interpretation is that the number of concurrent projects was too overwhelming, leading to the forced overtime, and a desire to leave and find a more manageable or financially rewarding situation. Those who left also had high evaluations, of greater than 0.75. They might be overworked and unhappy and therefore motivated to leave their job, but since they are doing their job well, and they are able to get another job.

Would promoting people help? While only 2% of people got promotions in the last 5 years, out of those who do only 6% leave. Otherwise, 24% of people leave. In comparison, out of the people with only two projects, 66% leave, which is 1567 people. Those with four or five projects mostly stay; 14% churn, which is 1021. And those with six or seven projects mostly leave with all people with seven projects leaving; 64% of these people leave, which is 911 people.

Machine Learning Methodology

The fact that we have some useful demographics about our employees, and that we know if the employee left or not means we can apply machine learning to model the data, and we can use supervised learning algorithms. From our preliminary data exploration, we found that the majority of people who leave cluster into one of three groups, depending on satisfaction level, working hours, and number of projects. This suggest that the relationship is strongly non-linear or conditional. As this is a classification problem (with two outcomes, will leave, or will stay) a suite of standard classification algorithms are applied: logistic regression, random forest, and support vector machines (SVM). These techniques work well when the data is "tall" - when there are more data points than features; in our case we have 14999 data points and nine features (20 when expanded into dummy variables).

When measuring how well the models perform, a metric used is accuracy - what percent of the test cases were predicted correctly. A baseline model against which we will compare these machine learning models is one where all people are predicted to stay. The percentage of people who leave in our dataset is 24%. Therefore the baseline model will be 76% accurate just predicting that everyone stays.

We have labels that tell us what the real answer is, but we want to test our model as if it were encountering new data. This is possible by dividing the data into two subsets, training the model on the larger subset, and testing prediction with the other - this is called cross validation. A way to train and test, while subjecting all data to training and testing, is to divide the data into K equal parts, where each part will be used as the test data in one run of the training-test cycle, with the remaining data used as training data. So the training and testing will be run K

independent times. This is called a K-Fold cross validator. Since the data is unbalanced (24% leave) we use a KFold algorithm that ensures that every fold contains also 24% leavers.

To get the best performance, we select the values of the input parameters that result in the best accuracy; this is accomplished easily with a grid search, meaning, trying every combination of manually input parameter values and returns the best parameter combination based on performance using cross validation.

For each of the models, the performance is characterized by standard prediction metrics such as:

- confusion matrix (true positives, false positives, true negatives, false negatives)
- plot the true positives rate vs the false positives rate, aka receiver operating characteristic (ROC) curve; the ideal case would look like a right angle with the bend in the (0,1) corner.
- precision-recall (PR) curve; precision is the true positives over all predicted positives, and recall is the true positives over all real positives. great for unbalanced datasets since it ignores true negatives. The ideal case would look like a right angle with the bend in the (1,1) corner.
- the area under the curve (AUC) for the ROC and PR curves (larger values are better)

To summarize, for each model, the same process was followed:

- grid search for optimizing fit parameters by testing all combinations of inputs, like C which is the inverse regularization coefficient
- run the best model with cross validation using K-fold methodology. This means for each "fold", train the model, and then test. See which features have the most impact on the target variable, for each fold.
- calculate performance metrics

Results

Logistic Regression

Logistic regression is a regression model where the output is categorical, and in the simplest case, has two values. In our case we have just two values in our churn variable. It is also a linear model; the accuracy of this model on our data was 77%. Recall that a baseline model that predicts everyone stays has 76% accuracy; so this model is not doing much better.

For the linear case, the majority of actual positives are predicted to be negatives; the true positive rate (aka recall) is a measly 37%. Only half of those predicted as positives are actual positives; the precision is 52%. Ideally these would be close to 100%. The accuracy of the model is 77%, which is about the same as the baseline model. The ROC curve shows the

model is better than random. The area under the curve (AUC) is 0.8 (ideal is 1.0). As mentioned above, the precision and recall of our dataset is very poor, which is reflected in the precision-recall (PR) curve and the poor AUC for the PR curve (0.45).

While logistic regression is a linear model - it is possible to fit non-linear behavior by adding features that are non-linear combinations of the original features. We tried the model with second degree polynomial terms, third degree polynomial terms, and simply interaction terms. The accuracy of these models were respectively 85%, 83%, and 91%. The issue of increasing the number of features is that it makes feature space very sparse, and harder to model with logistic regression.

The ROC curves showed that nonlinear models were better at separating negative and positive cases better. The best result came from the interaction terms, with the area under the curve (AUC) being 0.93 (ideal is 1.0). The PR curve is also improved over the linear case, with an AUC for the PR curve of 0.76.

Random Forest

Decision trees are great for nonlinear and conditional relationships. They are also fast, but prone to overfitting. In practice the ensemble technique of random forest is used instead of a single decision tree. Random forest runs a number of decision trees on a subset of the data, and averages to improve prediction and counteract overfitting. This model also provides a measure of feature importance.

Random forest, even the default model, trains and predicts very well on our data. It achieves 99% accuracy in the training and test data. The ROC and PR curves look textbook, with an AUC of both of 0.99, indicating that the model is very good at separating positive and negative cases.

Over the five folds in the cross validation, the feature importance stayed very stable, and the five features that are very strong in influence are: satisfaction level, last evaluation, number of projects, average monthly hours, time spent at the company.

Support Vector Machines

Support Vector Machines (SVM) work to delineate the boundary between categories by focusing on the data points that are on the edge of the boundary. They are great for non-linear models, since you can have arbitrarily more dimensions with little additional computational cost, because of the "kernel trick". We will try a series of kernels - linear, polynomial, and rbf. A negative about SVMs in this context is it is more difficult to glean learnings from SVM - the coefficients especially for higher order models don't necessarily indicate feature importance.

SVM with the third degree polynomial and rbf kernel is able to achieve 96% accuracy; the linear kernel performs about as well as linear logistic regression. The area under the ROC curve for

both is 97%, and the precision-recall curves are both over 90%, which is a much better performance than polynomial logistic regression.

Discussion

Random forest was the best model for our data, for prediction and for suggesting which features had the most impact. The data, when explored in pair plots, was highly non-linear, and collected in quilt-like patterns, which indicate conditional relationships, and that random forest is a good choice. The accuracy of the model even on test data approached 99%, with beautiful ROC and precision-recall curves. The features that random forest highlighted as the most important were satisfaction level, last evaluation, number of projects, monthly hours, and time spent at the company.

Logistic regression and SVM was also applied to the data. As might be expected, linear models like logistic regression and SVM with a linear kernel did not predict well, with accuracy little better than predicting that everyone stays. When interaction terms were added to the logistic regression, prediction vastly improved, but still only to 90% accuracy, which is still far below the 99% accuracy of random forest. SVM with the rbf kernel also did much better than the linear kernel, with an accuracy of 96% on test data, but again, random forest outperformed SVM, and also provides feature importance.

Table 1. Model performance metrics

Model	**accuracy score**	**AUC-ROC**	**AUC-PR**
baseline	0.76		
logistic regression (linear)	0.77	0.80	0.45
logistic regression (polynomial deg=2)	0.85	0.91	0.69
logistic regression (polynomial deg=3)	0.83	0.91	0.73
logistic regression (interaction terms)	0.91	0.93	0.76
random forest	0.99	0.99	0.99
SVM (linear kernel)	0.77	0.79	0.50
SVM (rbf kernel)	0.96	0.97	0.91
SVM (3rd deg polynomial kernel)	0.96	0.97	0.92

Recommendations

Because of uncertainties in motivations which cannot be answered by the current data set, we recommend to improve the resolution of the data being collected to include satisfaction level of work hours, work topic, sense of responsibility, and advancement. This will supplement existing data on average monthly hours, and if the employee was promoted in the last 5 years.

With regards to information we do have, we know that there are three clusters of people who leave their job:

1. 567 people who left with less than full time hours were not vastly unhappy, and had medium evaluations. It is unknown what led to these lukewarm satisfaction levels, and understanding if it was related to dissatisfaction about their hours, their tasks, or their evaluations would educate the employers on what actions they could take to retain these workers better.
2. 890 people who left worked overtime, had 4-5 projects, were very satisfied (rating of >0.7), and also have very good evaluations (>0.8). A possible interpretation is that the number projects felt manageable, hence the high satisfaction level, they worked overtime because they wanted to, and they left because they simply got better offers. The majority of leavers leave within 5 years and have not had a promotion in those 5 years. Possible actions are to provide perks that make the job more attractive, like higher salary (do a cost analysis of how much is lost by retraining a new person, and pay the high productivity employee this amount), more flexible working hours, more holiday, more stock options or equity, or anything else the company can offer at low to minimal cost.
3. 886 people who left worked overtime, had 6-7 projects, were very unsatisfied (rating of 0.1), but had very high evaluations (>0.75). A possible interpretation is that the number of concurrent projects was too overwhelming, leading to the forced overtime, and a desire to leave and find a more manageable situation. The employer could lessen the load of these unsatisfied high performers.

In terms of priority actions, the employer could reduce churn by reducing the workload of the third group who are overworked. All people with 7 projects leave, and the majority of people with 6 projects leave. By reducing from their workload they might reduce the churn rate of people with 6-7 projects from 64% to the churn rate of people with 4-5 projects 14%. Additionally, the churn rate of those with promotions is 6%. It would be helpful to know the cost to the company of losing these highly productive people and retraining new employees - this cost could be invested in keeping these highly productive people.