**MedView Pharmaceuticals CRISP-DM Project**

1 Business Understanding

1.1 Background

Our company, MedView Pharmaceuticals, is a medical research and sales company established in 1998. However, for the past 5 years there has been an increased employee churn rate in some of our departments and as such, the HR Manager has decided to intervene by deducing the cause(s) of this, as well as proffer a long lasting solution to this human resource flight by enlisting an algorithm that will help predict future churns so as to minimise turnover rates which can impact business continuity, increase hiring cost and affect the morale of the employees. This project is conducted using the company employee data from the HR department, detailing the demographic information of employees, as well as there salary, job satisfaction level and whether or not they are still with the company.

1.2 Key Stakeholders

1. HR Manager: Help the Human Resources identify the reason for employee attrition, predict and prevent it.
2. Management and Executives: Help the management in making sound and informed decisions on the retainment of employees.
3. Employees: This project will help improve employees’ morale and job satisfaction.
4. Data science team

1.3  Business Goals

The business goal is determined together with other department. This part should be the continuation of the background problem.

1. Gain insight on what leads to employee churns
2. Predict future churn prospects with at least 75% Accuracy, precision, recall, and F1-score
3. Prevent future employee attrition from data driven insights

1.4 Data Mining Goals

The data mining goal is determined by the data mining team and is a translation from the business goals.

1. Build predictive model with a minimum of 75% accuracy
2. Build predictive model with a minimum of 75% precision
3. Build predictive model with a minimum of 75% recall

1.5 Project Feasibility

The business problem is well suited for the analytics as a historical data of the employees is available. This data is credible and devoid of noise. Also, the predictive model in use is able to identify the actual factor responsible for the rise in employees churn in recent times. Lastly, the Human Resource department can and is poised to act on the insights from the findings to improve the retention rates. That is, the insights are actionable.

1.6 Redefining the Problem Statement

With regards to the project problem, there seems to be just the right amount of data available, however, constraints such as potential biases in data have been factored in, there are no legal issues and employee privacy is maintained and the project is in line with the data privacy policy of the company.

The redefined problem: “Predictive model for determining employee churn risk from HR historical data while maintaining ethical use of employee information”

1. Business benefits:  
   Reduction in the cost of hiring new staff to replace outgoing employees.
2. Improve staff morale and sense of job satisfaction.
3. Facilitate proactivity in HR decision making in the area of staff retention

2 Data understanding

In this phase, we collect, describe and explore the data for the project so as to determine its feasibility with regards to solving the business problem. The results of this phase will include:

1. Data description
2. Data exploration report
3. The data exploration report

2.1 Analytics Problem Framing

Predictive classification: With the given employee data, predict whether or not an employee will leave (Yes/No).

ABT (Analytics Base Table) Development with Attributes and Outputs

1. Attributes: Employee ID, Age, job role, department, education, job satisfaction, work-life balance, years at company, salary, etc.
2. Output: “Past Employee” (Yes/No) is the target variable.

A graph with blue squares

Description automatically generated

2.2 Project Assumptions:

1. Employee churn is a function of multiple variables such as job satisfaction, salary, etc.
2. The available data is representative of the entire staff.
3. Other factors such as external economic conditions are not considered in the analysis.

2.3 Key Success Metrics

1. The Accuracy of the classification model.
2. Precision and Recall so as to ensure correct predictions.
3. F1 Score for a balanced evaluation
4. Impact on the business: The reduction in the employee churn rate after model implementation.

3. Data Understanding and Preparation

The given data set contains 1470 employee records with 28 distinct features and is sourced from the Human Resource databases, performance records and employee surveys. The data set is available as “MedViewPharmChurn.csv”

3.1 Visual exploration of data

1. The distribution of job satisfaction, work-life balance, income, etc.
2. Correlation between features and Churn.

Data Pre-processing: Data cleaning and rescaling is done to handle missing values (if any) as well as convert categorical variables into numbers and the rescaling of salary and other tenure-related feature.

Documentation and Report Findings

1. Summary statistics of the dataset
2. Insights on some key attrition related factors
3. Exposure of certain data quality issues such as duplicates and imbalanced data, etc.

Problem Statement redefinement: This is adjusted based on insights from the data exploration.

4. Methodology Selection

4.1 Identification of Potential Models and Approaches

Logistic Regression, Decision Trees, Random Forest, and XGBoost were selected as they are good tools for classification based predictive analysis and the evaluation metrices used to compare which model is best.

4.2 Software Tools

Python Libraries and softwares:​

Pandas​, Scikit-Learn​, MatplotLib​, Numpy​, LogisticRegression​, DecisionTreeClassifier​, RandomForestClassifier​, XgBoost, etc, and Auto-Model in Rapid Miner were imported and used to perform the analysis while being benchmarked by the Rapid miner Auto Model to determine the efficacy of the models.

4.3 Test Approaches

Models were tested for accuracy, precision, recall and F1-Score and results displayed in the subsequent sections.

4.4 Select Best Approach

Choose Model performance based solely on Accuracy metrics and business interpretability.​

Random Forest emerge with 96% Accuracy making it the best model by that metrix.

A graph with blue rectangular bars

Description automatically generated with medium confidence

**5. Model Building**

**5.1 Model structures identification:**

The project takes advantage of different predictive models to predict whether an employee would likely leave the company and the following models were evaluated:

Logistic Regression: This model provides a probabilistic interpretation and is useful with regards to binary classification projects and quite interpretable.

Decision Trees: This model helps in the identification of key decision rules and is also quite interpretable.

Random Forests: This model uses multiple decision trees and increases generalization while reducing the incidence of over-fitting.

XGBoost: Extreme Gradient Boosting offers faster training, regularization and a more advanced performance on structured dataset.

**5.2 Data Features Considered and utilised in the project:**

1. Demographics of the workers (e.g., age, gender, marital status)
2. Work history (tenure, job role, department)
3. Compensation of workers (salary level, raises, bonuses)
4. The Performance ratings
5. Promotions accorded
6. Work-life balance scores by workers
7. Overtime and leave patterns of the employees

**Running and the evaluation of the model:**

Our dataset (ChurnPredictionNEW.csv) was thus processes:

Data Processing:

Missing Values handling: There was no missing values in the dataset.

Encoded categorical variables with one hot encoding was performed.

Data scaling: the dataset contained well scaled data leaving no need for the standard scaler to be called upon.

**Splitting of the data:**

The dataset was split 80/20 train-test split for the proceeding evaluations.

Stratification to help preserve class distribution due to churn imbalance.

**Evaluation metrices used:**

Each of the models were evaluated based on the following:

1. **Accuracy**: This measures the overall correctness of the model, but is not sufficient due to the possible issues of class imbalance.
2. **Precision**: This helps determine how many of the churners predicted actually left.
3. **Recall:** This is a measurement of sensitivity and finds out the number of people who left, how many were correctly identified
4. **F1-**Score: This simply measures the harmonic mean of both precision and recall.

**(Note: other minor metrices was also added to the model for robustness such as MSE, RMSE, MAE AND R-squared)**

**Random forest has the highest Accuracy of 96%.**

**Overall metrices visualization:**

A screenshot of a graph

Description automatically generated

Model Cross Validation based on F1-score visualized:

A graph of blue rectangular bars

Description automatically generated with medium confidence

In the measurement of the harmonic mean of both precision and recall (F1-score) XGBoost was the best model as far as cross validation with 0.4259 mean score.

**Model and data calibration:**

Class imbalance was handled using oversampling of the minority category (‘pastEmployee[Yes]). The outcome of the data balancing is seen below:

**A graph of a graph with blue squares

Description automatically generated with medium confidence**

**Documentation and communication of findings:**

Assumptions: The project proceeded with the assumption that the features contained in the dataset are representative of the future behaviour of employees. It is known that the best predictor of future behaviour is past behaviour, ceteris paribus. With this assumption, we can predict that the relationships captured in the historical data will hold going forward.

Limitations: Factors that were not captured during the project include external factors such as personal life events, market trends, etc, and behavioural information like survey feedbacks, emails, etc, were also not included.

Performance summary:

A comparison of the Confusion matrix of each model is visualised below. Also, the Feature importance ranking is displayed below showing that Overtime is ranked highest.

Confusion matrices

A chart with numbers and labels

Description automatically generatedA chart with different colored squares

Description automatically generatedA chart of a forest confusion matrix

Description automatically generatedA chart of confusion matrix

Description automatically generated

Random Forest has the best confusion matrix label prediction outcome.

Feature importance ranking

A graph of a number of blue bars

Description automatically generated with medium confidence

Feature importance indicates that Overtime, Job level, Stock option level and job role are key predictors of employee churn and should be paid ample attention to by HR as they have more influence on employee attrition.

**6. Deployment**

**6.1 Perform Business Validation**

Model predictions were compared with the HR insights in order to verify practical relevance.

Steps taken

1. Prediction for a recent quarter was anonymously shared with the HR.
2. The HR reviewed the files and gave feedback on the predicted risk.
3. Over 80% of the high risk predictions actually matched the HR concerns.

**Pilot Test**

1. A **pilot program** was conducted with a particular department (approx. 100 employees).
2. Intervention offered to top 10% of the high risk employees.
3. Exit interviews were monitored for the following months to gauge predictive accuracy.

**Report & Recommendations**

**Final Deliverables:**

Having seen the performance of different models (Logistic Regression, Decision trees, Random Forests, and XGBoost) with python code, below is a visual of the evaluation and model efficiency of the same models but using Rapid miner Auto modelling tool:

Note: The models performed better with the python code on Colab Notebook, than with the Rapid miner Auto model.

Auto Model Outcome:

A screenshot of a computer

Description automatically generated

While Linear regression model has the fastest runtime, XGBoost was the best performing model with the best gain (46) and lowest classification error (11.7%).

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Colab Notebook outcome:

A screenshot of a graph

Description automatically generated

Comprehensive technical report:

Overall, Random Forest is the best model with the highest Accuracy of 96%, Precision 95%, Recall 98%, F1-score 96%, and confusion matrix label prediction outcome, even though XGBoost is the highest in the F1 Cross validation mean score according to the python code.

With regards to the Auto model, XGBoost is the best model with highest gain (46) and lowest error (11.7%), Accuracy 88.3, Precision 73.8, Recall 50.1, f1- score (sensitivity) 50.1, only seconded by Logistic regression model.

Churn probability and prediction:

A graph with blue and white text

Description automatically generated

The model produced a close correlation between Employee churn (pastEmployee) with churn probability and prediction.

Distribution of predicted churn probabilities:

A diagram of a distribution of churn probabilities

Description automatically generated

A diagram of a number of blue dots

Description automatically generated

**Recommendations to HR:**

HR should focus on the top three influences on employee attrition viz: Overtime, Job level and Stock option level. Discussions should be had with the employee about how these would favour them the most and with issues other them with regards to these factors.

There should be targeted interventions which can be enforced by scheduling a one-to-one meetings with employees that are flagged as high-risk.

Also, HR should create retention incentives by offering flexible working and learning opportunities to staff members.

And lastly, culture check-ins such as pulse surveys should be conducted for the flagged teams.

**Ongoing Monitoring Plan:**

As a proactive step going forward, the model should be retrained at least quarterly with updated employees dataset and re-evaluated for performance that match up to the benchmark of this current model. Also, HR need so ensure that churn risk score is integrated into its dashboard for proper monitoring. Finally, an alarm system should be set up for when risk crosses a set threshold.

Conclusion: Using the CRISP-DM model, the data science team has succeeded in building a set of (4) models both in Colab notebook using python code, and another with Rapid miner Auto model, to help the company MedView pharmaceuticals analyse, predict and give insights to prevent further employee attrition which is the goal of the HR manager, with a significant degree of accuracy above the threshold KPI. The python models outperformed the Auto model massively.