Washington State University

Data Science Project Report

Uncovering "The Great Resignation"

Submitted by:
Priyanka Gosh Dastidar - 011804812
Nithyashree Senguttuvan - 011808333
Kunal Sanghvi - 011809708

January 25, 2023

Contents

1 Abstract

Beginning in early 2021 in the wake of the COVID-19 epidemic, employees voluntarily left their positions in large numbers, a phenomenon known as "The Great Resignation". We intend to present some proof and some explanations for this after evaluating the data we have found. Many have theorized as to why people have grown more likely to leave their existing companies. We will be concentrating on the American workforce for our initiative, particularly in the industrial, healthcare and hospitality, and educational fields. Many people have thought about leaving their jobs at some time. We would want to consider why as we continue with our analysis. Was it the long commute, the low income, or something else? Was the workplace unhealthy? We will look at a handful of the many causes for resignations that exist. We have been discussing this issue as a group since the start of the Great Resignation as it has inspired each of us to explore a different line of work and because we have all been directly or indirectly impacted by this economic trend. We will be able to anticipate future economic trends more accurately if we are aware of the variables that led to the Great Resignation.

Employee attrition counts the number of employees who leave their positions voluntarily or involuntarily for various causes. A widespread exodus from labor during the COVID-19 epidemic is called "the Great Resignation" today. Although there are many contributing elements, they may be divided into the Five R's, which are five categories - Retiring, Relocation, Reconsideration, Reshuffling, and Reluctance. During the epidemic, the number of older workers retiring increased. People move from major towns to smaller ones in search of a less stressful existence. People abandon their employment because their ideals do not match those of the organizations and toxic work environments. Better pay, a better work environment, business culture, and flexible scheduling are all things that employees are looking for. During the epidemic, working from home has become the new standard.

2 Introduction

There have been instances of susception that many people have stayed put in their jobs during the pandemic, were just waiting for the right time to make their exit from the organization. With that intent in mind, we tried to uncover the cause that might have led to attrition. From all datasets found, we considered US Bureau of Labor Statistics data, to evaluate the possible cause of the attrition and how, with our findings, companies could formulate strategies to curb employee turnovers. When an employee exits from a company, it becomes hard for the business. To backfill their position, the company incurs costs to hire a new employee like recruiter fees, interviewing time and training them. When a position gets vacated, the company also pays for the cost of lowered productivity and that is not it. New hire will need time to reach the existing employee's level of productivity and their proficiency. If we talk numbers, the cost that company incurs to replace an employee paying $20,000 a year starts at 10% of their annual pay and to backfill a senior employee, company could end up paying around 200%. So, to uncover the reason behind "The Great Resignation" becomes a particularly key factor. Through our approach we hope to clarify questions like: -

1. What could be the factors contributing to an employee's turnover?
2. Could we draw a correlation between attrition and age, job satisfaction, hikes, number of years working etc.?
3. What will the future trend to resignation look like?
4. Which departments will most be affected by resignation?

We have used several concepts from Data Science mainly: - Explanatory Data Analysis (EDA), Visualizations, Supervised Learning, few concepts from Machine Learning etc. There were several ways to reach a conclusion, on data gathered and making the correct prediction from data. We have used EDA to analyze dataset, derive its characteristics and to be able to understand any patterns within the data. Furthermore, it helped us detect outliers and anomalous events.

3 Problem Definition

The problem which we are trying to focus on is the great resignation scenario. The most often reported reasons for leaving a job are income stagnation despite rising living expenses, few chances for career progression, hostile work environments, a lack of benefits, rigid policies regarding remote work, and persistent job discontent. Seeing these problems and the increasing attrition rate, the team decided to uncover the great resignation scenario. To answer all these problems, we decided to implement a solution that includes a predictive model to foresee the probable future occurrence of the event and the extent to which it could affect an organization. Furthermore, it would take stances at data such as: -Time charge of employees who have resigned, their designation, performance, compensation, pay increased or not, and training opportunities provided to those employees. With insight from the predictive model and visualizations in hand, the organization can predict the possible risk of attrition beforehand. It will enable them to form strategies to hold back employees, thereby reducing employee turnover. So, uncovering correlations derived from data will benefit organizations across the globe.

4 Models

To train any model, the most important thing which is required is the dataset. Also, the dataset should be structured, and should not have any missing values or NA entries along with relations should be there between all the different tables in the dataset so that there is some meaningful outcome. The most difficult part was to find an appropriate set that has all the possible values along with necessary fields which could help us to do a complete prediction on the great resignation scenario. After exploring for a week or two we finally found an appropriate data set from the US Bureau of Labor Statistics which contains the attrition dataset till 2022. We started to explore the dataset and we discovered that it has many entries linked with multiple different tables/datasets. We started our research again to find some dataset which is a subset of the dataset present in the US Bureau of Labor Statistics to gain some explanation of the huge dataset in front of us. Then we came to a website where we could find a reasonable explanation about the dataset present on the official website and the explanation was related to the

2015 data present. Therefore 2015 dataset was too huge, and we used a subset of the complete dataset to train the model for testing purposes in the future we can increase its data till 2022 and not overfit the data to the model.

The tables created are mentioned in the figure below, which contain the data used to train the model.

File name	Description
data_dictionary	Definition of all variables available for study
employee_survey_data	Employee survey inputs
manager_survey_data	Manager survey inputs
general_data	Employee descriptive variables and attrition
in_time	Employee clock-in times
out_time	Employee clock-out times

Figure 1: Data types

- Data dictionary consists of all the definitions of the headings/variables which are present all over the tables/datasets. For eg: Age, Attrition, Business Travel, Department, Gender, Work-life balance, etc. Along with the variable names, there is the meaning of the variables along with the levels if any. Therefore, there is a complete description available of the dataset variables.
- Employee survey data table has four headings/variables Employee Id, Environment Satisfaction, Job Satisfaction, and Work-life balance. Each row just contains a number whose complete explanation is provided in the data dictionary file.
- Similarly, every table is linked in such a way that it can be easily explained with the data dictionary table.

Once the data is gathered then we need to process it in such a way that it is out of missing values and in proper shape to use. Therefore, we use data-wrangling techniques. Data wrangling is the process of cleaning, transforming, and organizing data in a way that makes it ready for analysis and visualization. This often involves working with large and complex datasets and may require a combination of manual and automated processes to ensure that the data is in a usable format. Some common tasks involved in data wrangling include filtering and selecting relevant data, handling missing or incomplete data, merging, and reshaping datasets, and converting data into a standard format. Data wrangling is a crucial step in the data analysis process, as it ensures that the data is ready for further analysis and interpretation. One of the ways we used data-wrangling is shown in the figure below where we have changed the orientation and combined the data of in and out time.

Now comes the exploratory data analysis part. Here we have done EDA in two diverse ways. First, we have done exploratory data analysis in terms of purifying the data and imputing the data if there are any missing values present and the later exploratory data analysis is done on the overall dataset to conclude on the prediction made by the model and the actual trend of the attrition is same or not. Therefore, there are two data analyses conducted in an exploratory way on the dataset.

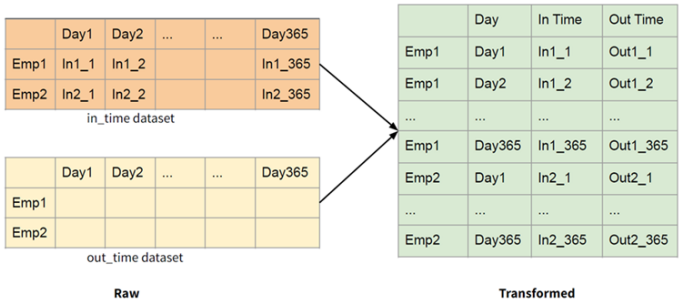

Figure 2: Data Wrangling

There were a couple of exploratory data analysis processes and feature engineering performed and the flow along with the findings are mentioned below:

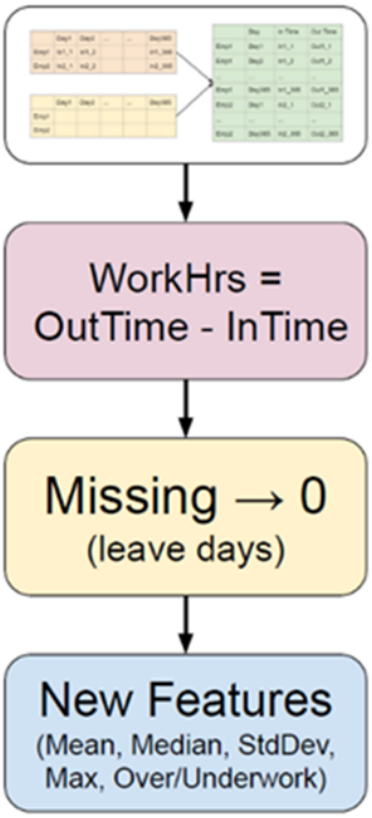

Figure 3: EDA

- Wide dataset layout with recorded work times per workday per employee for 2015
- Missing data in the dataset corresponding to public holidays or employee leaves.
- Standard hours in the general dataset can be compared with average actual employee work time to signal over or under work.

- Missing feature - Relationship Satisfaction, compared to the data dictionary

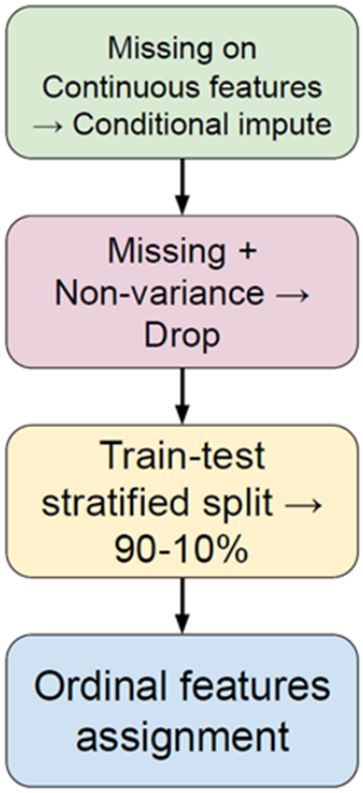

Figure 4: Modelling

- Features with missing values (fig [no.]) attributed to <1% of data distribution
    1. Ordinal + Continuous features
- Non-variance features - Employee Count, Over18, Standard Hours (fig [no.])
- Imbalanced distribution on the response variable, attrition.
- Ordinal features - Education and all features in survey datasets.

col	num_missing	pct_missing
EnvironmentSatisfaction	25	0.567
JobSatisfaction	20	0.454
WorkLifeBalance	38	0.862
NumCompaniesWorked	19	0.431
TotalWorkingYears	9	0.204

Figure 5: Missing values

Once the data wrangling is done on the dataset and the dataset is ready, now we can go towards the preparation of the model. Here we have used the pycaret library in python to identify the best model among a bunch of models and a baseline model is generated.

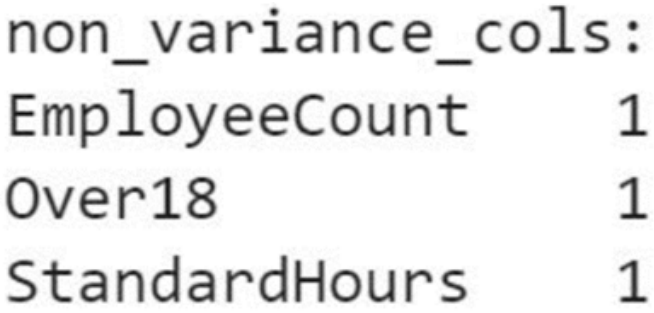

Figure 6: Non variance features

5 Implementation

5.1 Model

Pycaret is a popular open-source machine-learning library in Python that makes it easy to train, compare, and deploy machine-learning models. It is commonly used for tasks such as classification, regression, clustering, and natural language processing. One of the advantages of PyCaret is that it provides a user-friendly interface that allows users to build and evaluate machine learning models without needing to have a deep understanding of the underlying algorithms quickly and easily. This makes it an excellent choice for data scientists who want to quickly prototype and test different models without spending a lot of time on coding and debugging. The flow of the pycaret model training is shown in the figure below:

Pycaret trained around 16 models with 10 folds and holdout data for testing and based on the F1 metric the best model selected is the extra trees classifier along with the highest accuracy. The output of all the models based on the parametric values are shown in the figure below:

All the best values in the individual parameter are marked in yellow color for all the models. Seeing the most in the extra tree classifier model is selected as the best model and saved as a .pkl file. This model also produces the confusion matrix for train data, test data and validation data and it is depicted in the following figure:

Thus, this model is further used for prediction of attrition to obtain results.

5.2 Visualization

Data visualization is the process of presenting complex information using simple diagrams and charts. Because of this, it can create stories based on data while allowing people to recognize the many patterns and correlations that have been found in the data, which are nothing more than the insight obtained from the data. Following that, these insights are used to solve problems. It makes it easier for data analysts and scientists to analyze data and come to smart conclusions. It is the best way to get information over to non-technical people so they can comprehend it and draw judgments based on the evidence.

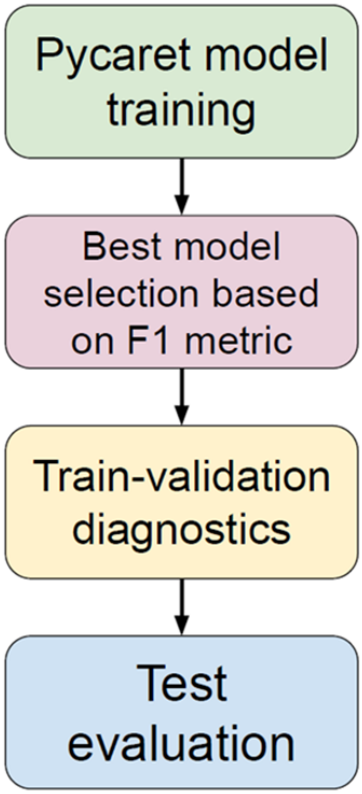

Figure 7: Model

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
et	Extra Trees Classifier	0.9757	0.9859	0.8615	0.9898	0.9200	0.9058	0.9096
gpc	Gaussian Process Classifier	0.9631	0.9313	0.8547	0.9189	0.8837	0.8619	0.8640
rbfsvm	SVM - Radial Kernel	0.9001	0.9835	0.3938	0.9836	0.5633	0.5191	0.5921
rf	Random Forest Classifier	0.8879	0.9451	0.3467	0.9510	0.5031	0.4552	0.5315
qda	Quadratic Discriminant Analysis	0.8293	0.7926	0.4697	0.4851	0.4753	0.3738	0.3749
nb	Naive Bayes	0.8326	0.7263	0.2686	0.4873	0.3451	0.2585	0.2741
lda	Linear Discriminant Analysis	0.8473	0.7966	0.2346	0.6001	0.3353	0.2674	0.3062
knn	K Neighbors Classifier	0.8197	0.7992	0.2353	0.4213	0.2969	0.2044	0.2175
ridge	Ridge Classifier	0.8451	0.0000	0.1025	0.6746	0.1770	0.1419	0.2208
lr	Logistic Regression	0.8381	0.7122	0.0649	0.3552	0.1096	0.0828	0.1211
mlp	MLP Classifier	0.7033	0.5330	0.2022	0.0835	0.0616	0.0036	0.0138
dt	Decision Tree Classifier	0.8293	0.5000	0.0089	0.2091	0.0148	0.0019	0.0215
ada	Ada Boost Classifier	0.8241	0.5720	0.0089	0.0105	0.0096	-0.0066	-0.0066
svm	SVM - Linear Kernel	0.8352	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
gbc	Gradient Boosting Classifier	0.8352	0.6476	0.0000	0.0000	0.0000	0.0000	0.0000
lightgbm	Light Gradient Boosting Machine	0.8352	0.5128	0.0000	0.0000	0.0000	0.0000	0.0000

Figure 8: Accuracy rates

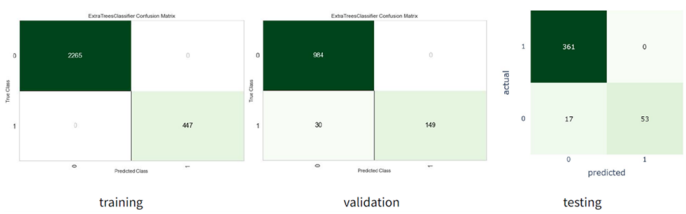

Figure 9: Confusion Matrix

5.2.1 Visualization for Model

Understanding the meaning of the data is made simpler through data visualization. Using this, research is performed and based on the findings, opinions are brought up. Thanks to visualization, anybody may perform exploratory data analysis on accessible datasets. It makes explicit the applications of each notion. They enable the detection of anomalies and correlations between data and assist us in making better judgments.

Survey Data

About 1400 of the workers were surveyed to get the survey data, and the results were used to generate six subplots.

• Education Levels: Undergraduate, College, Bachelor, Master, and Doctoral
• Work Involvement
• Job Satisfaction
• Environment Satisfaction
• Relationship Satisfaction: Low, Medium, High, and Very High:
• Work/Life Balance: Bad, Good, Better, and Best

Exploratory Data Analysis

For the exploratory part of our study, we utilized Matplotlib and the graph style from FiveThirtyEight to analyze historical data about the three industries that were most and least affected by the Great Recession.

5.2.2 Geospatial Visualization

Geospatial visualizations focus on the relationship between location and data to convey knowledge. Any positional information that might disperse the location of the data is helpful for spatial analysis. The key difference between geographical representations is scale. A circuit diagram on a microchip may explore position, but it is not geographical because it does not correspond to Earth or another planetary body. A map of Saturn's surface, for example, uses latitude and longitude to overlay variables on a map to gain understanding, while a map of the stars is not considered to be a geographical representation. Maps are the primary focus of geographic visualizations. In addition to acting as a container for more information, they let people change their visual focus by introducing context via shapes and color. They aid in identifying issues, keeping track of changes, comprehending trends, and predicting for certain areas and times. The real connections between the data items are shown through geospatial representations. The viewer is thus susceptible to two common errors: autocorrelation and scaling, as their perception of the data may change due to changes in the map's scale. Even with unconnected data, a perspective may generate a link between data points that are close to one another on a map.

Mapping

The Labor data has been mapped over the country United States of America, using

the folium package, an interactive map package that resembles Google Maps features. The data has been filtered based on years and description that brings in the mapping of states on the map with high attrition rates for a particular year and sector.

6 Analysis

To derive correlation between data we have used Tableau. Tableau being simple and easy to use with its simple dragging and dropping technique. Analytics are powerful and the user experience is very intuitive. Also, another reason to use tableau is that we had the license for it. Below were the observations gathered from the dataset using EDA

- The first dashboard displayed the correlation of attrition with: -

1. Age – From the dataset, we tried to figure out if there happens to be a relation of a person's age with attrition. From the visualization we could see that people in the age bracket between 18 to 30 years and 31-43 years are more susceptible to attrition. People belonging to the age bracket of 57 years and above are less likely to contribute to attrition.
2. Total Working Years- From the pie-chart it's clearly visible that people working between zero to five years and six to ten years are contributing more to attrition than people who have more years of working experience Industries.
3. Percent Salary Hike – We could notice from the statistical graphic that employees who got salary hike in range of 10%-14% are contributing more to attrition. Perhaps, they have the quest to earn more and are less satisfied with their pay range as compared to employees getting more salary hikes.

- The second dashboard displayed the correlation of attrition with:-

1. Department – Department plotted against attrition, gives out the percentage of attrition from each department. From the pie-chart we could see Research and Development along with sales contributes more to attrition than Human Resources.
2. Job Role – Even Job Roles have a significant influence on attrition rate. Laboratory Technicians and Sales Executives are contributing more to attrition. Other Job Roles are less likely to contribute towards attrition.
3. Marital Status – From the pie chart we could establish the conclusion that single and married employees are dissatisfied with their job and hence are more likely to contribute towards attrition than divorced employees.

- The third dashboard displayed the correlation of attrition with:

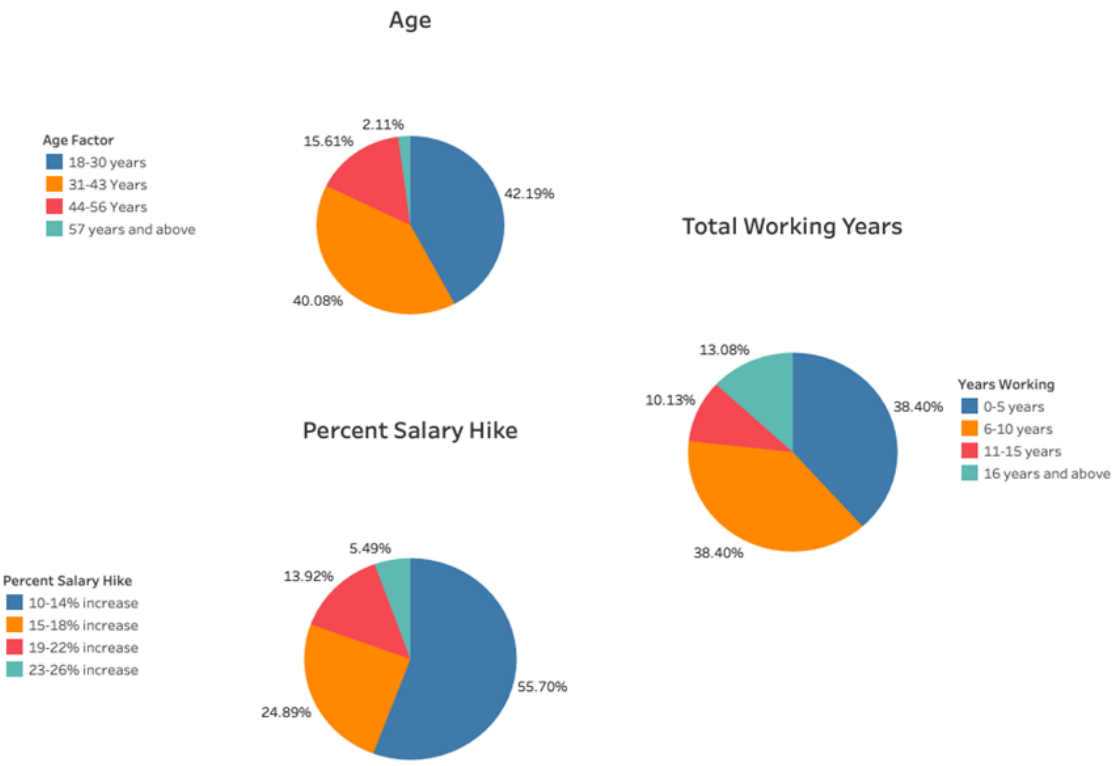

Figure 10: Tableau Analysis for first dashboard

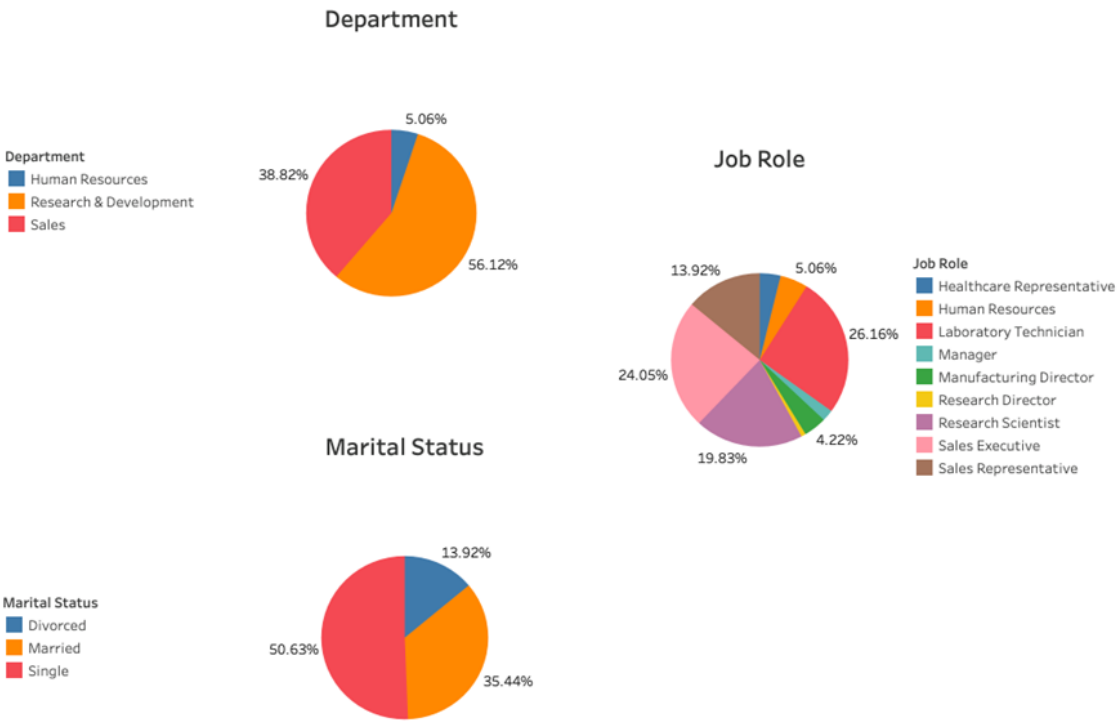

Figure 11: Tableau Analysis for second dashboard

1. Gender- As we could see Male employees significantly contribute more towards turnover than female employees. From, this we could derive that male employees have a higher tendency to opt for resignation than females.

2. Overtime – Another category contributing towards attrition. Although, the pie-chart suggests a marginal difference in attrition between employees who did overtime with employees that have worked normal hours in the office.

3. Travel Frequency – From the pie-chart, we could establish that people who traveled rarely and frequently are more likely to contribute towards attrition than the rest of the employees.

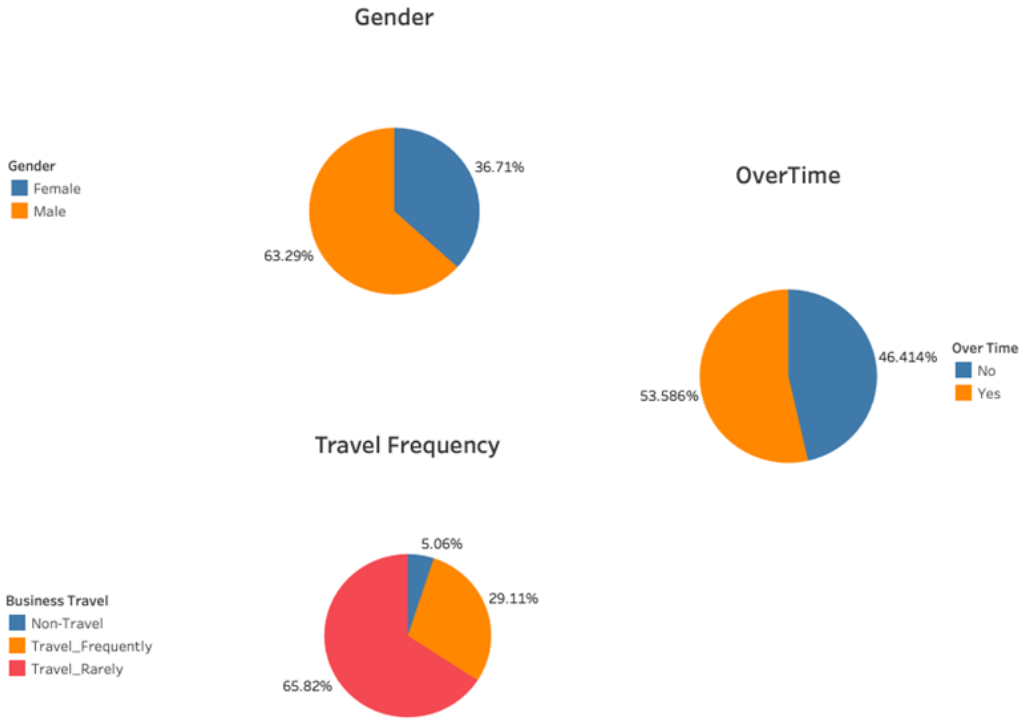

Figure 12: Tableau Analysis for third dashboard

7 Results and Discussion

7.1 Model

The pycaret trains a total of 16 models with the dataset which is improvised using data-wrangling methods and several parameters are generated by pycaret for those models such as Accuracy, AUC, Precision, Recall, F1 Score, Kappa, and MCC. We have given the most important to the F1 score and then we check the accuracy of the models, thus out of those 16 models which pycaret trained with a 10-fold we can clearly observe that

the extra trees classifier has the highest F1 score along with the accuracy. Therefore, in this way, we select the best model now, yet we can implement further feature engineering and hyperparameter tuning to increase accuracy. From the above image we can

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
et	Extra Trees Classifier	[illegible]	[illegible]	[illegible]	0.9898	[illegible]	[illegible]	[illegible]
gpc	Gaussian Process Classifier	0.9631	0.9313	0.8547	0.9189	0.8837	0.8619	0.8640
rbfsvm	SVM - Radial Kernel	0.9001	0.9835	0.3938	[illegible]	0.5633	0.5191	0.5921
rf	Random Forest Classifier	0.8879	0.9451	0.3467	0.9510	0.5031	0.4552	0.5315
qda	Quadratic Discriminant Analysis	0.8293	0.7926	0.4697	0.4851	0.4753	0.3738	0.3749
nb	Naive Bayes	0.8326	0.7263	0.2686	0.4873	0.3451	0.2585	0.2741
lda	Linear Discriminant Analysis	0.8473	0.7966	0.2346	0.6001	0.3353	0.2674	0.3062
knn	K Neighbors Classifier	0.8197	0.7992	0.2353	0.4213	0.2969	0.2044	0.2175
ridge	Ridge Classifier	0.8451	0.0000	0.1025	0.6746	0.1770	0.1419	0.2208
lr	Logistic Regression	0.8381	0.7122	0.0649	0.3552	0.1096	0.0828	0.1211
mlp	MLP Classifier	0.7033	0.5330	0.2022	0.0835	0.0616	0.0036	0.0138
dt	Decision Tree Classifier	0.8293	0.5000	0.0089	0.2091	0.0148	0.0019	0.0215
ada	Ada Boost Classifier	0.8241	0.5720	0.0089	0.0105	0.0096	-0.0066	-0.0066
svm	SVM - Linear Kernel	0.8352	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
gbc	Gradient Boosting Classifier	0.8352	0.6476	0.0000	0.0000	0.0000	0.0000	0.0000
lightgbm	Light Gradient Boosting Machine	0.8352	0.5128	0.0000	0.0000	0.0000	0.0000	0.0000

Figure 13: Accuracy rates

conclude following points:

• The accuracy of extra tress classifier is the highest with 97.57%.
• The AUC is 0.9899 which means that the predictions are 98.99% correct.
• The recall is also around 0.8615 which is good positive prediction number out of all possible positive predictions.
• The precision is not the best where SVM (Support Vector Machine) – Radial Kernel takes over it.
• The F1 score is the highest 0.92 and we have given importance to this parameter the most.
• The Kappa and MCC are comparatively the best among all.

7.2 Python Code Visualization

7.2.1 Survey Data

The bar plot below and a few pie charts had been generated to explore the data of the features that could contribute to the model in accurate prediction of the cause.

- Education Levels: Employers have a “Below College” qualification and next to it is the “Bachelor” level. This results in people either resigning for better positions or salary.
- Work Involvement: Employers are involved at the actual expected rate of 2.
- Job Satisfaction: Some people are not satisfied with their job, which accounts for about 75% of people who are satisfied, which might be an issue.
- Environment Satisfaction: Similarly, to Job Satisfaction, few people are not satisfied with their job, which accounts for about 75% of people who are satisfied, which might need an improvement in the environment.

- Relationship Satisfaction: Though the vast majority agree that this is High, and Very High, we should investigate people with low and medium for ways of improvement.
- Work/Life Balance: Many people believe that it has been Better over the years.

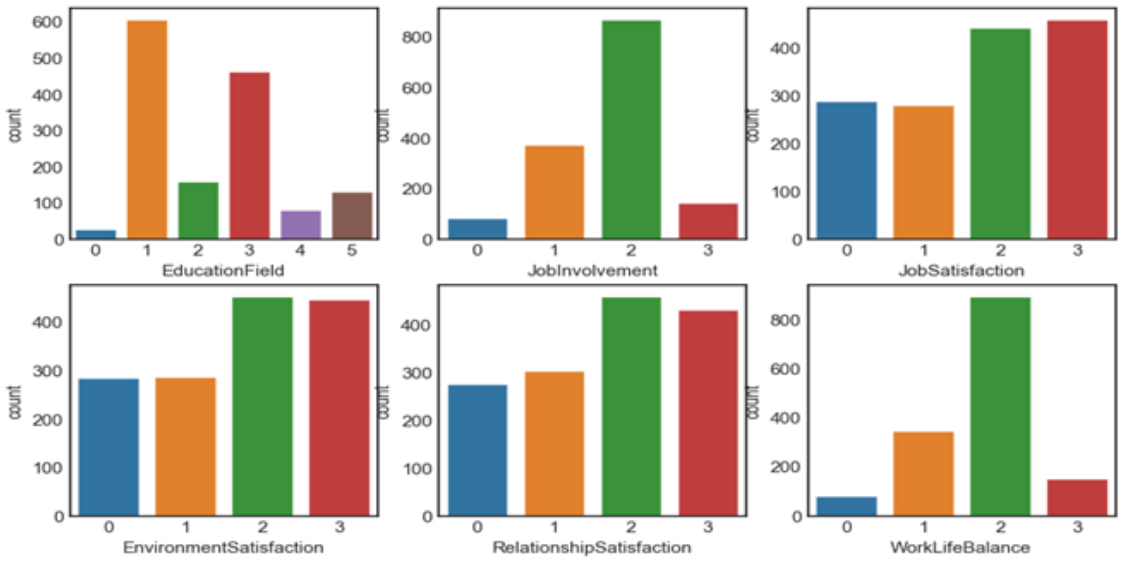

Figure 14: Bar chart

7.2.2 Exploratory Data Analysis

Manufacturing, education and health services, and hospitality are the sectors that have been most impacted throughout this time. Except for manufacturing, all three industries suffered severely during the Great Resignation, and it has only been since the beginning of 2022 that they have begun to steadily recover. Finance, IT, and professional and

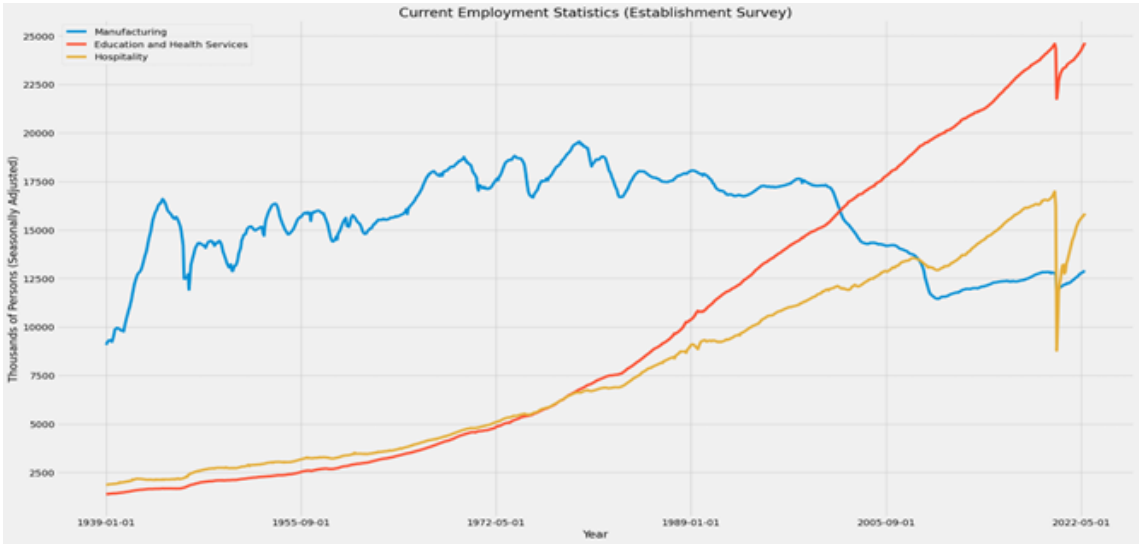

Figure 15: Most affected

business were the industries least impacted during this time period. All three, except

for professional and business, had a modest decline at the beginning of the pandemic, but it rapidly recovered thanks to factors like the increase in remote work.

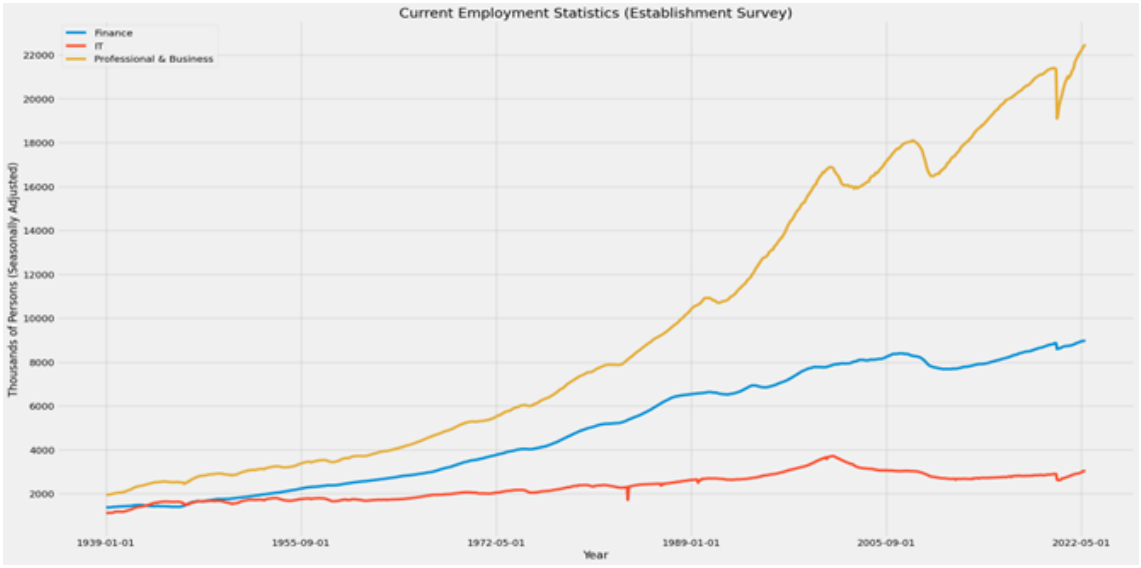

Figure 16: Least affected

7.2.3 Geospatial Visualization

The following image displays the overall attrition rate in all states of US, based on the data. This is not useful much as it shows all the states, spread out. The following visu-

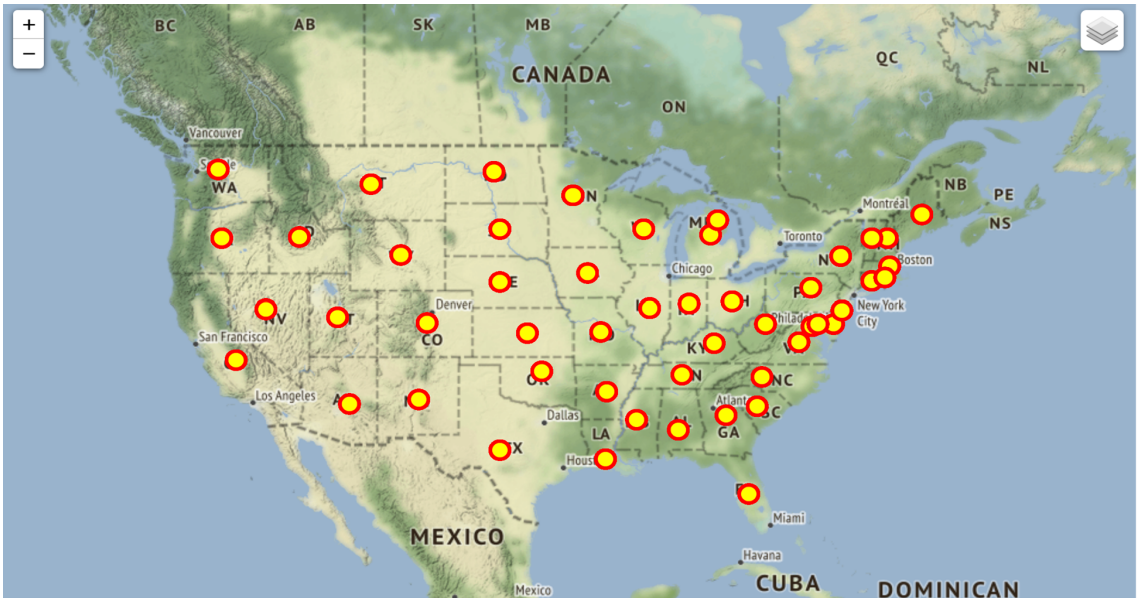

Figure 17: Overall attrition rates over states

alization is on Government and government enterprises in the year 2016 with attrition rates greater than 50000. The following visualization is on Private employment - Health care and social assistance on all the years with attrition rates greater than 50000.

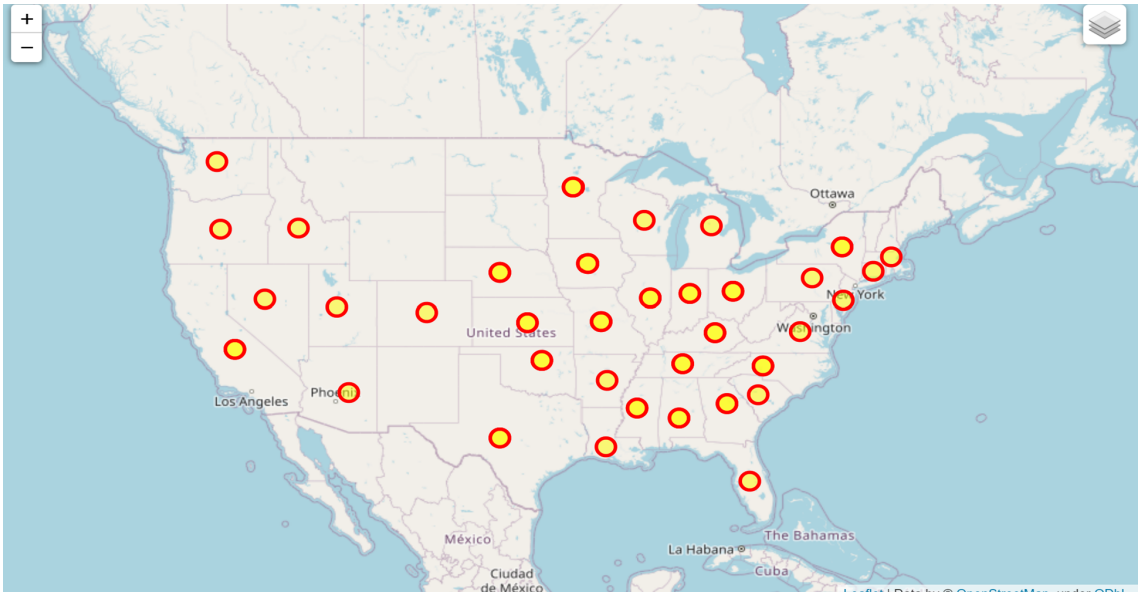

Figure 18: 2016 > 50000

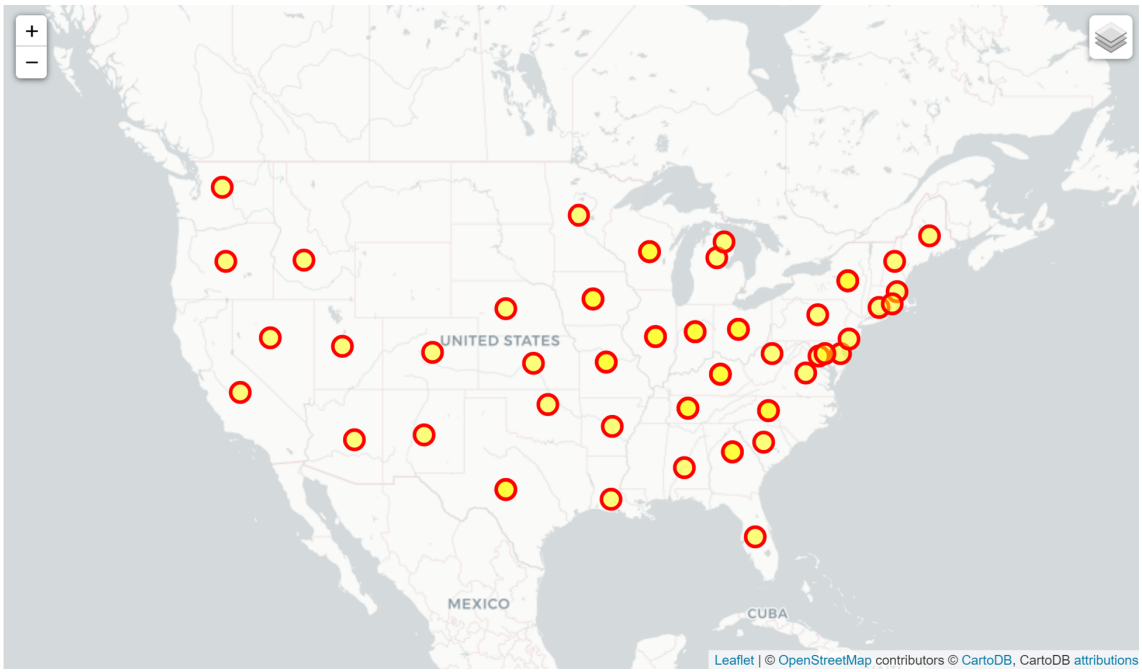

Figure 19: > 50000

8 Related Work

The future improvement for the model can be done is feature selection and hyper parameter tuning which can further increase the accuracy of the baseline model and can also help us with a great prediction power for the attrition.

With respect to visualization, a more interactive feature map with a different package in Python may have been incorporated to have options to choose from.

9 Conclusion

The best model generated by pycaret is saved as a .pkl file therefore it can be loaded later, and this method saves time that is spent training the model. Also, there is always improvement available for tuning the model, yet we have achieved a 97.57% of accuracy without categorical feature engineering and hyperparameter tuning.

The visualization by Python code has supported the model in identifying the features and spread on the attrition rates over the years in the states of US. Visualizations done in Tableau gives us the correlation of attrition with other related categories. With an insight in hand, an organization will be able to predict beforehand the possible risk of attrition. It will enable the organization to derive mention strategies to hold back employees thereby reduction employee turnovers. So, uncovering correlation derived from data will benefit the organization.

9.1 Code Link

https://github.com/nithyashreesenguttuvan/Data-Science-Project.git

References

[1] https://help.tableau.com/current/pro/desktop/en-us/actionsdashboards.htm

[2] https://www.bls.gov/data/

[3] https://www.workstream.us/blog/hiring-and-firing-statistics