|  |  |
| --- | --- |
| D:\General Documents\Logo's\Updated logo 2021.jpg | **SIDDAGANGA INSTITUTE OF TECHNOLOGY, TUMAKURU**  **(An Autonomous Institute Affiliated to Visvesvaraya Technological University, Belagavi, Approved by AICTE, New Delhi, Accredited by NAAC and ISO 9001:2015 certified)**  **Department of Computer Science & Engineering**  **(Program Accredited by NBA)** |

**Industrial Training report**

**On**

**“MACHINE LEARNING”**

Submitted in Partial Fulfilment of the Requirements for the Award of Degree of

**Bachelor of Engineering**

**in**

**COMPUTER SCIENCE & ENGINEERING**

Submitted by

**Mohammed Sayeed**

**1SI18CS057**

Training carried out

at

INTERNSHALA

|  |  |
| --- | --- |
| Convener | Industry Expert |
| Dr. Y. S. Nijagunarya | Kunal Jain |
| Professor | Founder |
| Department of CSE | Analytics Vidhya |
|  |  |

**Department of Computer Science and Engineering**

**2021-2022**

|  |  |
| --- | --- |
| D:\General Documents\Logo's\Updated logo 2021.jpg | **SIDDAGANGA INSTITUTE OF TECHNOLOGY, TUMAKURU.**  **(An Autonomous Institute Affiliated to Visvesvaraya Technological University, Belagavi, Approved by AICTE, New Delhi, Accredited by NAAC and ISO 9001:2015 certified)**  **Department of Computer Science & Engineering**  **(Program Accredited by NBA)** |

**CERTIFICATE**

It is certified that the Industrial training work entitled **Machine Learning** carried out by **MOHAMMED SAYEED**, USN**: 1SI18CS057**, a bonafide student of **Siddaganga Institute of Technology,** in partial fulfilment for the award of **Bachelor of Engineering in** **Computer Science** **and Engineering** of the Visvesvaraya Technological University, Belagavi during the year 2021-2022. It is certified that all corrections / suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library. The industrial training report has been approved as it satisfies the academic requirements in respect of Industrial training / Professional practice work prescribed for the said Degree.

**Dr. A. S. Poornima**

Professor & H.O.D

Dept. of C.S.E

SIT, Tumakuru

External Viva

Name of the examiners Signature with date

1. S P Gururaj 1.

2. 2.

**COMPANY CERTIFICATE**



**ACKNOWLEDGEMENT**

The experience I had in the internship has been of great value in terms of learning and industrial experience. I am also grateful for having a chance to meet so many talented professionals who led me through out that time.

I take this opportunity to express my deepest gratitude and special thanks to the **Internshala** for the training.

I express my deepest thanks to my internal guide, **Dr. Shobha K, Assistant Professor, Dept. of CSE,SIT,Tumakuru** for his guidance and giving necessary advice to prepare this report.

I also express my profound thanks to the **Dr A.S. Poornima, Head of the Department**, **Dept. of CSE, SIT, Tumakuru,** for allowing me to take industrial training at Daedal Crafters. I express my special thanks to our **Dr.S.V Dinesh, Principal of SIT**,Tumakuru.

I am fortunate for being part of SIT, the institution that stood by my side and assisted me towards successfully completing the industrial training. I also take this opportunity to thank the **Department of Computer Science and Engineering, SIT,Tumakuru,** for their tremendous help and guidance.

I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

By

**Mohammed Sayeed**

**1SI18CS057**

**ABSTRACT**

Industrial training is an important phase of a student life. A well planned, properly executed, and evaluated industrial training helps a lot in developing a professional attitude. The aim and motivation of this industrial training is to receive discipline, skills, teamwork, and technical knowledge through a proper training environment, which will help a person to develop a responsiveness of the self-disciplinary nature of problems in information and communication technology.

Machine learning is the science of getting computers to act without being explicitly programmed. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that we probably use it dozens of times a day without knowing it.

Machine learning is important because it gives enterprises a view of trends in customer behaviour and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.

Hence, this training has given us a great opportunity to learn various topics about the Machine Learning that has been widely used in industry currently. I have learned various topics like Introduction to Python, Data exploration and analysis, Linear Regression, Dimensionality Reduction, Logistic Regression, Decision Trees, Ensemble models and Clustering. I have also worked on two projects - Customer Churn Prediction and House Price Prediction.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
|  | **CONTENT** | **PAGE NO.** |
| **1.** | **PROFILE OF THE ORGANIZATION** | **1** |
|  | 1.1 About the company | 1 |
|  | * 1. Products and services | 1 |
|  | * 1. Company Strategy | 1 |
|  | 1.4 Domain | 2 |
|  | 1.5 Clients for the company | 2 |
| **2.** | **MODULES OF TRAINING** | **3** |
|  | 2.1 Introduction to Machine Learning | 3 |
|  | 2.2 Data | 4 |
|  | 2.3 Introduction to python | 6 |
|  | 2.4 Data exploration and pre-processing | 7 |
|  | 2.5 Linear Regression | 10 |
|  | 2.6 Introduction to Dimensionality Reduction | 12 |
|  | 2.7 Logistic Regression | 13 |
|  | 2.8 Decision Trees | 14 |
|  | 2.9 Ensemble Models | 16 |
|  | 2.10 Clustering | 17 |
| **3.** | **HANDS-ON PROJECT DURING TRAINING** | **19** |
|  | 3.1 House Price Prediction | 19 |
|  | 3.2 Customer Churn Prediction | 26 |
| **4.** | **REFLECTIONS** | **35** |
|  | 4.1 Industrial training outcomes | 35 |
|  | 4.2 Non-Technical Outcomes | 35 |
|  | 4.3 Time Management | 36 |

**CHAPTER 1**

**PROFILE OF THE ORGANIZATION**

* 1. **ABOUT THE COMPANY**

Internshala is an internship and online training platform. This website helps students to find best internships for their choice with organizations in India. It is India’s no.1 training and internship platform with 40000+paid internships in Engineering, MBA, media, law, arts, and other streams. Basically, Internshala means Internship resources. this site offers internship searching and posting, and other career services such as counselling, cover-letter writing, resume writing and training programs to students. It’s headquarter based in Gurugram, India. It is founded by Sarvesh Agrawal, an IIT madras alumnus in 2010.

* 1. **PRODUCT AND SERVICES**

Internshala offers following products & services to its users:

1. Internship and job posting: Organization & individuals looking to hire interns and/or full-time employees can post their internship and/or job requirement on Internshala.
2. Internship and job search: Internship and job seekers from all over India can search and apply for internships and jobs relevant to their coursework and interest free of cost.
3. Internshala Trainings: Internshala offers a variety of online training programs across multiple disciplines (Machine Learning, Web Development, Digital Marketing, Java, French and several others) which students can do from the comfort of their homes and learn the skills needed in today’s industry.

Internshala reserves the right to add/remove any product and service at its sole discretion or change the price charged for such offerings.

**1.3 COMPANY STRATEGY**

* **Purpose:** The purpose of the Internshala is to facilitate the students with trainings on their interested topics and to provide the chance of finding internship for the interested ones.
* **Vision:** Internshala is a dot com business with the heart of dot org.
* **Mission:** Internshala is a technology company on a mission to equip students with relevant skills & practical exposure through internships and online trainings.
* **Core values:** Internshala helps a student to turn his passion into profession by making him/her graduate fully assured, confident and stake a place in the world.
* **Goals:** Flexibility, Convenience, Affordability, Chance for a fresher to seek a job and finally quality.

**1.4 DOMAINS**

Internshala works on domains like:

* Trainings - One can learn new-age skills on the go with their short-term online trainings.
* Job-Oriented Specialisations - Specialize in industry-oriented programs that make one ready for a career in their dream profile.
* Fresher Jobs - Get premium fresher jobs with a minimum CTC of 3LPA on one’s fingertips.
* Internships - One can find internships from great companies to give a kickstart to their career.

**1.5 CLIENTS FOR THE COMPANY**

Internshala is used by 25,000 businesses to recruit interns. Some of the notable names include Reliance, HDFC Bank, Ibibo group, Make My Trip, Kolkata Knight Riders and many Members of Parliament (MPs) etc. On student side, Internshala is used by 750,000 students across 5,000 colleges in India. So far growth of Internshala has come from Word of Mouth from happy customers (~50% of new business is referrals) and we intend to stay laser focused on that. In addition, we have also started advertising on relevant business & HR forums to get more corporate client onboard.

**CHAPTER 2**

**MODULES OF TRAINING**

I have completed 6 weeks industrial training on “Machine Learning” and the curriculum of the training was organised in the following manner.

|  |  |
| --- | --- |
| Week 1 | Introduction to Machine Learning |
| Data |
| Introduction to python |
| Week 2 | Data exploration and pre-processing |
| Week 3 | Linear Regression |
| Introduction to Dimensionality Reduction |
| Week 4 | Logistic Regression |
| Decision Trees |
| Week 5 | Ensemble Models |
| Clustering |
| Week 6 (Projects and Final Exam) | House Price Prediction |
| Customer Churn Prediction |
| Final Exam |

**2.1 Introduction to Machine Learning:**

Machine learning (ML) is the study of computer [algorithms](https://en.wikipedia.org/wiki/Algorithm) that can improve automatically through experience and by the use of data. It is seen as a part of [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence). Machine learning algorithms build a model based on sample data, known as [training data](https://en.wikipedia.org/wiki/Training_data), in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, [email filtering](https://en.wikipedia.org/wiki/Email_filtering), [speech recognition](https://en.wikipedia.org/wiki/Speech_recognition), and [computer vision](https://en.wikipedia.org/wiki/Computer_vision), where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Some popular applications of Machine Learning are:

* Facebook newsfeed
* Facebook photo auto tag
* Product Recommendation

**How Machine Learning Works?**

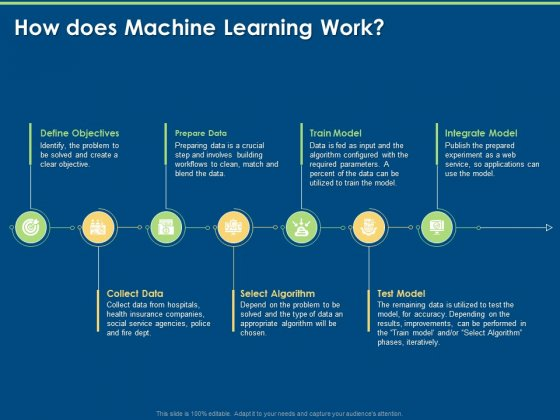
****

Fig 1

**Types of Machine Learning:**

1. Supervised Learning: Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labelled training data consisting of a set of training examples.
2. Unsupervised Learning: Unsupervised learning is a type of algorithm that learns patterns from untagged data. The hope is that through mimicry, which is an important mode of learning in people, the machine is forced to build a compact internal representation of its world and then generate imaginative content from it.

**2.2 Data:**

Data are individual facts, statistics, or items of information, often numeric. In a more technical sense, data are a set of values of qualitative or quantitative variables about one or more persons or objects, while a datum is a single value of a single variable



Fig 2

**Variables:** A variable represents one specific characteristic of data or tells one specific information about the data under consideration. The value of the variable can change and usually changes across different data points over time.

E.g., Call duration, Contact name etc.

Types of variables:

* Numeric or Quantitative variables. E.g., Call duration
  + Continuous variables. E.g., Temperature
  + Discrete variables. E.g., Number of calls made in a day
* Categorical or Qualitative variables. E.g., Contact name
  + Ordinal variables. E.g., Rating of product
  + Nominal variables. E.g., Colour of the Phone
* Independent variables: E.g., Number of calls made in a day
* Dependent variables: E.g., Battery percentage (Depends on the number of calls made in a day)

**Types of Data:**

* Structured Data: Examples of structured data include names, dates, addresses, credit card numbers, stock information, geolocation, and more. Structured data is highly organized and easily understood by machine language.
* Unstructured Data: Unstructured data is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well.

**Graphical and Analytical Representation of data:**

Data analysis is a process of inspecting, [cleansing](https://en.wikipedia.org/wiki/Data_cleansing), [transforming](https://en.wikipedia.org/wiki/Data_transformation), and [modelling](https://en.wikipedia.org/wiki/Data_modelling) [data](https://en.wikipedia.org/wiki/Data) with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.

Types of analysis:

* Descriptive analysis
* Diagnostic analysis
* Predictive analysis
* Prescriptive analysis

Types of graphical data representation:

* Line Graph
* Bar Graph
* Histogram/Frequency Polygon
* Pie-Chart
* Scatter Plot
* Box and Whisker plot
* Time Series

**Limitation of Traditional Data Analysis:**

* Lack of alignment within teams
* Lack of commitment and patience
* Low quality of data
* Privacy Concerns
* Complexity and bias

**2.3 Introduction to Python:**

Python is an interpreted, high-level, general-purpose programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python’s elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

Concepts of python that are covered are:

1. **Understanding Operators: -** Operators are symbolic representation of Mathematical tasks.
2. **Variables And Datatypes: -** 
   * Variables are names bounded to objects.
   * Data types in python are Numbers, Strings, Lists, Tuples, Sets, Dictionaries
3. **Conditional Statements: -**
   * If-else statements (Single condition)
   * If- elif- else statements (Multiple Condition)
4. **Looping Constructs: -**
   * For loop
   * While loop
5. **Functions: -** 
   * Functions are re-usable pieces of code.
   * Created for solving specific problem.
   * Two types: -
     + Built-in functions.
     + User- defined functions.
   * Functions cannot be reused in python.
6. **Data Structures: -**
   * Two types of Data structures:
     + Lists: It is an ordered data structure with elements separated by comma and enclosed within square brackets [].
     + Dictionary: It is an unordered data structure with elements separated by comma and stored as key-value pair, enclosed with curly braces {}.
7. **Understanding Libraries: -**
   * **Pandas: -** It is the most popular library of python used in Data Science. Pandas has a wide usage in data manipulation and analysis. It also has features like dataset joining and merging, data filtration, reshaping of datasets and many more.
   * **NumPy: -** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
   * **Matplotlib: -** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.
   * **Seaborn: -** Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**2.4 Data Exploration and Pre-Processing:**

Data exploration is an approach similar to initial data analysis, whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data, rather than through traditional data management systems.

Why is data exploration important?

* Understand the data and make sure it is ready to be used in a model.
* A model is as good as the data on which it is built.
* It is structured and step-by-step approach in understanding and preparing the data

**Outliers:** An outlier is a data point that is noticeably different from the rest. They represent errors in measurement, bad data collection, or simply show variables not considered when collecting the data.

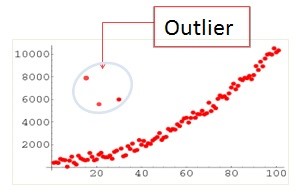


Fig 3

We can use box-whisker plots to identify outliers.

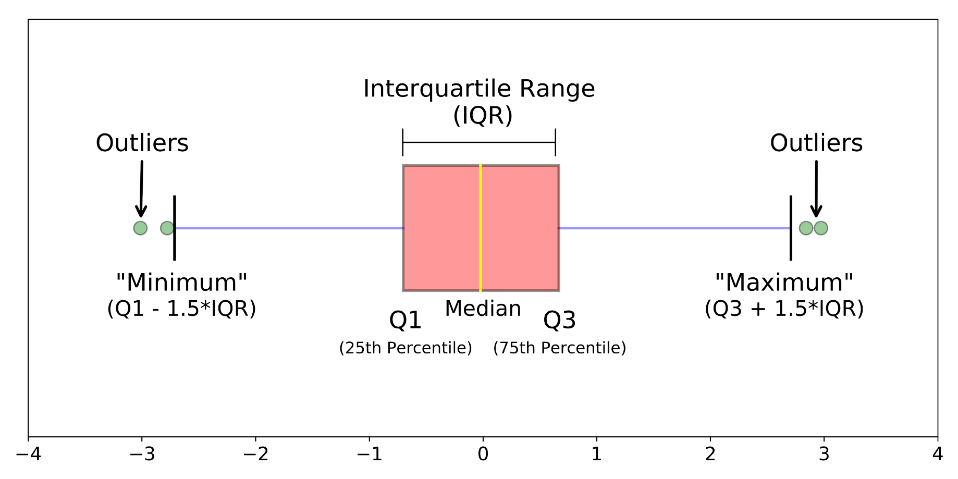


Fig 4

Why outliers occur?

* Genuine variability in data.
* Recording errors

Treating Outliers:

* Deletion: The entire row containing the outlier is removed.
* Capping/Imputing: The outlier is replaced with median/average values.
* Data Transformation: The variable is transformed to its log, square or cube roots hoping that the transformed values will be in range.
* Binning: Different bins are formed based on the values of the variable.

**Variable Transformation:** It refers to the process of making changes to a variable in a way that it becomes more useful and meaningful for analysis and modelling purposes.

Need for variable transformation:

* Outlier treatment
* When a variable does not represent the information, it contains in the best possible way.
* To derive new variable from existing ones or combining variables to transform a variable to make it more useful

**Correlation:** Correlation is a statistical term describing the degree to which two variables move in coordination with one another. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation.

**ANOVA:** Itstands for Analysis of Variance. ANOVA is a statistical test used to analyse the difference between the means of more than two groups. There are two types of testing in ANOVA. They are: -

1. One-way ANOVA: - A one way ANOVA is used to compare two means from two independent groups using the F-distribution. F-Distribution is a probability distribution of the F-Statistic. An F-statistic is a value you get when you run an ANOVA test or a regression analysis to find out if the means between two populations are significantly different.
2. Two-way ANOVA: - A Two Way ANOVA is an extension of the One-way ANOVA. With a One-way, one can have one independent variable affecting a dependent variable. With a Two-way ANOVA, there are two independents.

**Train-Test Split of the data:** Training data is data on which the model is built. Training data is used by the model to learn and usually a larger part of the data is taken as the training data. Once the model is trained, its performance is examined using another subset of original data called the test data. The Model predicts the target values for the test data and these predictions are compared with original data for accuracy prediction and further analysis.

**Feature Scaling of data:** Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Techniques for Feature Scaling:

* Standardization: It involves rescaling the feature values so that they have the properties of a standard normal distribution.
* Min-Max Scaling: The variable is transformed such that its value lies between 0 and 1.
* Normalization: The variable is transformed such that its value lies between -1 and 1. It is also called mean normalization.

**2.5 Linear Regression**

* Linear Regression is a statistical method to estimate a linear relationship between a target variable Y and a set of non-target variables X.
* The representation is a linear equation that combines a specific set of input values X the solution to which is the predicted output for that set of input values X. As such, both the input values and the output value are numeric.
* As it plots a linear graph, its name is Linear Regression.
* Two types of Linear Regression are explained here: -
  + Simple Linear Regression: - Here only one X variable will be there. Formula for simple linear regression is given as: -

Where *b0* is the constant and *b1*is the slope or co-efficient.

* + Multiple Linear Regression: - Here there are many X variables. Formula is given as:

Where *bi* is the respective slopes/co-efficient for corresponding *xi* variable.

The idea of Linear Regression is to fit a best line whose sum of square of distance from all the points is minimum.

Assumptions for Linear Regression:

* Linearity: The relationship between X and the mean of Y is linear.
* Homoscedasticity: The variance of residual is the same for any value of X.
* Independence: Observations are independent of each other.
* Normality: For any fixed value of X, Y is normally distributed.

**Gradient descent:** Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. The idea is to take repeated steps in the opposite direction of the gradient of the function at the current point, because this is the direction of steepest descent.

**Feature Engineering:** It is the technique of generating new features from the existing features from the existing features. It is the science of extracting more information from the existing data. It consists of

* Feature pre-processing
  + Feature transformation
    - Dummy Encoding
    - Label Encoding
* Feature generation
  + Feature Extraction
  + Combination of Features

**Model Evaluation Matrices:**

1. Mean Absolute Error:
2. Mean Squared Error:
3. Root Mean Square:
4. R2: It is the proportion of the variation in the dependent variable that is predictable from the independent variable.
5. Adjusted R2: It penalizes the result for adding variables which do not improve the existing model.

**2.6 Introduction to dimensionality Reduction**

The number of input variables or features for a dataset is referred to as its dimensionality. Dimensionality reduction refers to techniques that reduce the number of input variables in a dataset ensuring that the reduced data conveys maximum information.

Challenges of highly dimensional data:

* Visualization becomes tougher
* Increase in computational complexity
* All the variables might not be important factor for predicting the target variable
* Difficult model interpretability
* Difficult data exploration

Techniques for Dimensionality Reduction:

1. **Missing value ratio:** Some features are removed if there they contain more missing values than a certain threshold. An example of threshold can be 70% missing data in a feature.
2. **Low variance:** In predictive modelling, the independent variables with low variance are eliminated.
3. **High correlation:** If two independent variables in the data are highly correlated, one of them is removed as they produce same effect in predicting the target value. It Is not a reliable technique as it only looks at one-one correlation. A reliable technique which considers things at an aggregate level is called Variation Inflation Factor (VIF). The VIF quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.
4. **Backward feature elimination:** It is used to remove redundantvariables from the model. It strikes a balance between the model performance and model simplicity.
5. **Forward feature selection:** Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model based on adjusted-R2 score till addition of a new variable does not improve the performance of the model.

**2.7 Logistic Regression**

Logistic regression is a classification model rather than a regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. In order to restrict the values of prediction between 0 and 1 and to make predictions as probabilities, we use logit function also known as sigmoid function. A simple logit function can be:

**Evaluation Metrices:**

1. **Confusion Matrix:** A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

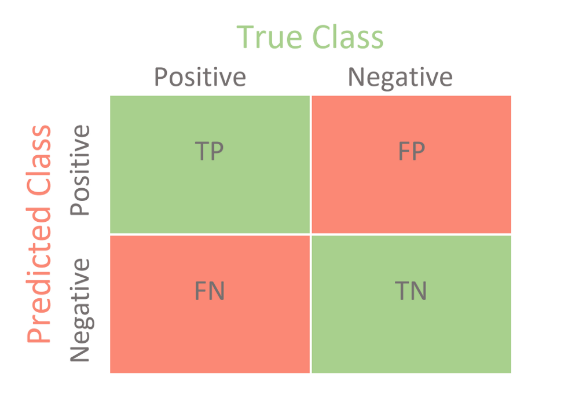


Fig 5

1. **Accuracy:** In a set of measurements, accuracy is closeness of the measurements to a specific value.
2. **Precision:** It is used when avoiding false positives is more essential than encountering false negatives.
3. **Recall:** It is used when avoiding false negatives is more essential than encountering false positives.
4. **F1-Score:** It is the harmonic mean of precision and recall.
5. **Log loss:** Log-loss is indicative of how close the prediction probability is to the corresponding actual/true value (0 or 1 in case of binary classification). The more the predicted probability diverges from the actual value, the higher is the log-loss value.
6. **AUC-ROC curve:** AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve. AUC provides an aggregate measure of performance across all possible classification thresholds.



Fig 6

**2.8 Decision Trees**

**Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.**

**Working of decision trees:**

Decision Trees can be used for both classification and regression. The methodologies are a bit different, though principles are the same. The decision trees use the CART algorithm (Classification and Regression Trees). In both cases, decisions are based on conditions on any of the features. The internal nodes represent the conditions and the leaf nodes represent the decision based on the conditions.

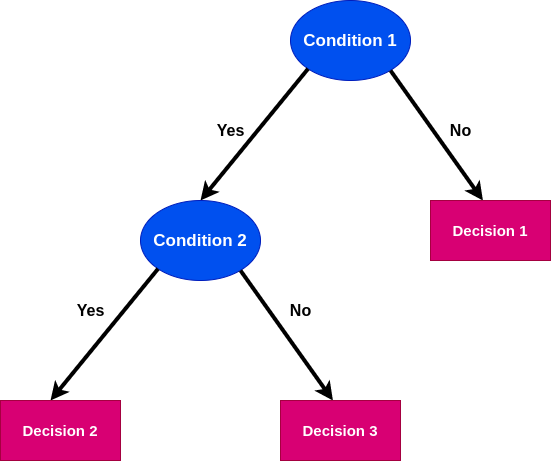


Fig 7

Criteria in a decision tree split (Classification Decision Tree):

1. **Gini impurity:** The Gini impurity measure is one of the methods used in decision tree algorithms to decide the optimal split from a root node, and subsequent splits. A Gini Impurity measure will help us make this decision. Gini Impurity tells us what is the probability of misclassifying an observation.
2. **Information gain using entropy:** Information gain is the reduction in entropy or surprise by transforming a dataset and is often used in training decision trees. Information gain is calculated by comparing the entropy of the dataset before and after a transformation. Entropy in machine learning is related to randomness in the information being processed in the project.

Criteria in a decision tree split (Regression Decision Tree):

1. **Variance:** Reduction in Variance is a method for splitting the node used when the target variable is continuous, i.e., regression problems. It is so-called because it uses variance as a measure for deciding the feature on which node is split into child nodes. Variance is used for calculating the homogeneity of a node.

Some advantages of decision trees are:

* Simple to understand and to interpret. Trees can be visualised.
* Requires little data preparation. Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
* The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
* Able to handle both numerical and categorical data. However, scikit-learn implementation does not support categorical variables for now. Other techniques are usually specialised in analysing datasets that have only one type of variable.
* Able to handle multi-output problems.
* Uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by Boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret.
* Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model.
* Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

**2.9 Ensemble Models**

Ensemble modelling is a process where multiple diverse models are created to predict an outcome, either by using many different modelling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data.

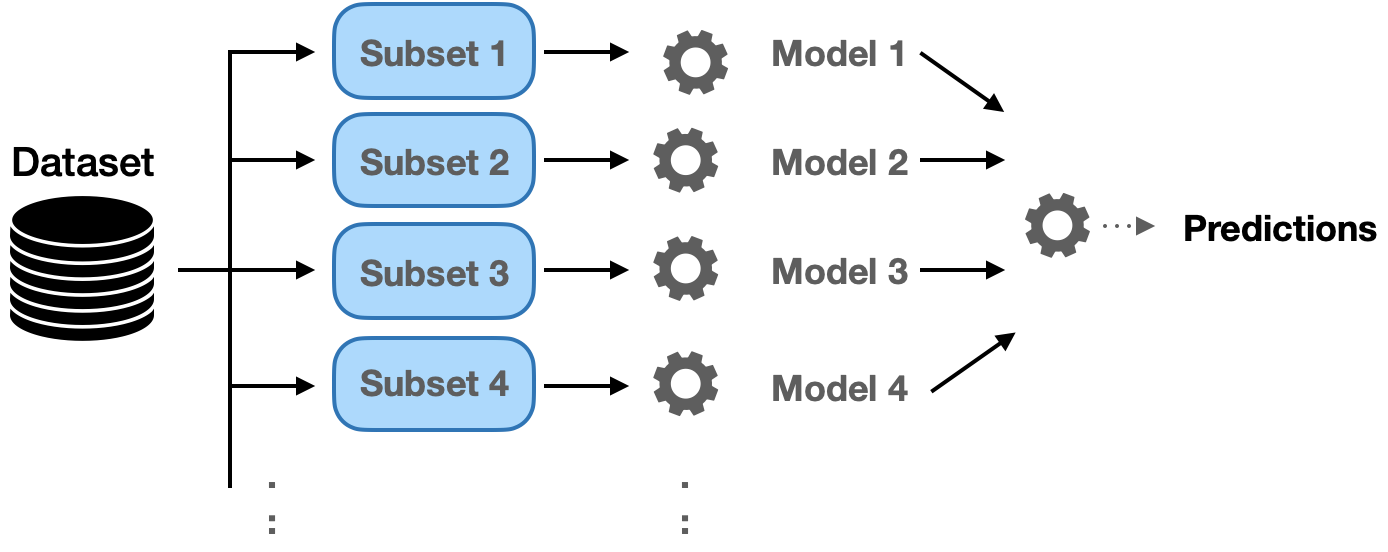


Fig 8

Advantages of Ensemble models:

* Captures most of the diverse signals or patterns
* Less incorrect predictions
* Reduce overfitting
* Better performance

Disadvantages of Ensemble Models:

* Increased complexity
* Not interpretable as a whole
* Increased time complexity and computational requirements

**Random Forest:** Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

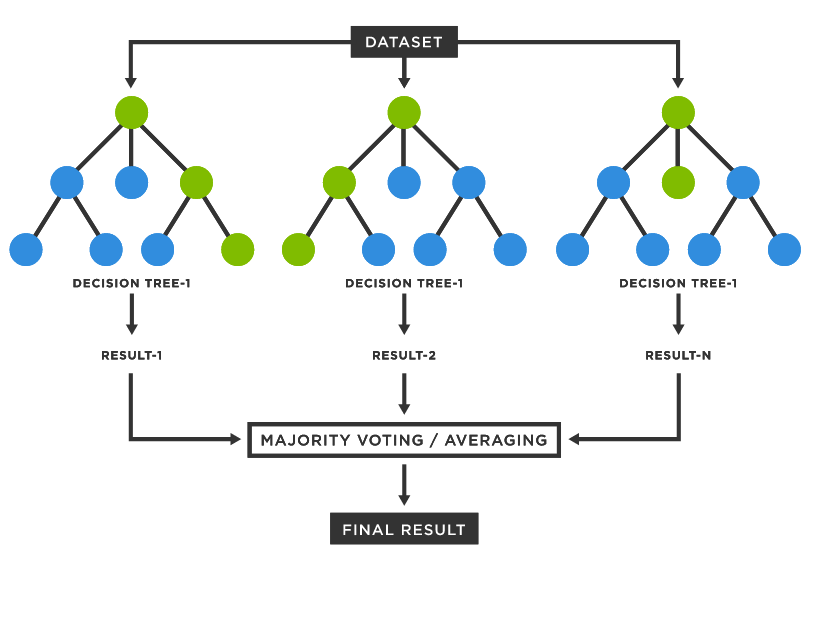


Fig 9

**2.10 Unsupervised Machine Learning - Clustering**

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as a way of grouping the data points into different clusters, consisting of similar data points.

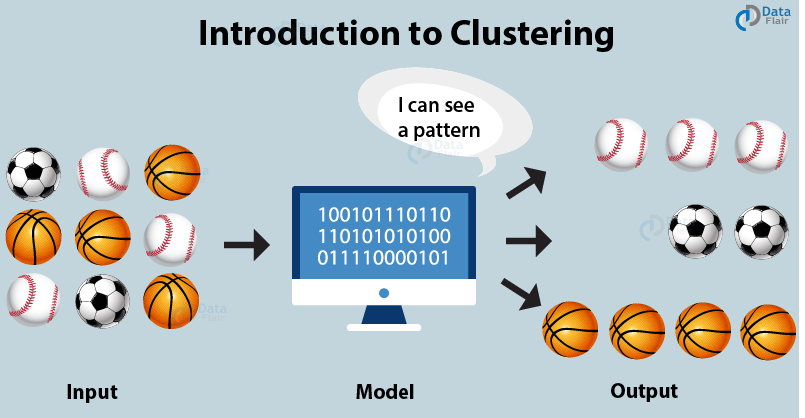


Fig 10

Applications of clustering:

* Customer segmentation
* Document clustering
* Image segmentation
* Recommendation engines

Evaluation of clustering models:

1. **Inertia**: Inertia is the sum of squares for all dataset points to their closest centroid. You can measure inertia after each convergence. If we repeat steps 2-5 multiple times and measure the inertia each time, the optimal clusters are the final assignments with the smallest inertia.
2. **Dunn Index:** The Dunn Index is a method of evaluating clustering. A higher value is better. It is calculated as the lowest inter-cluster distance divided by the highest intra-cluster distance.

**K-Means clustering algorithm:** K-means clustering is a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. Data points are clustered based on feature similarity.

**CHAPTER 3**

**HANDS-ON PROJECTS DURING TRAINING**

**3.1 House Price Prediction using Linear Regression**

House prices increase every year, so there is a need for a system to predict house prices in the future. Prediction house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location.

**Dataset:**

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. There are a total of 21 columns and 21613 rows in the dataset. Each row represents the details of one house that was previously sold. The following figure shows the first 5 entries in the dataset considered.

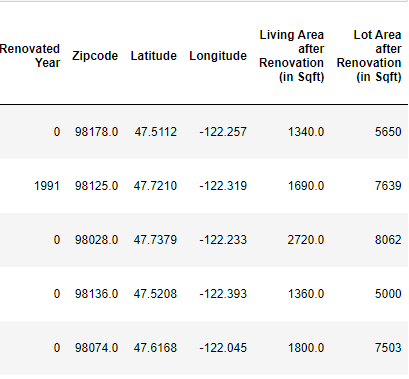
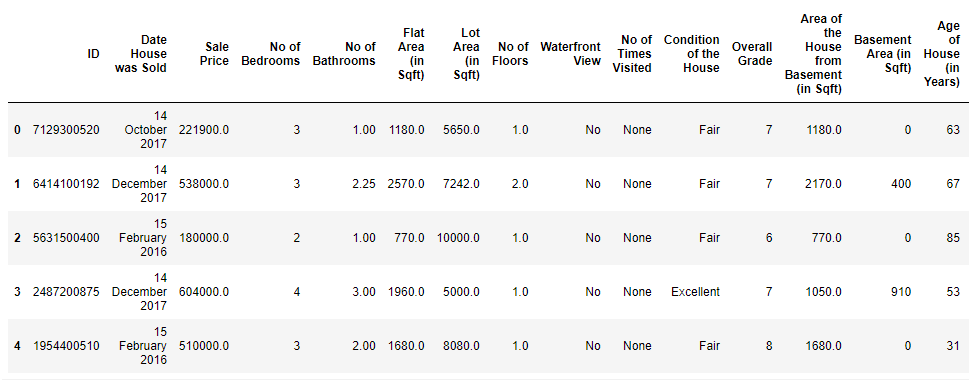


Fig 11

**Implementation:**

Importing the required libraries for the project

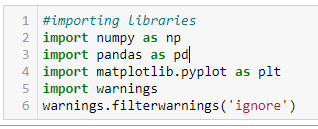


Fig 12

Quantitative description of the Target variable:

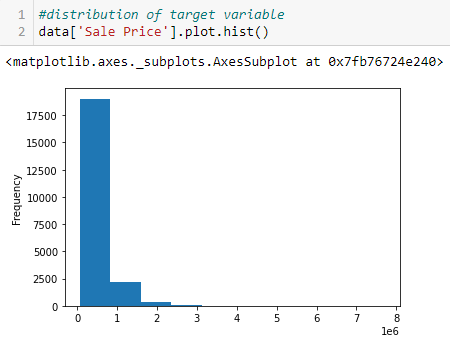
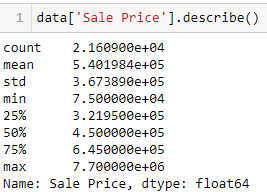


Fig 13

**Checking and removing the outliers:** We first calculate the 25% and 75% quartiles and we then find the outliers in the target variable. Limit\_imputer function caps the outliers to the upper and lower limits.

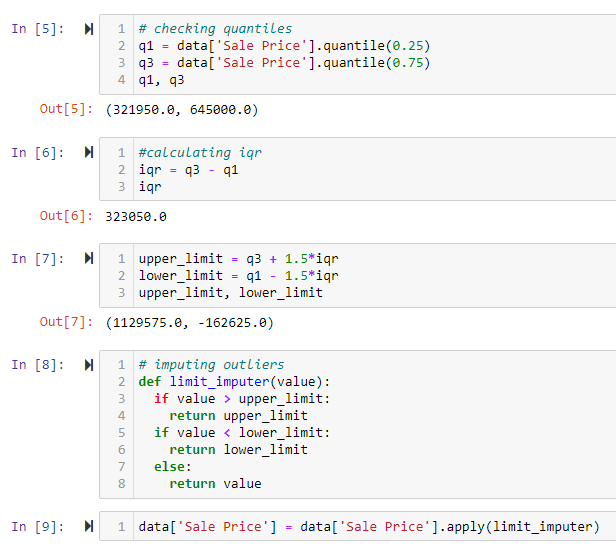


Fig 14

Quantitative description of the Target variable after treating the outliers:

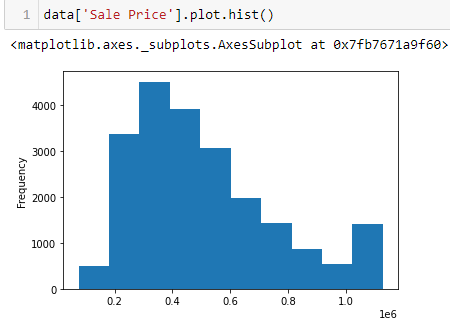
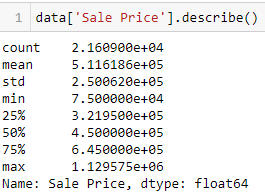


Fig 15

Dropping the rows with missing values:

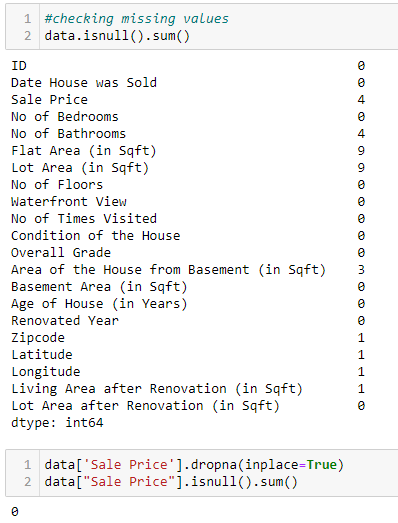


Fig 16

**Imputing missing values in numerical variables:** Here we use inbuilt imputer from sklearn. We have replaced the missing values with the median value of the respective attributes.

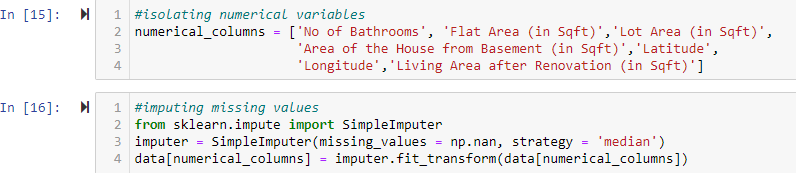


Fig 17

**Variable Transformations:**

1. **Zip code:**

Missing values are replaced with most frequent occurrence in the dataset for Zip code attribute.

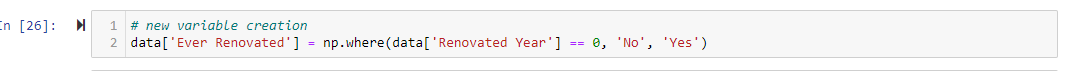


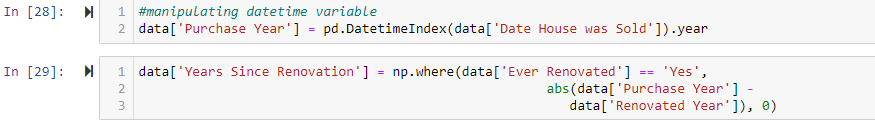
1. **Data extraction from date-based attributes:**

The following variables are created:

* Ever Renovated
* Purchase Year
* Years since Renovation

The redundant variables are then dropped.





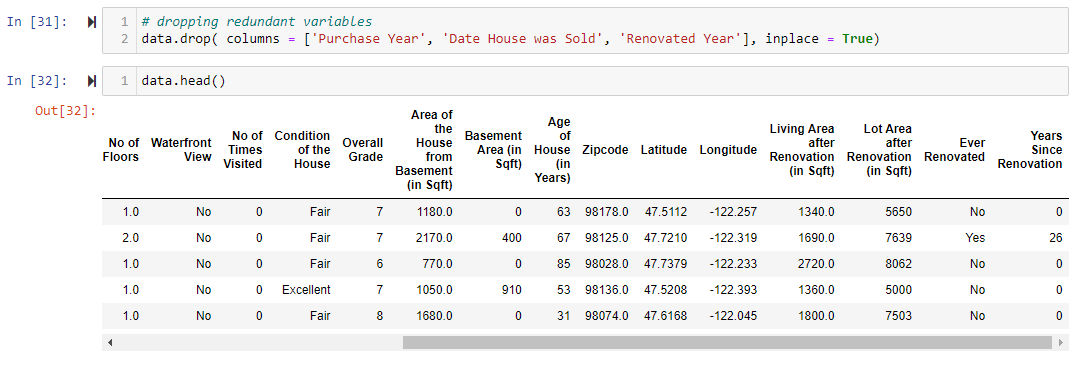


Fig 18

1. **Conversion from string to categorical variables.**
2. **No of Times Visited:**

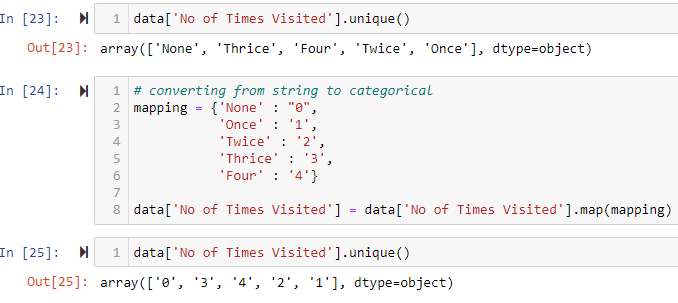


Fig 19

1. **Waterfront View, Condition of the House, Ever Renovated:**



**Building the Model – Linear Regression:**

1. **Removing skew in numerical variables:**

Distribution of variables before removal of skew

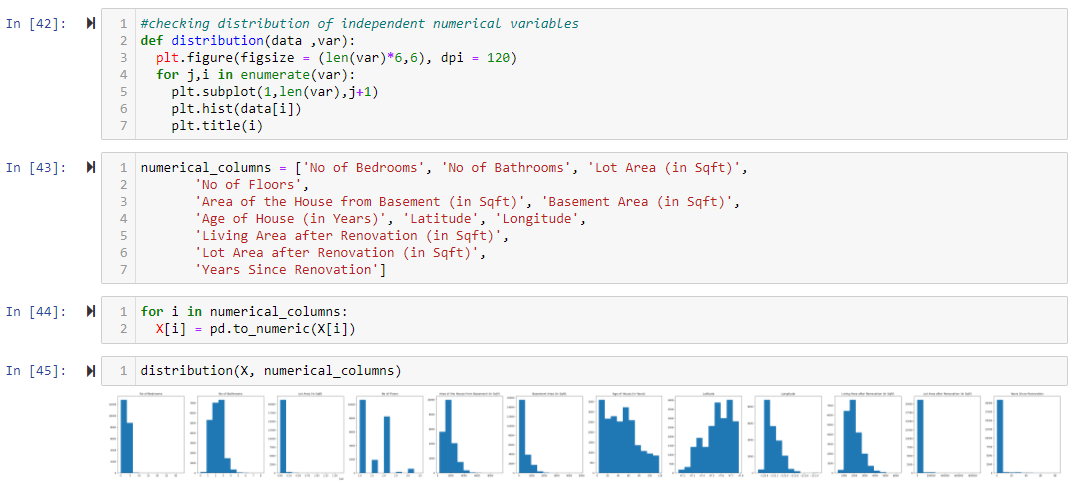


Fig 20

Distribution of variables after removal of skew using the right\_skew function:

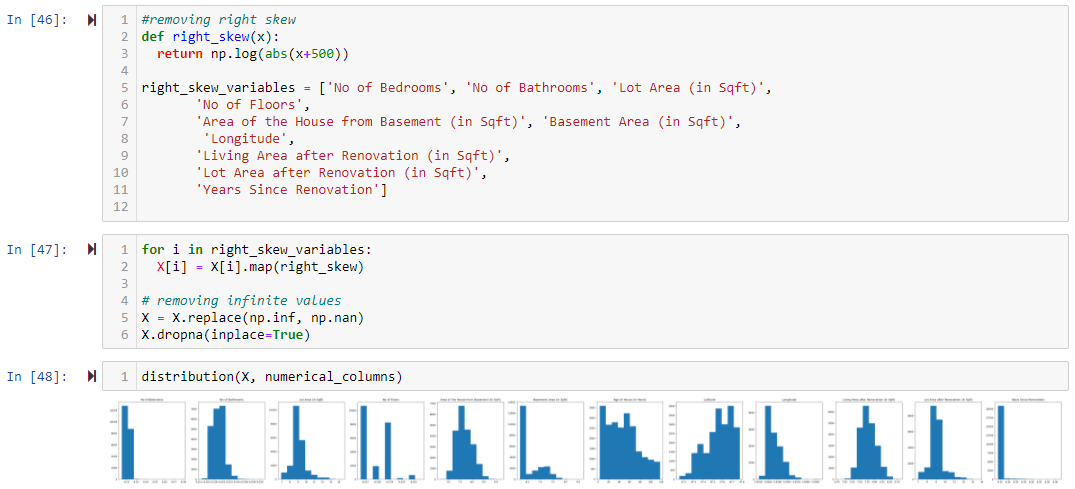


Fig 21

1. **Scaling the dataset:**

We use the StandardScalar function sklearn for scaling the dataset. It Standardizes features by removing the mean and scaling to unit variance.

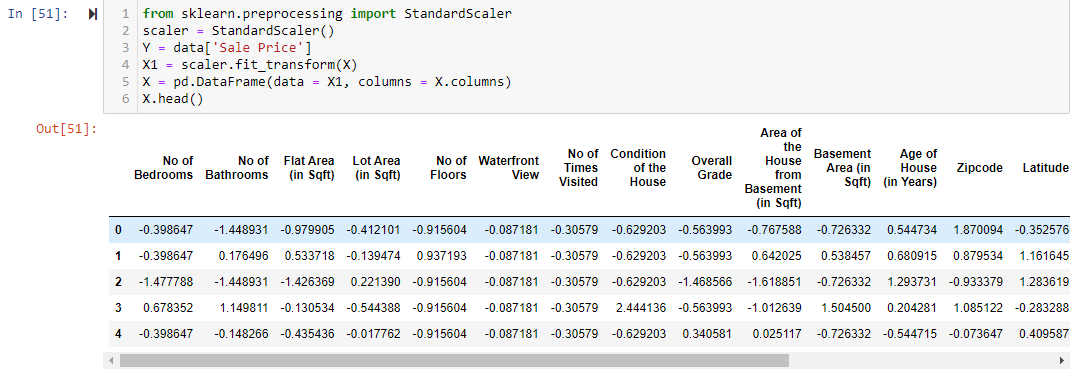


Fig 22

1. **Checking and Removing Multicollinearity:**

Here we remove the highly correlated columns of the dataset by considering VIF:

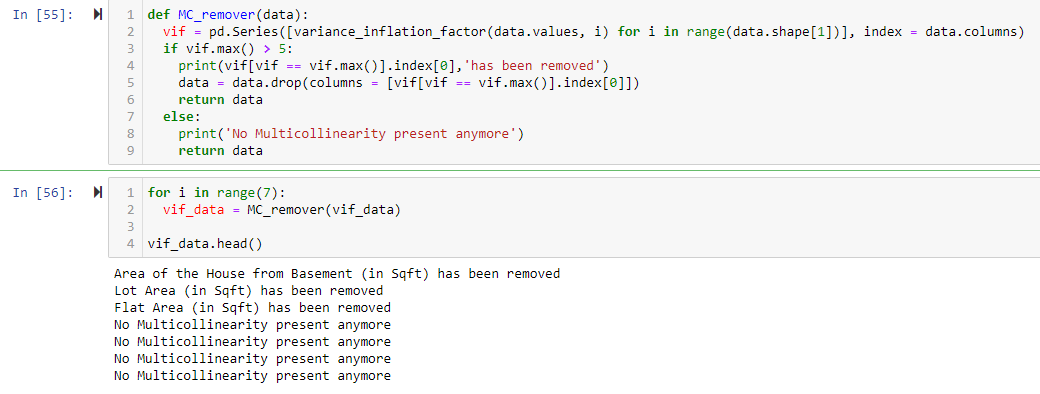
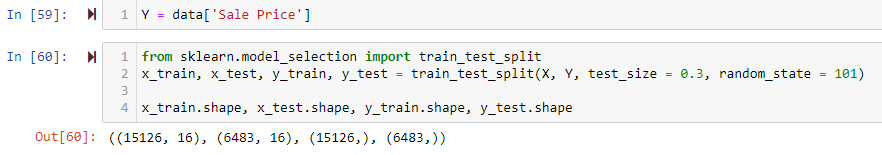


Fig 23

1. **Splitting the dataset:**

Here we consider a 70-30 split.



1. **Training the Model:**

Here, we use LinearRegression model from sklearn and train the model on the training data. We then calculate the accuracy of the built model using the test data.

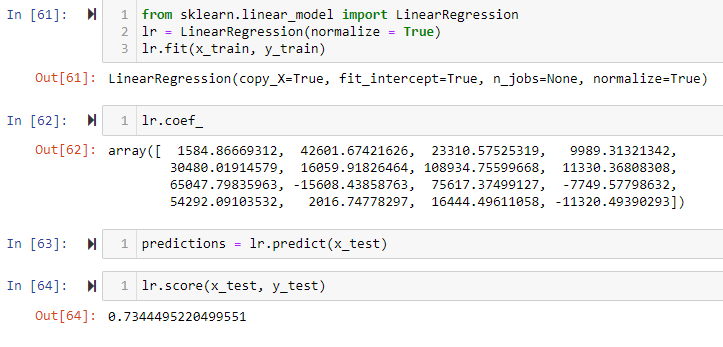


Fig 24

**Model Evaluations:**

1. **Residuals:**

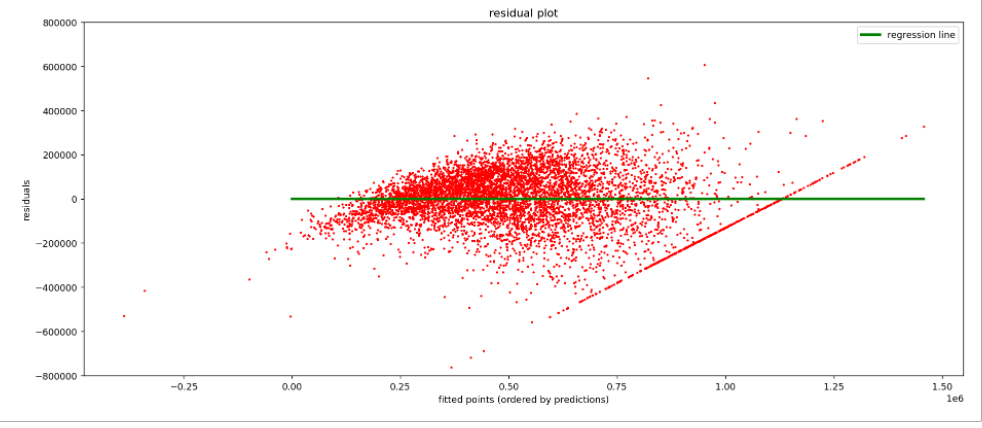


Fig 25

1. **Distribution of errors:**

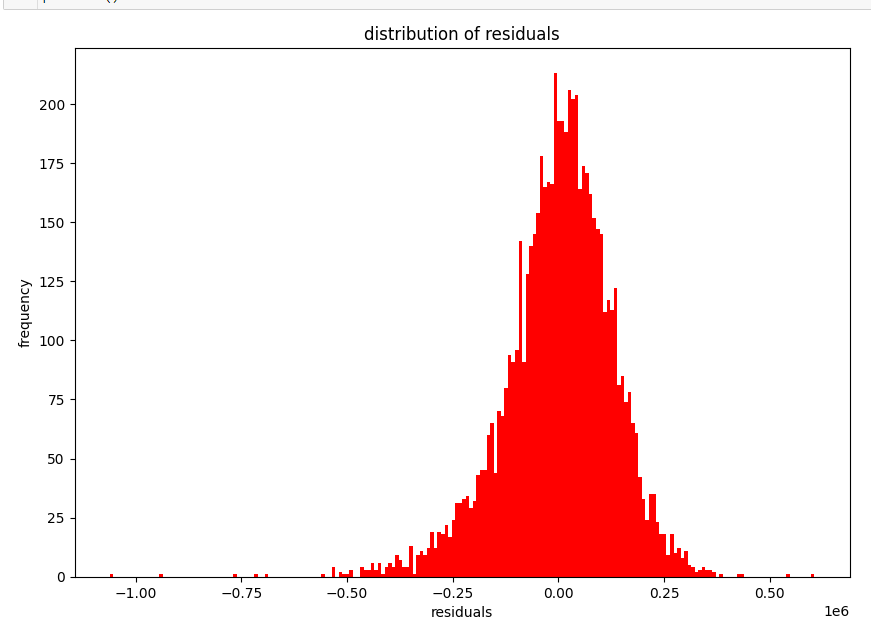


Fig 26

1. **Model Coefficients:**

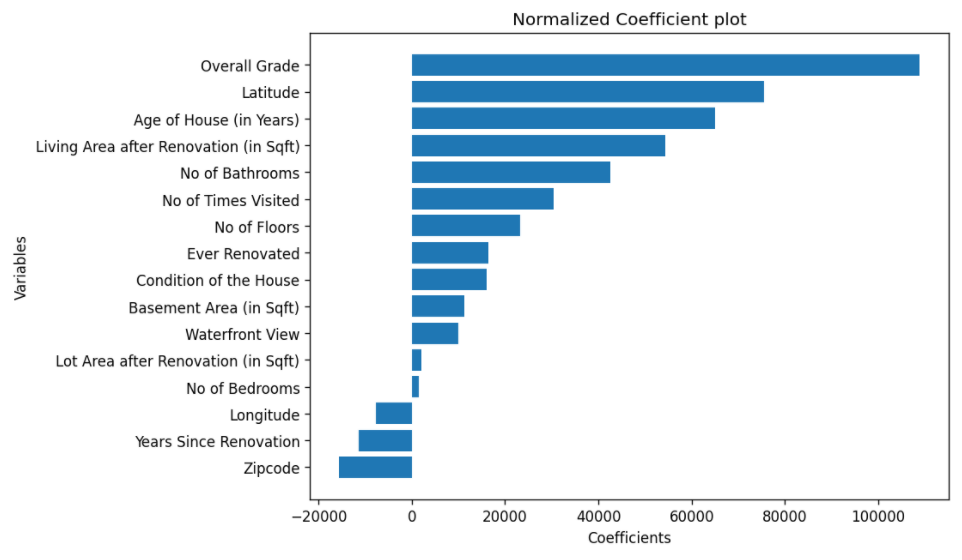


Fig 27

**3.2 Customer Churn Prediction using Logistic Regression**

Churn is defined in business terms as ‘when a client cancels a subscription to a service they have been using.’ So, Churn Prediction is essentially predicting which clients are most likely to cancel a subscription i.e., ‘leave a company’ based on their usage of the service. From a company point of view, it is necessary to gain this information because acquiring new customers is costlier than retaining old ones.

The output in the case of Churn prediction is a simple yes or a no. That makes it a classification problem where you have to predict 1 if the customer is likely to churn and 0 otherwise. Many factors influence the reasons for a customer to Churn like a new competitor in the market offering better prices, what they are getting is not up to the mark and so on.

**Dataset:**

Each row represents a customer, each column contains customer’s attributes described on the column Metadata. The data set includes information about:

* Customers who left within the last month – the column is called Churn
* Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
* Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges
* Demographic info about customers – gender, age range, and if they have partners and dependents.

**Implementation:**

We import the necessary libraries. The data contains 28382 rows and 21 columns.

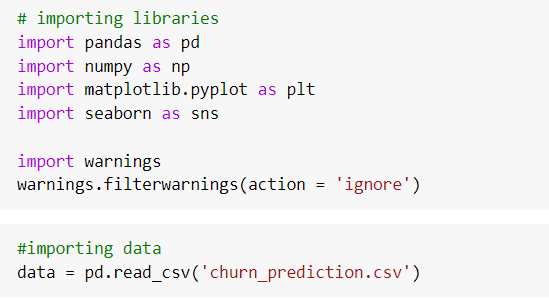
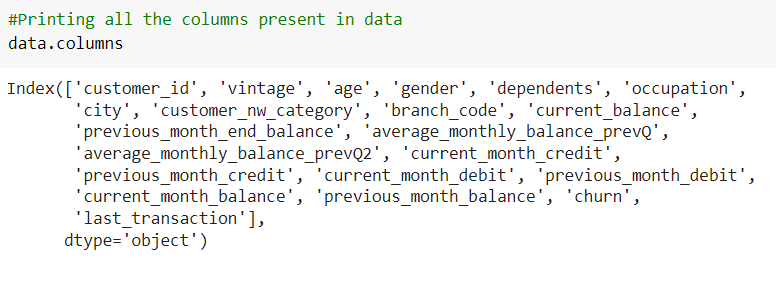


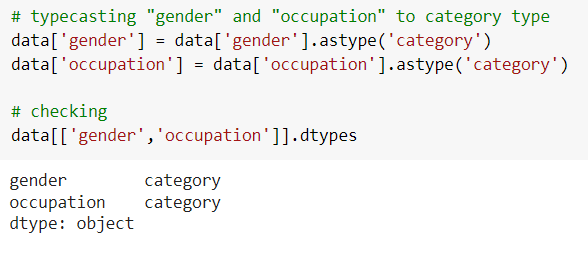
Fig 28

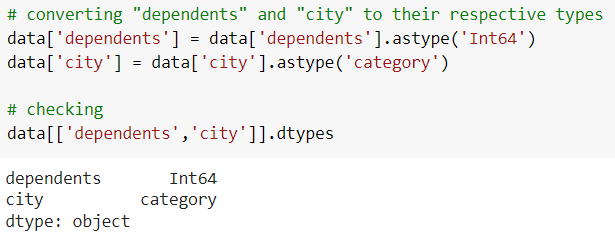
The columns of the dataset are

**Variable Identification and typecasting:**

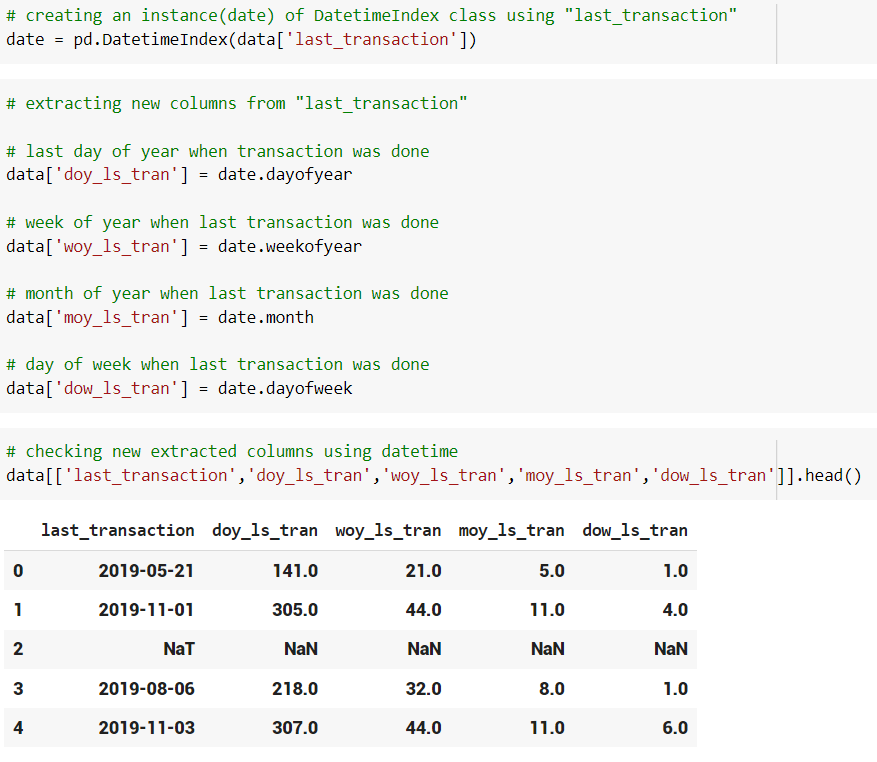
Certain variable’s datatypes are changed based on the requirement







**Date time datatype:**

The first column is the complete date of the last transaction which was done by the any given customer.

The next columns represent the day of year, week of year, month of year, day of week when the last transaction was done.

Breaking down the date variable into this granular information will help us in understand when the last transaction was done from different perspectives. Now that we have extracted the essentials from the last\_transaction variables, we will drop it from the dataset.

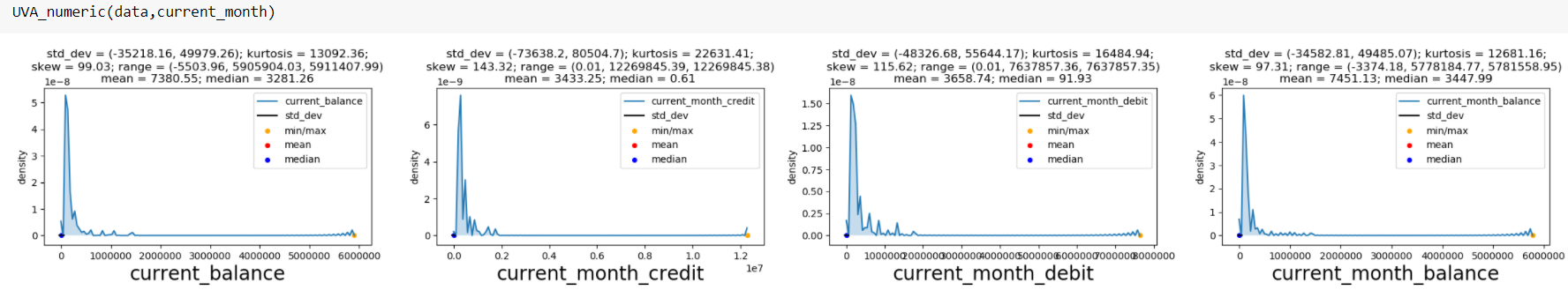
**Current month:**

Fig 29

**To remove outliers:**

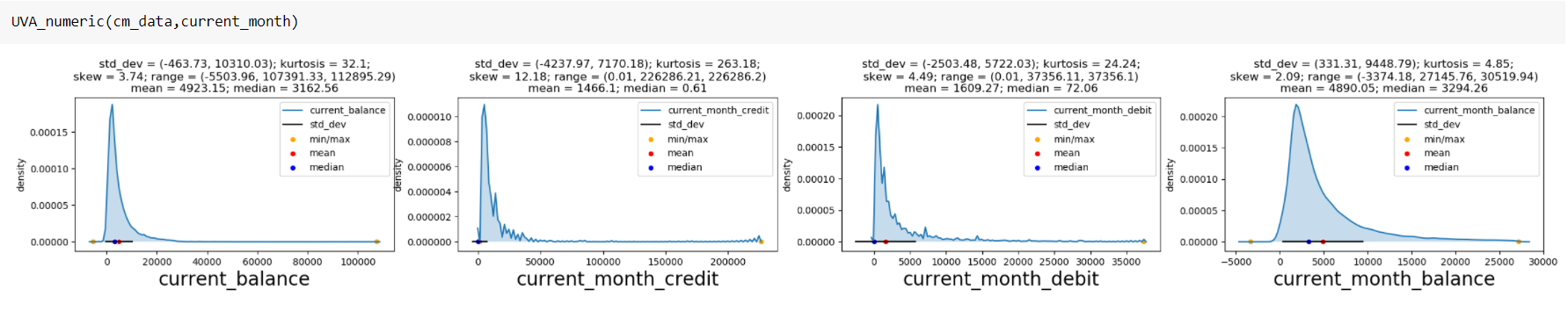
**After removing outliers:**

Fig 30

Plots are skewed after the removal of outliers.

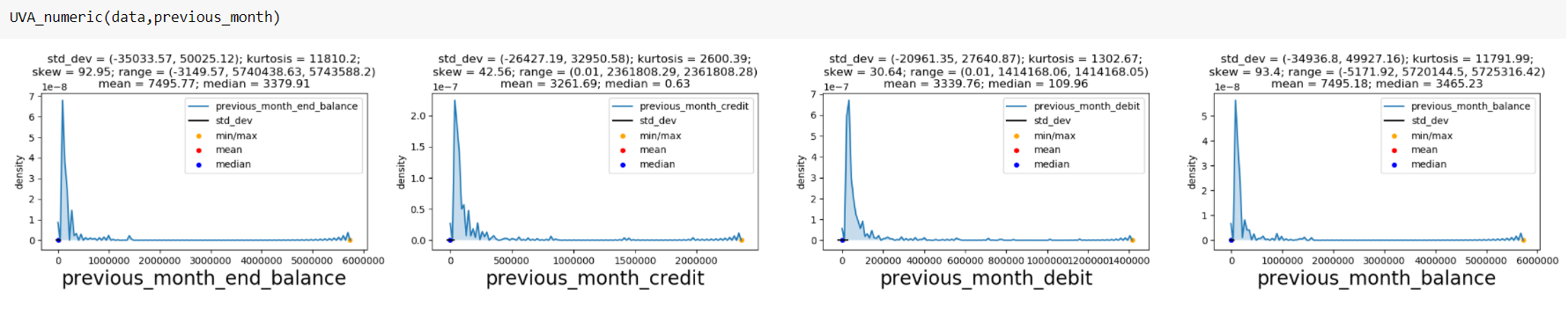
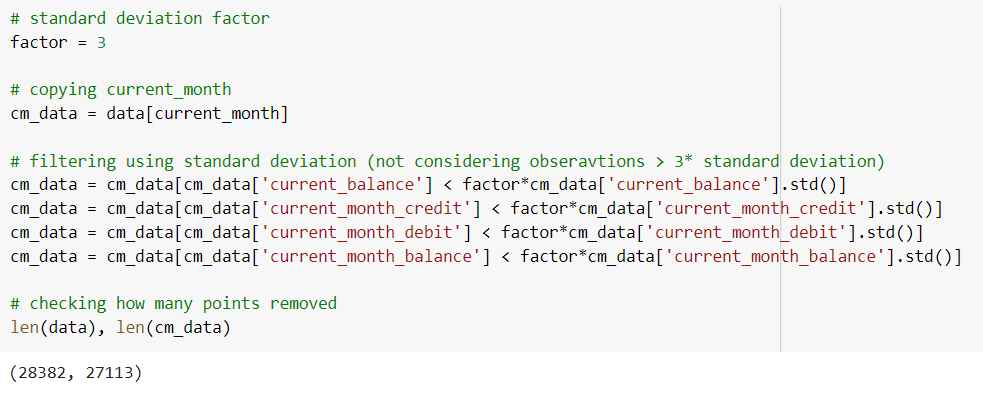
**Previous month:**

Fig 31

****This looks very similar to current\_month. Most of the customers perform low amount transactions.

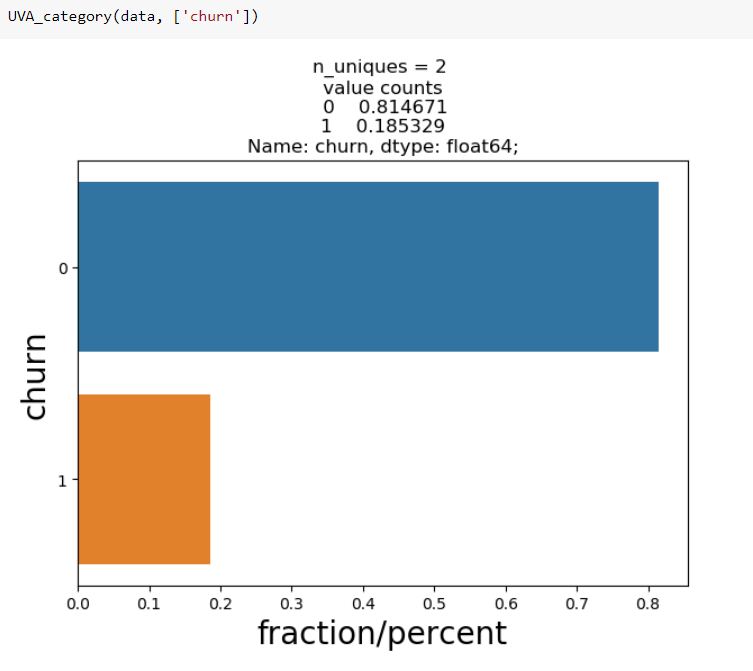
**Churn:**

Fig 32

Number of people who churned are 1/4 times of the people who did not churn in the given data.

**Univariate Missing values:**

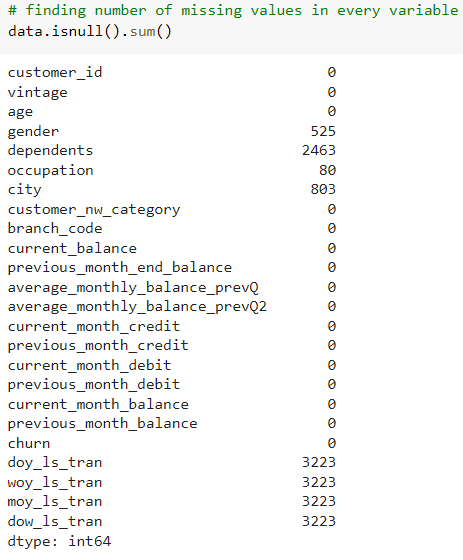
****

Fig 33

**Bivariate Analysis:**

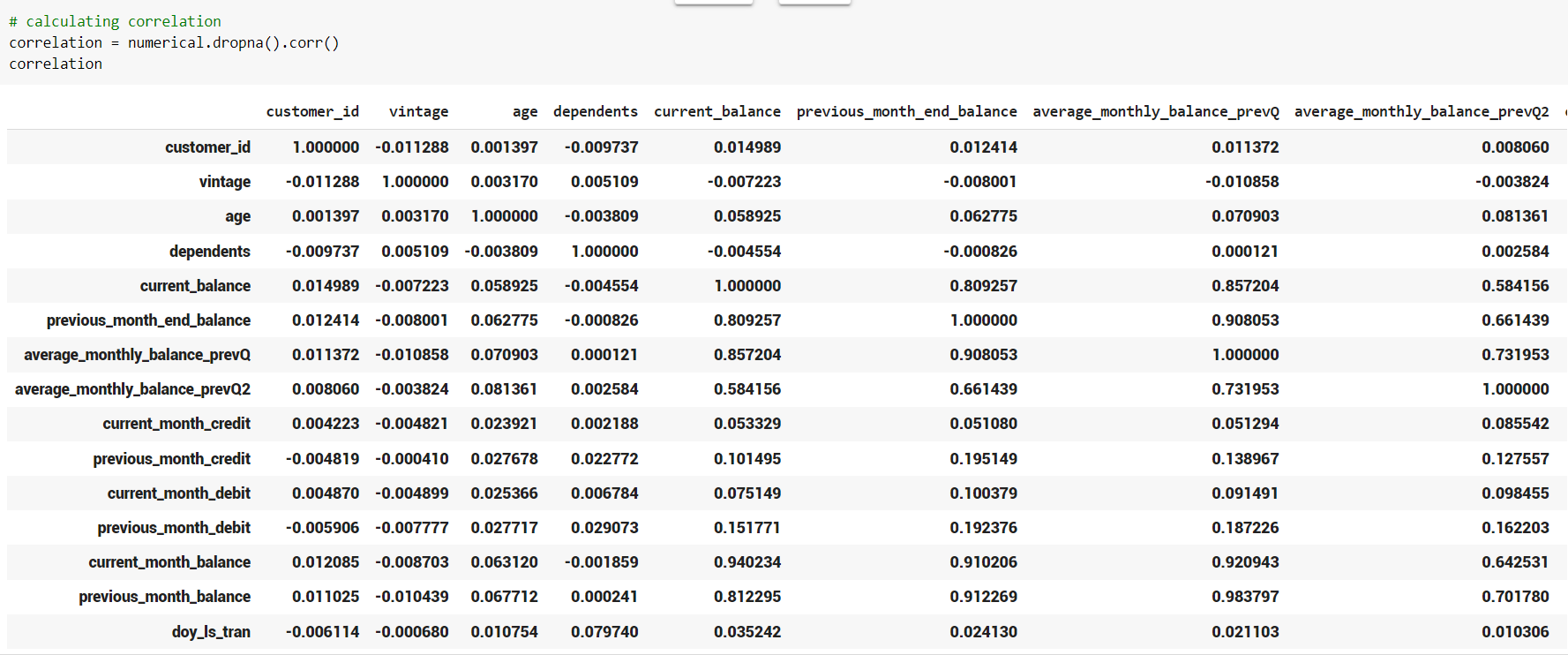
1. **Correlation Matrix:**

Fig 34

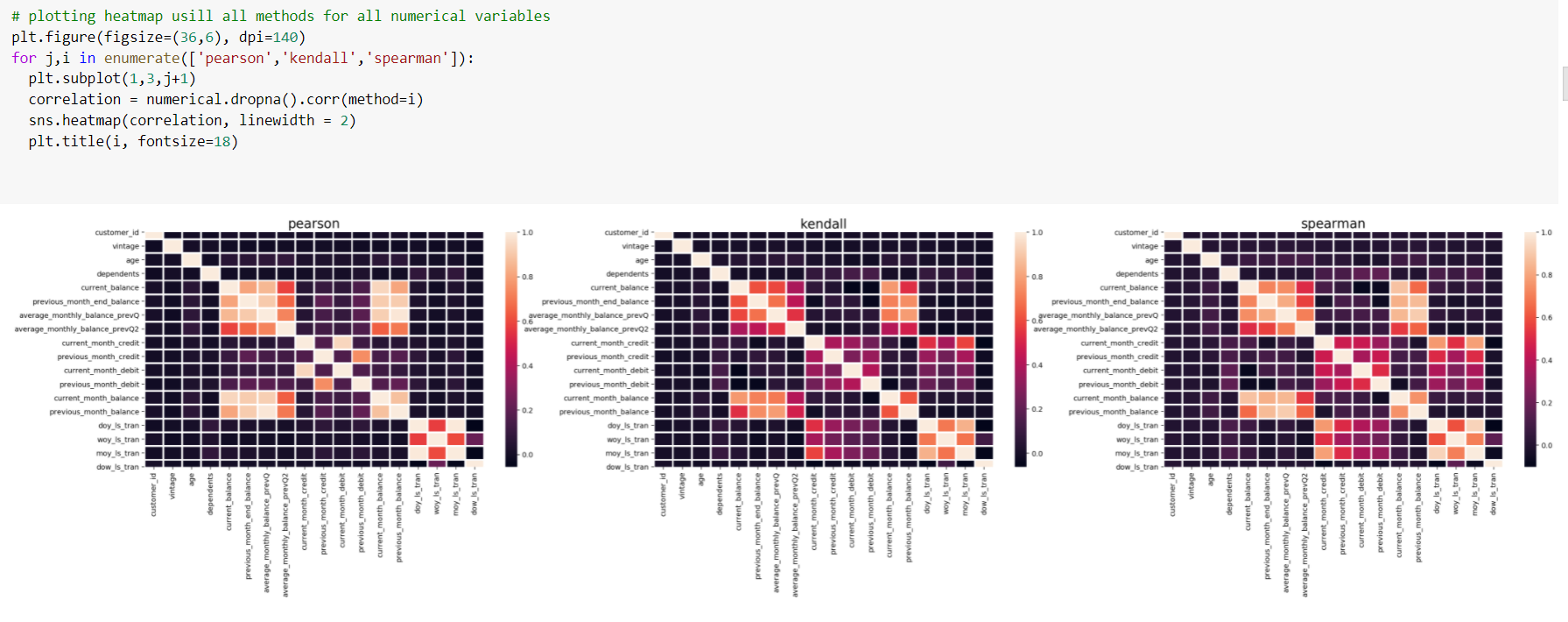
1. **Heatmaps:**

Fig 35

Kendall and Spearman correlation seem to have very similar patterns between them, except the slight variation in magnitude of correlation.

Too many variables with insignificant correlation.

Major correlation lies between the transaction variables and balance variables

Transaction variables like credit/debit have a strong correlation among themselves.

Balance variables have strong correlation among themselves.

Transaction variables like credit/debit have insignificant or no correlation with the Balance variables.

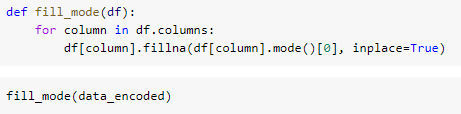
**Encoding:**

Encoding the variables using get dummies pandas function so every variable has numerical value attached to it.



**Missing values with mode:**

Filling missing values with the mode of the data



Segregating dependent and independent variables

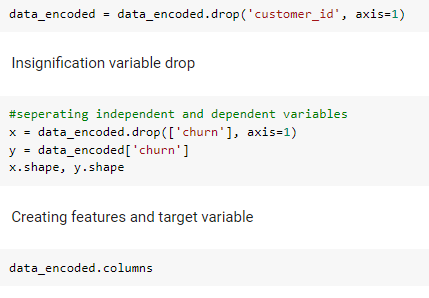


Fig 36

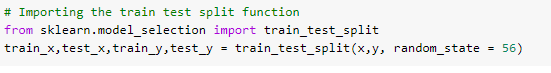
**Splitting data to train and test set:**

Fig 37

**Normalising using max\_min\_scaler:**

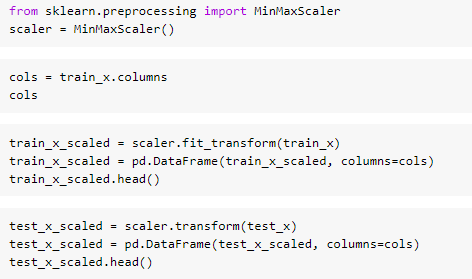
Transform **features by scaling each feature to a given range**. This estimator scales and translates each feature individually such that it is in the given range on the training set.

Fig 38

**Model Building:**

**Logistic Regression: Logistic regression** is a process of modelling the probability of a discrete outcome given an input variable. The most common [logistic regression models](https://www.sciencedirect.com/topics/computer-science/logistic-regression-model) a binary outcome; something that can take two values such as true/false, yes/no, and so on.

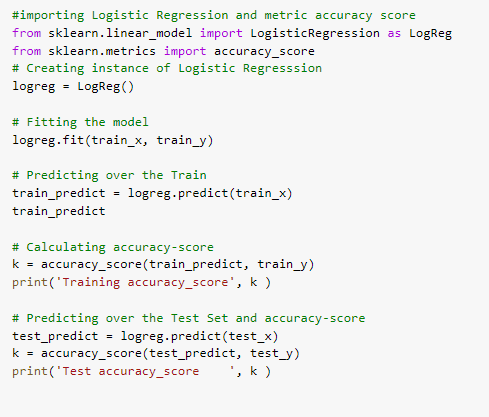


Fig 39

****

Fig 40

Logistic regression works pretty well and able to provide accuracy more than 80% for both training and test set.

**Regularization:**

Regularizations are techniques used to reduce the error by fitting a function appropriately on the given training set and avoid overfitting.

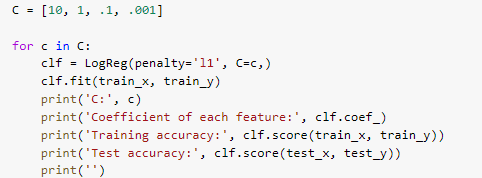


Fig 41

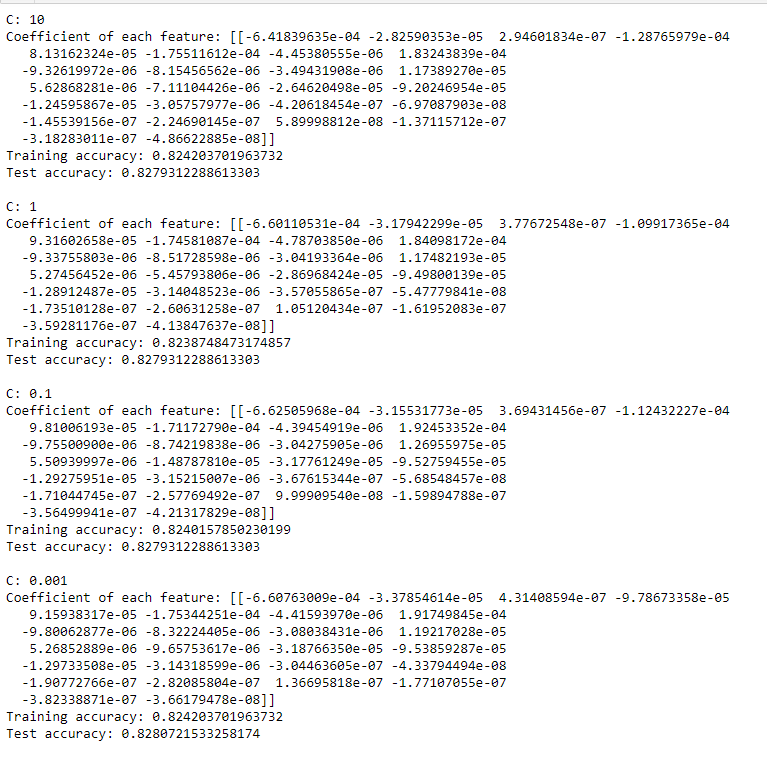


Fig 42

* As the value of C increases, Regularization constant decreases (C is inverse of Regularization Constant)
* So, most of the features associate themselves with 0 coefficient value
* This can be a feature selection technique as well while building model.

**CONCLUSION:**

Churn prediction by analysing the data is done. Considering all the factors affecting, customer churn is predicted using machine learning techniques.

The different types of Machine Learning methods including Linear Regression, and Logistic Regression are compared and analysed for optimal solutions.

**CHAPTER 4**

**REFLECTIONS**

During my internship at *Internshala*, I was fortunate enough to get experience and learn the different sides of the working environment and I am thankful for having got the opportunity to get trained with them.

Successfully completion of any type of technology requires helps from a number of people. I have also taken help from different people for the preparation of the report. Now, there is little effort to show my deep gratitude to those helpful people.

First, I express my sense of gratitude and indebtedness to our Training mentor – Kunal Jain, from the bottom of my heart, for his immense support and guidance throughout the training. Without his kind direction and proper guidance this study would have been a little success. In every phase of the project his Supervision and guidance shaped this training to be completed perfectly. I’m very grateful for choosing such an informative training. I really hope that this training helps me in the future to pursue my career in the field of Machine Learning.

**4.1 INDUSTRIAL TRAINING OUTCOMES**

Industrial training outcomes are:

* Develop relevant programming abilities in Python.
* Demonstrate proficiency with statistical analysis of data.
* Execute statistical analysis with professional statistical software.
* Demonstrate skill in data management.
* Develop the skill to build and assess data-based model.
* Understand the concept and application of a Machine Learning Algorithm.
* Apply Machine Learning concepts and methods to solve problem in real-world contexts and communicate these solutions effectively.

**4.2 NON-TECHNICAL OUTCOMES**

The non-technical outcomes that an industrial training will teach are:

* Behaving Professionally.
* Behaving ethically.
* Listening effectively.
* Time Management.
* Being attentive throughout the lecture.
* Allocating time effectively.
* Adapting effectively to changing conditions.

**4.3 TIME MANAGEMENT**

Time management may be aided by a range of skills, tools, and techniques used to manage time when accomplishing specific tasks, projects, and goals complying with a due date. Time management is usually a necessity in any project development as it determines the project completion time and scope. So, it is necessary to come up with a technique to manage our time.