

Project one

Dsc 680



September 18, 2022

Colton Cook

**Table of Contents**

**Business Problem………………………………………………………………………………………………………………….2**

**Background/History………………………………………………………………………………………………………………2**

**Data Dictionary……………………………………………………………………………………………………………………..2**

**Data Explanation…………………………………………………………………………………………………………………..3**

**Methods……………………………………………………………………………………………………………………………….3**

**Analysis…………………………………………………………………………………………………………………………………4**

**Modeling Setup……………………………………………………………………………………………………………………..7**

**Feature important ………………………………………………………………………………………………………………..7**

**Conclusion…………………………………………………………………………………………………………………………….8**

**Assumptions………………………………………………………………………………………………………………………….9**

**Limitations…………………………………………………………………………………………………………………………….9**

**Implementation Plan……………………………………………………………………………………………………………10**

**Challenges……………………………………………………………………………………………………………………………10**

**Future Uses/Additional Applications……………………………………………………………………………………10**

**Ethical Assessment………………………………………………………………………………………………………………10**

**Business Problem**

Looking into whether managers got positive or negative reviews from employees. Positive and negative reviews are important in a company as if managers have positive feedback, it should mean a happier workforce. An unhappy workforce leads to attrition of the workers and can be very expensive to replace. Its not only the time to replace, but also the knowledge the worker leaves and the time to get a new employee up to speed. All of this is expensive, so avoiding this is best practice.

**Background/History**

Solving the employee attrition problem would go a long way to keeping business costs down. The cost of replacing an individual employee can range from one-half to two times the employee's annual salary. The annual overall turnover rate in the U.S. in 2017 was 26.3%, based on the Bureau of Labor Statistics. So, a 100-person organization that provides an average salary of $50,000 could have turnover and replacement costs of approximately $660,000 to $2.6 million per year. Fifty-two percent of voluntarily exiting employees say their manager or organization could have done something to prevent them from leaving their job. This seems like a problem worth trying to solve.

**Data Dictionary**

|  |  |
| --- | --- |
| **Feature** | **Description** |
| Age | Person’s age |
| MnagerReview | 0-negetive 1-positive |
| BusinessTravel | Travel for work |
| DailyRate | Amount Per day for salary |
| Department | Which one they belong too |
| DistanceFromHome | How far to drive to work |
| Education | 1-Highschool 2-some college 3-bacholers 4- masters 5- PhD |
| EducationField | Board field of education |
| EmployeeCount | Count of employee |
| EmployeeNumber | Number of Employee |
| EnvironmentSatisfaction | 1 low 4 high |
| Gender | Female 0 Male 1 |
| HourlyRate | Amount paid per hour |
| JobInvolvement | 1 low 4 high |
| JobLevel | Level of senority |
| JobRole | Job Title |
| JobSatisfaction | 1 low 4 high |
| MaritalStatus | 0 divorced 1 married 2 single |
| MonthlyIncome | Amount paid per month |
| MonthlyRate | Rate per month |
| NumCompaniesWorked | Number of companies worked at |
| Over 18 | If over 18 |
| OverTime | Worked Overtime |
| PercentSalaryHike | Amount of raise |
| StockoptionLevel | Has stocks or not |
| TotalWorkingYears | Number of working years total |
| TrainTimesLastYear | Number of Trainings |
| WorkLifeBalance | Satisfaction with work/life balance |
| YearsAtCompany | Years at current company |
| YearsInCurrentRole | Years at current role |
| YearsSinceLastPromotion | Time pasted sine last promotion |
| YearsWithCurrentManager | Years with Manager |

**Data Explanation**

Finding data for this problem wasn’t easy as private companies don’t like to make their data public for various reasons. I came across this dataset at IBM that a few departments used for human analytics. The dataset is a combination of data tables. It looks like various attributes divided between company and demographic information. Demographic being age, gender, marriage, education level, degree field, etc. The company information is salary, title, hours worked, length of time with company. Dataset has 1471 rows and thirty-five columns.



**Methods**

Exploratory data analysis was done using python. Used typical python libraries for this process in pandas, numpy, matplotlib, seaborn. Fairly typical process for exploring the data. Modeling was handled with sklearn. Looked into fairness of certain features with IBM specific library. All the code was done in a jupyter notebook.

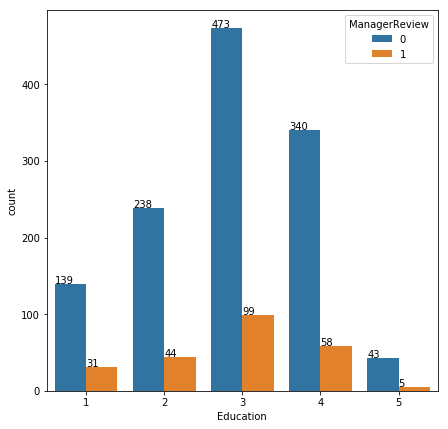
**Analysis**

After cleaning up the data and getting it into a tubular format, I decided to look at various distributions of the data around some of the features, starting with education and leaving a review.

Findings:

41% of employees having bachelor's degree are likely to leave a bad review.

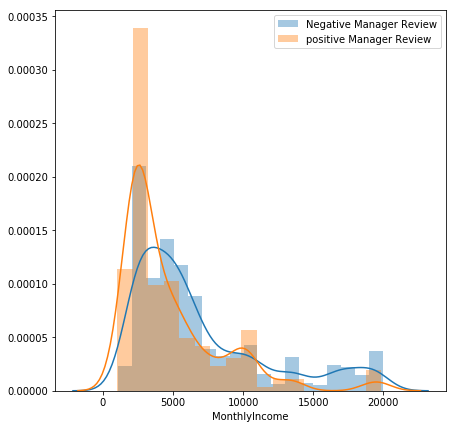
24% of employees having masters are the next in line to leave a bad review.



Monthly Income vs Manager Review

Findings:

It looks like people who are paid less are less likely to leave the company.



Business Travel

There are 3 categories in this:

No travel (0).

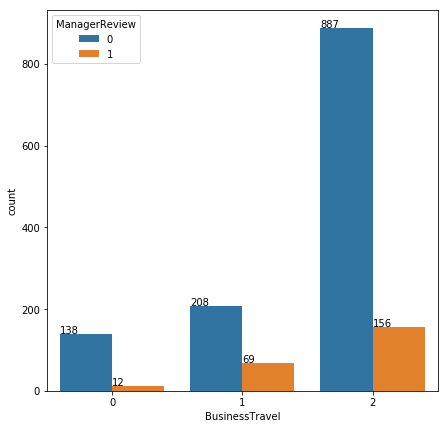
Travel Frequently (1).

Travel Rarely (2).

Bad Review: No = 0 and Yes = 1

Findings

Travel doesn't seem to be a factor in whether an employee leaves a bad review or not. Traveling rarely has the highest number of employees do it at 156. Percentage wise travel frequently has about a 25 percent negative review rate.



**Modeling Setup**

I decided to use Adaboost. Many weak and inaccurate classifiers are combined to produce a highly accurate prediction. The classifier is serially trained. Samples that are misclassified in previous round are given more focus. Initially weight is equal for all the samples. Weight of misclassified instances are increased each time and weight of correctly classified instances are decreased, this will let more misclassified sampled to be selected for the next round. After each classifier is trained, the weight is assigned to the classifier as well based on accuracy. More accurate classifier is assigned higher weight so that it will have more impact in outcome.

**Feature important**

MonthlyIncome 0.03467794373118038

OverTime 0.03161674257128877

Age 0.0273395112893986

JobRole 0.02666682541149834

MaritalStatus 0.021742815129764947

JobLevel 0.02097721981054601

WorkLifeBalance 0.01841327649386071

TotalWorkingYears 0.01823200187203544

JobInvolvement 0.015016626492141194

YearsWithCurrManager 0.014878309748851626

**Feature Important with maximum Chi-Square**

MonthlyIncome 127922.29369381821

TotalWorkingYears 230.72161773754925

YearsAtCompany 142.10005430324915

YearsInCurrentRole 117.5225958913567

YearsWithCurrManager 110.6715338985734

Age 84.15527681001525

OverTime 63.84506671452294

DistanceFromHome 63.77214163101213

JobRole 51.72130983521353

StockOptionLevel 25.26882603175403

**Conclusion**

I used the chi-square feature set in the end as I felt it was a bit more representative of the data. I ran a grid search for to find best estimators looking through 100-1000. Ended up using 100. The results are as follows:

Table

Description automatically generated with medium confidence

The model could predict up to 88 percent whether an employee would leave or not. Top three roles facing attrition:

26% of employees who are likely to quit belong to Laboratory Technician group.

24% of employees belong to Sales Executive group.

19% of employees belong to Research Scientist group.

The model developed will be able to predict whether an employee will stay or not. This will help company to know the status of an employee in advance and take necessary actions to prevent loss that will incur.

**Assumptions**

The data is representative of the company. It isn’t a big chunk of data, with IBM having around 400,000 employees. So, it would be ideal if the data was bigger.

The features are representative of the population.

**Limitations**

It is only a snapshot of data. Would be nice to see this over time more and compare if external factors like covid changed things.

Features not being largely correlated, only weakly.

**Implementation Plan**

This model could be deployed in production. Could be run on batch runs once a month and see if some employees are more likely than others to leave the company and set up some kind of plan to prevent that.

**Challenges**

There weren’t any challenges to this. One small one, I used to use a library called cufflinks and it doesn’t work anymore so that was a bummer.

**Future Uses/Additional Applications**

It gives a snapshot of how to retain employees and what factors lead to them leaving. It would be better if there was more of a time element to it.

**Ethical Assessment**

The lack of certain categories in gender is a issue now. Some other considerations probably should be considered like married or has kids shouldn’t be included in the dataset. So the dataset shouldn’t include info that could be used against various groups, or at least decisions need to be made with considerations to those issues.

**References**

<https://www.gallup.com/workplace/247391/fixable-problem-costs-businesses-trillion.aspx>

<https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/great-attrition-or-great-attraction-the-choice-is-yours>

http://oldtm.lbp.world/SeminarPdf/140.pdf