Literature Survey

**Team Members:**

1. Shrutesh Reddy L - PES2UG20CS464
2. Sumant Kulkarni - PES2UG20CS469
3. Akash Hegde - PES2UG20CS492

**Research papers studied by Sumant Kulkarni (PESUG20CS469)**

**LITERATURE PAPER 1 :**

TITLE OF RESEARCH PAPER : Research on a Customer Churn Combination Prediction Model Based on Decision Tree

* Customer Churn Prediction is a prominent issue facing companies. Preventing customer churn, trying to retain and retain customers has become an very serious issue for business operations and development.Most of the current customer churn predictions use a single prediction model, which makes it difficult to accurately predict customer churn. Based on the prediction results and confidence of decision tree this paper designs a combined prediction model of customer churn and conducts empirical research on the effectiveness of the model. The prediction results show that compared with the single customer churn prediction model, the combined prediction model has higher accuracy and better prediction effect, and can more intuitively display the basic characteristics of the churn customers.
* Based on the information gain theory, the decision tree is one of the most widely used data classification algorithms at present.The decision tree prediction process is performed in two steps:one is to build and evolve a decision tree using the training set; the other is to test the attribute values of each node, classify the input data, and use the attribute values of this class to complete the estimation of the prediction object
* Aiming at the problem that a single model is difficult to achieve high-precision customer churn prediction, this paper uses the prediction results and confidence of decision tree prediction model and neural network prediction model to build a combined customer churn prediction model. The empirical results show that the combined prediction model can not only have a better interpretation ability like a decision tree model, but also a higher prediction accuracy rate of a machine learning model, which can better make up for the shortcomings of a single prediction model, and can also get more stable and accurate prediction results.

**LITERATURE PAPER 2:**

TITLE OF RESEARCH PAPER : A Comparative Study Using KNN and Decision Trees

* Churn prediction represents one of the most important components of Customer Relationship Management (CRM).Customer Relationship Management (CRM) represents the main focus of attention for many organizations around the world, due to the competitive nature among these organizations. In the purpose of retaining customers and maintaining their satisfaction, researchers of many fields including business intelligence, marketing and information technology were motivated to investigate the best methods that deliver the best services for customers.Machine Learning algorithms had been implemented in the purpose of optimally predicting the possible churning customers and making the right decisions at the right moments. Researchers had conducted several studies on various types of algorithms and results were found very promising. Customer Relationship Management (CRM) represents the main focus of attention for many organizations around the world, due to the competitive nature among these organizations.
* Many challenges have been faced by researchers while testing the techniques and algorithms to extract the optimal churn rate and identify churn customers. One of the most complicated issues is dealing with incomplete datasets that have missing values .In addition, large datasets can be hard for researchers to organize and handle because of the existence of noisy data records. Decision tree technique is one of the most used classification algorithm due to its ability to operate in an environment where there are large amounts of noisy data, and its ability to avoid over-fitting issue in classification .Study focuses on the performing comparison between Decision Trees and K-Nearest Neighbor classification algorithms. Both of the algorithms have the same input data with the same training and testing ratios. Input data for both algorithms is processed through filters to reduce noise and remove undesired data. Afterwards, the cleaned data is split into training and testing sets. Training sets are then modeled using algorithms to deliver the desired output.

**LITERATURE PAPER 3 :**

TITLE OF RESEARCH PAPER : Machine Learning Based Telecom-Customer Churn Prediction

* Customer churn or attrition refers to the percentage of customers who will discontinue with a company service during a given time frame. Churn rate is calculated by dividing the number of customers a company lost over a given period of time by the number of retained customers at the beginning of that time period. Churn prediction is a key predictor of the long term success or failure of a Business. In this research, machine learning techniques are explored in order to predict telecom customer churn. Ubiquitous techniques like Random Forest Classifiers and SVMs are compared with relatively newer architectures like XGBoost and Deep Neural networks to classify if a customer will churn or not. The efficiency of these models are further explored by passing them through a grid search. The Random Forest model works best for this particular use case with a prediction accuracy of 90.96% on the testing data before grid search.
* Churners are customers who will be switching from one telecom service provider to another. Prediction of telecom churners has been an area of interest for researchers and many researchers have worked on various techniques to predict telecom customer churn. Telecom industry has been battling the threat of losing more than 25% of its customers every year, which is believed to result in huge revenue loss. Random forest performs well on a large dataset and handles missing variables without deletion of variables.Feature selection is performed on the dataset and a few columns are dropped.The dataset used was focused on a customer retention program which included various customer attribute fields and also a column of customer churn.
* The Churn Prediction is carried out pre-processing steps on this data set and the same was used as input data for implementing all the techniques.

**Research papers studied by Shrutesh reddy L (PESUG20CS464)**

**LITERATURE PAPER 1 :**

TITLE OF RESEARCH PAPER :Customer Churn Prediction In Telecommunication Industry Using Random Forest Classifier

* Nowadays data has become the important aspect in each and every field. In this the data about the telecommunication industry is collected and then the raw data is classified into churn and the non churn customers.In existing system they uses the algorithm called LDT and UDT which train the system blindly with too many attributes which are not necessary for the computation and performance achieved is about 84 percent.To resolve this problem in existing system we use an efficient algorithm known as Random Forest Classifier whose efficiency is about 95 percent.
* In churn prediction all the data should to labeled in order to make the corret decision to retain the customers the core concept of this project.
* Churn prediction is one of the best way in finding the feedback from the customers and it sustain the customers by giving some offers.
* The churn prediction is the process of collecting the customer feedback about the product and tries to retain the customer within their organization.
* In this research paper it has been told that RANDOM FOREST achieves most efficiency.
* Accuracy=(TP+TN)/(TP+TN+FP+FN)

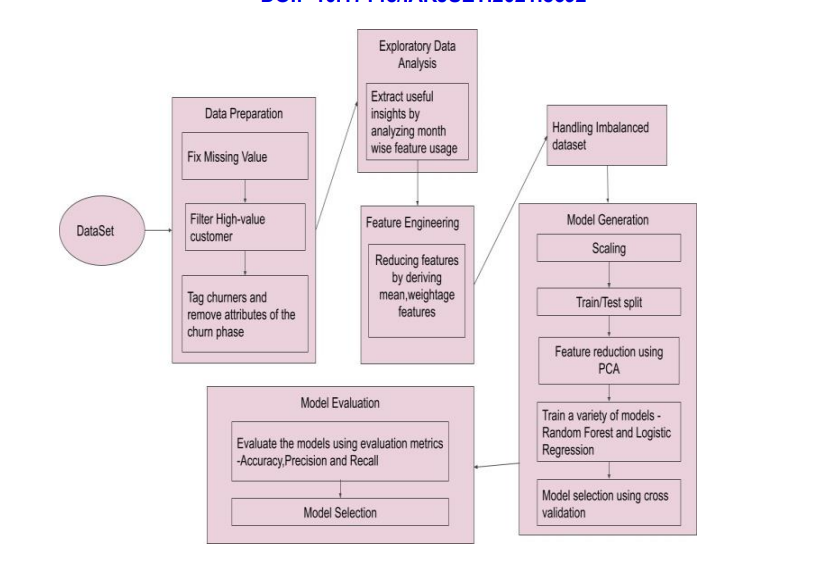
**LITERATURE PAPER 2 :**

TITLE OF RESEARCH PAPER :Customer Churn Analysis and Prediction in Banking Industry using Machine Learning

* The Banking industry faces challenges to hold clients. The clients may shift over to different banks due to fluctuating reasons, for example, better financial services at lower charges, bank branch location, low-interest rates, etc. Thus, prediction models are utilized to predict clients who are probably going to churn in the future.
* In this paper, different models of machine learning such as Logistic regression (LR), decision tree (DT), K-nearest neighbor (KNN), random forest (RF), etc. are applied to the bank dataset to predict the probability of customer who is going to churn.
* By exploring raw data and performing data analysis, one can predict in advance whether the customer agitates or proceed with the organization.
* Different models mentioned in this research paper are:
* Decision tree (DT): It a type of supervised learning where a target variable is also defined. DT is mostly used in classification problems.The tree is split into internal nodes and leaf nodes based on splitting criteria. T
* K nearest neighbor (KNN): It is also known as a lazy learning algorithm. KNN is used for both regression and classification predictive problems.
* Random Forest (RF): It makes use of the multiple DT form a king decisions. It combines the output of multiple (randomly created) DT to generate the final output as it uses the concept to fensemble learning.

**LITERATURE PAPER 3:**

TITLE OF RESEARCH PAPER :CUSTOMER CHURN PREDICTION.

* Many machine learning techniques and algorithm plays an important role for companies in today’s commercial conditions because gaining a new customer’s cost is more than retaining the existing ones.
* The customer who cease a product or service for a given period is referred as churner.
* It is very crucial to make the data useful because unwanted or null values can cause unsatisfactory results or may lead to producing less accurate results. In the data set, there are a lot of incorrect values and missing values. We analyzed the whole dataset and listed out only the useful features.
* 
* Feature selection is a crucial step for selecting the required elements from the data set based on the knowledge. The dataset used here consists of many features out of which we chose the needed features, having only significant features and reducing the number of irrelevant attributes increases the performance of classification.
* Load the dataset and print the first 5 records of the dataframe to check the loaded dataset.
* Evaluate the Random Forest model based on their performance metrics like accuracy, precision and recall that is more important to identify churners than the non-churners accurately.

**Research papers studied by Akash Hegde (PESUG20CS492)**

**LITERATURE PAPER 1 :**

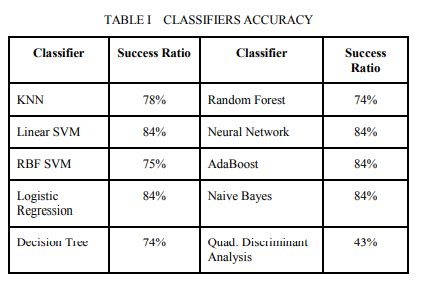
TITLE OF RESEARCH PAPER : “Research on Customer Churn Intelligent Prediction Model based on Borderline-SMOTE and Random Forest."

* With the rapid development of bigdata and development in the world’s finance, the country’s banks face very serious challenges and with the pressure is gradually increasing. With the rise of payment apps and other software storage apps , the most serious problems encountered by banks is the loss of customers . Nowadays, research on customer churn prediction is developing day by day, and the proposed models are getting more optimized.
* For good classification prediction ability, Hu Yong pei proposed a customer churn prediction model based on AP clustering and random forest. The results show that this model’s actual prediction result is more accurate than general decision trees.
* In the bank customer data, it is found through research that churns customers often account for a small part of all customers, so the customer data presents a typical imbalance.
* It is essential to find a model that can deal with data imbalance and have higher classification accuracy. Among the existing methods for dealing with unbalanced data, the BorderlineSMOTE algorithm is an improved oversampling algorithm, which has a better processing effect on the data.
* This paper proposes balancing data and training a random forest model using the Borderline-SMOTE method to predict bank customer churn. The OOB error rate unique to the random forest, AUC, Precision, Recall, and F-mean commonly used by classifiers are used as the evaluation indicators of the model. At the same time, KNN, single decision tree and Naive Bayes are used as the comparison objects of this model.
* The model is trained after using the oversampling method to process the data. The experiments show that the prediction model combining the Borderline-SMOTE method and random forest proposed in this paper has the best overall results. When predicting unbalanced data such as bank customers, which is effectively balanced, the performance is improved by about 4% compared with other prediction models. In future work, it is considered to improve the characteristics of bank customers and improve some algorithms of the essential devices in the model, so that the model’s performance in solving such problems is completed.

**LITERATURE PAPER 2 :**

TITLE OF RESEARCH PAPER : “Towards Accurate Predictions of Customer Purchasing Patterns.”

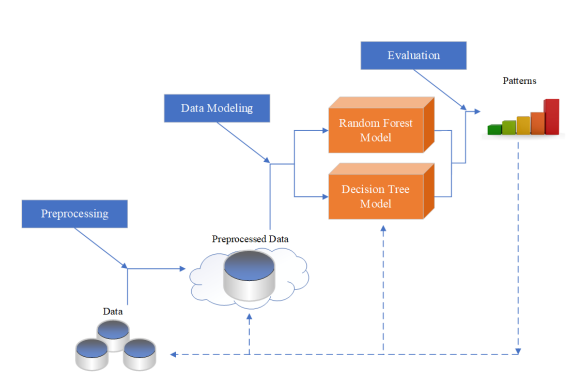
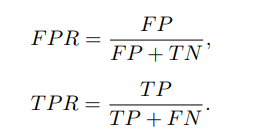
* Retailers are becoming increasingly aware of the value of their data and the importance of customer relationship management (CRM). Connecting customer data with wider sources allows customer identification, segmentation, product association, prediction, visualization of trends, recommenders, loyalty programs and so on.
* The predictive capabilities of the classifiers were assessed using linear regression, Lasso and regression trees. Prediction accuracy on untrained customers was generally better than 80%. The models implemented (and compared) for classification were: Logistic regression
  + - * Linear SVM
      * Decision Trees
      * Random Forest Classifier
      * Multi-Layer Perceptron (Neural Network)
* Postcode data was then used to classify solely on demographics derived from the UK Land Registry and similar public data sources. Prediction accuracy remained better than 60%.
* The Churn Rate, or rate of attrition, is the percentage of customers who make no further purchases within a given time period. For purposes of this study, we define it as the proportion of customers who, having made a purchase, make no further purchases in the following 12 months. For this topic the related terms are : Customer lifecycle value (CLV) and Customer Relationship Management.
* Many experiments had been conducted on many customers and the results of their success ratio on different classification models are recorded and they are given below :



* In this paper we have shown the feasibility of identifying the probabilities of customer purchase repetition, making it possible to target important customer groups. We have further shown how public data sources can be used to augment internal data and thereby achieve improved marketing and profitability. Thus far we have only scratched the surface. In future we intended to conduct more detailed analysis by adopting unsupervised and more importantly exploratory techniques to further our understanding of the factors that influence customer behaviour in a more generic context.

**LITERATURE PAPER 3 :**

TITLE OF RESEARCH PAPER : “Customer Churn Prediction by Classification Models in Machine Learning.”

* A classification problem is when the output variable is a category, such as “churn” or “non-churn” or “spam” and “not spam” . In classification problems, some information is drawn based on input data, which the output variable can be identified according to the information. The classification model predicts the category of unknown data. Classification problems are widely applied, for example:
* assessing the risk of diabetes among individuals
* classifying the user as churn or not
* classifying the emails into spam or not spam.
* In this paper, decision tree and random forest models are employed to predict customer churn. Actually decision tree is a graphical representation of the relationships among different properties for classification tasks. And the random Forest means, an effective ensemble learning method, which can be used for classification tasks. A random forest consists of many decision trees, where each decision tree corresponds to a category.
* The random forest classifier has many applications, in them few are mentioned below:
  + - * big data classification
      * churn prediction and
      * intrusion detection
* For predicting the customer churn rate , first we have to do the data processing .
* This includes feature selection and sample partition. After this we have to do the data cleaning part that means, removing the NaN values and dropping the unnecessary columns from the dataset.
* After this we have to do the data modelling, which includes the decision tree classifier, random forest classifier, logistic regression and many.
* The total churn prediction model that has been built is captured in an architectural style. That is given below:
  + 
* We can calculate FPR and TPR ratio using these formulas:
  + - * + 
* In this paper, a decision tree model and a random forest model were employed to predict customer churn for a company operating in the chemical industry. The underlying risk of customer churn and the factors of customer churn were analyzed to appropriately train the decision tree and random forest models. Our experimental results indicated that the most important churn factor was low-priced count (LC).
* The higher the number of LC customers, the lower the probability of churn. A comparative evaluation revealed that the random forest model had higher prediction precision than the decision tree model when considering the training error and generalization error. We could conclude that the random forest model had a good fit to the data, compared with the decision tree model.
* Overall, the random forest model was superior to the decision tree model in predicting the unknown instances in the testing set. In the future, we will compare the quality of the decision tree and random forest models through prediction of the customer churn volume