**Employee Attrition Predictor**

# **Index**

1. Overview of the Data Science project lifecycle
2. Machine learning problem framing
3. Data collection and integration
4. Feature visualization and analysis
5. Feature selection
6. Feature Engineering
7. Linear regression baseline model training
8. Other ML model feasibility evaluations
9. Finalised model tuning
10. Business goal evaluation
11. Prediction and story telling
12. Production deployment (AWS Sagemaker)
13. Model maintenance

# **Chapter 1: Overview of the Data Science Project Lifecycle**

Data science project lifecycle is completely different from conventional SDLC as we have a lot of maintenance to take care of.

Let’s have a look at the development cycle that I have followed here:

# **Chapter 2: Machine learning problem framing**

## *Business Requirement:*

## Predict the employee attrition using the HR data.

1. *ML Requirement:*

Attrition is the target variable represented as a Boolean field. Using features provided in the CSV file predict whether the employee will leave the organisation or not.

1. *Assumptions:*
2. Since we do not have test data separately, we will use the Train-Test Split from SciKit-Learn for generating the test data (used to compute the performance metrics).
3. Data is fictional and can result in a very low accuracy score- we will then opt for some-other metrics for model evaluation.

Exit Criteria:

The accuracy should be higher than 50% or linear model baseline – whichever is lower.

# **Chapter 3: Data collection and integration**

This has already been done by the HR department using survey from employees. We have a .csv file with all the possible properties required for this model.

## **Property description table:**

|  |  |
| --- | --- |
| **PROPERTY** | **Explanation** |
| Age | Age of employee |
| Attrition | Employee got separated |
| BusinessTravel | Travelling for business |
| DailyRate | Rate per day for employee |
| Department | Department working in |
| DistanceFromHome | Distance from home |
| Education | Level of education: |
| EducationField | Education field |
| EmployeeCount | Indicating that the record is for 1 person only |
| EmployeeNumber | Employee Number |
| EnvironmentSatisfaction | Environment Satisfaction rating |
| Gender | Gender |
| HourlyRate | Rate per hour |
| JobInvolvement | Job Involvement rating |
| JobLevel | Level of Job |
| JobRole | Job role |
| JobSatisfaction | Job Satisfaction rating |
| MaritalStatus | Marital Status |
| MonthlyIncome | Monthly Income |
| MonthlyRate | Monthly Rate |
| NumCompaniesWorked | Companies worked earlier |
| Over18 | Age over 18 |
| OverTime | Doing over time |
| PercentSalaryHike | Salary Hike |
| PerformanceRating | Performance Rating |
| RelationshipSatisfaction | Relationship with employer Satisfaction rating |
| StandardHours | Standard hours |
| StockOptionLevel | Company stocks owned |
| TotalWorkingYears | Working Years |
| TrainingTimesLastYear | Number of training taken last year |
| WorkLifeBalance | Work Life Balance rating |
| YearsAtCompany | Years with current company |
| YearsInCurrentRole | Years in current role |
| YearsSinceLastPromotion | Years since last promotion |
| YearsWithCurrManager | Years with current manager |

## 3.1. Data Cleaning:

Now we have understood the properties so we can go to the Data Cleaning before doing any visualizations

### Finding NaN values

### Cleaning the column values for clear understanding

* Business Travel – removing the additional “Travel” string from all the values
* Removing “Employee count” as it does not change for employees
* Removed “Employee Number” as it was unique and has no significance in model
* Removed the “Standard Hours” as it is same for all records

# **Chapter 4: Feature visualization and analysis**

## **Feature Description Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **FEATURE** | **DATA TYPE** | **CODE DATA TYPE** | **FEATURE TYPE** |
| Age | Real Number – not decimal | Decimal |  |
| Attrition | Yes or No | Boolean | Boolean |
| BusinessTravel | “Non-Travel”, “Frequently” or “Rarely” | Categorical | Nominal |
| DailyRate | Currency - decimal | Decimal |  |
| Department | “HR”,” R&D” and “Sales” | Categorical | Nominal |
| DistanceFromHome | Distance Unit | Decimal |  |
| Education | 1.'Below College'  2.'College'  3.'Bachelors'  4.'Master'  5.'Doctors' | Ordinal Category | Ordinal |
| EducationField | HR, Life sciences, medical etc | Categorical | Nominal |
| EmployeeCount | Count | Number | Useless |
| EmployeeNumber | Unique identifier for employee | Number | Useless |
| EnvironmentSatisfaction | 1 'Low'  2 'Medium'  3 'High'  4 'Very High' | Ordinal Category | Ordinal |
| Gender | Male or Female | Boolean | Boolean |
| HourlyRate | Currency | Decimal |  |
| JobInvolvement | 1 'Low'  2 'Medium'  3 'High'  4 'Very High' | Ordinal Category | Ordinal |
| JobLevel | Range from 1 to 5 | Ordinal Category | Ordinal |
| JobRole | Sales executive, manager etc | Categorical | Nominal |
| JobSatisfaction | 1 'Low'  2 'Medium'  3 'High'  4 'Very High' | Ordinal Category | Ordinal |
| MaritalStatus | Single, married or divorced | Categorical | Nominal |
| MonthlyIncome | Currency | Decimal |  |
| MonthlyRate | Currency | Decimal |  |
| NumCompaniesWorked | Count | Number |  |
| Over18 | Yes or No | Boolean |  |
| OverTime | Yes or No | Number |  |
| PercentSalaryHike | Hike in % | Decimal |  |
| PerformanceRating | 1 'Low'  2 'Good'  3 'Excellent'  4 'Outstanding' | Ordinal Category |  |
| RelationshipSatisfaction | 1 'Low'  2 'Medium'  3 'High'  4 'Very High' | Ordinal Category |  |
| StandardHours | Number of hours | Number |  |
| StockOptionLevel | number | Number |  |
| TotalWorkingYears | Number | Number |  |
| TrainingTimesLastYear | Number | Number |  |
| WorkLifeBalance | 1 'Bad'  2 'Good'  3 'Better'  4 'Best' | Ordinal Category |  |
| YearsAtCompany | Years associated with company | Number |  |
| YearsInCurrentRole | Years associated in current role | Number |  |
| YearsSinceLastPromotion | Years since last promotion | Number |  |
| YearsWithCurrManager | Years with current manager | Number |  |

## 4.1. Feature Selection: 1st Iteration:

I’ll attempt to predict using all features for the first time considering below properties for the categorical data.

For simplicity I have considered these as nominal for my first attempt

|  |  |
| --- | --- |
| **Property Iteration 1** | **Assumption** |
| Business Travel | Nominal |
| Department | Nominal |
| Education field | Nominal |
| Job Role | Nominal |
| Marital Status | Nominal |

Encoding these features:

1. Label encoding all the categorical data
2. One Hot encoding the nominal data
3. Ordinal data can be left as it is so that the value weights can be utilised

# **Chapter 5. Feature Engineering**

Major part at this stage is encoding the data as ML models require numerical data to work on.

## 5.1. Label Encoding

Using sklearn.preprocessing LabelEncoder we can encode the categorical data which in our case is [Gender, Over18, OverTime, BusinessTravel, Department, EducationField, JobRole, MaritalStatus]

## 5.2. One Hot Encoding

Using sklearn.preprocessing OneHotEncoder we