A Machine Learning Approach CUSTOMER CHURN PREDICTION IN BANKING INDUSTRY

Vibhore Singh   
Master of Data Science  
University Of CanberraCanberra, Australia  
u3248455

Meenakshy Baiju   
Master of Data Science  
University Of CanberraCanberra, Australia  
u32

*Abstract*— Customers are the most valuable asset for a business such as banking and customer churn is the biggest concerns the banking industry is facing today. Bhattacharya in [1] suggests that retaining a customer is about five to six times less expensive than acquiring a new one. So, it is important to identify which customers are going to churn and take necessary actions to reduce this. This project involves the development of a machine learning based model to predict customer churn in a dataset of banking customers with the objective of helping banks retain their customers. We use the dataset containing details of the demographic, banking activity, financial status etc. of 10,000 bank customers and apply various ML models to predict the probability of customer who is going to churn. The research culminates with the comparison of the models based on accuracy and F1 statistic and Random Forest emerged as the most efficient model. The project employs a Pattern Recognition and Machine Learning (PRML) approach to identify churn-prone customers.

Keywords—Machine Learning; Churn Prediction; Pattern Recognition; Banking Analytics

# Introduction

Customer churn is when customer stop using a company’s service. By reducing the customer churn by 5%, we can increase the profits of an enterprise by an upwards of 25%. [2]. Customer churn is one of the most significant factors responsible for impacting not only the bank's profits but also the reputation. [3] As compared to the average revenue per user (ARPU) which is one of the most significant key performance indicators, churn has a greater impact on the customer lifetime value (CLTV). [4]

ML tools can learn from the historical data to identify patterns that are associated with the customer churn. Computational ML techniques help companies to build accurate customer profiles based on customer behavior and analyse this data to get insights so that the businesses can predict which customers are likely to churn, thus banks can employ retention strategies and improved CRM. [5]

In this paper a predictive model is proposed to find the patterns in the dataset of 10,000 bank customers that can help to detect the early signs of churn in customers and give predictions based of factors such as financial status, bank activity, demographic etc. and the banks can take preventive measures to prevent churn. Since we have an unbalanced dataset, our preferred algorithms for this project are Logistic regression (LR), decision tree (DT), Support Vector Machine (SVM), random forest (RF). Along with those, some feature selection and preprocessing techniques make our predictions more accurate.

The remainder of paper is organized as follows:

Section II provides the background and related work. Section III explains the methodology. The Results are presented in Section IV and section V provides the conclusions.  
  
**Questions to be investigated:**

1. Can we identify customers who are likely to leave the bank?
2. Can we identify the factor that has the strongest effect on churn?
3. Can we use the models on real time data by analysing the historical patterns?
4. Which is the most efficient machine learning model for this task?

**Goals**:

1. To identify customers who want to churn.
2. To identify the factors responsible for their churn.
3. To choose a model that will efficiently classify the customers as churn/non-churn.
4. To give predictions for the future based on the previous bank data.
5. To avoid future churn by addressing reasons responsible for churn.

# Background and Relevant Work

Numerous research studies have been conducted to predict customer churn in the banking industry by using popular machine learning algorithms like Artificial Neural Network, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression. The prediction models are used to forecast churn customers in the future. These models make use of historical data from previous churners and seek to find patterns with current clients. If any similarities are discovered, the current clients are then categorised as potential churners.

In [5], the authors stated the significance of balancing the data while conducting classification as imbalanced data resulted in low recall scores. After using the Synthetic Minority Oversampling Technique (SMOTE), Random Forest proved to be the best in case of evaluation metrics with 85% accuracy. Moreover, it was observed that age and membership activity were the most salient [6] features.

The authors proposed a [5]Deep Neural Network model for anticipating customer churn of a retail bank in Iran. Here, the model trains the data for 30 days and then predicts the churners in the next month. To compare, Traditional Machine learning algorithms were utilised. DNN model overperformed other classifiers such as Logistic Regression, Decision Tree, and Naïve Bayes with 84% accuracy rate. According to the researchers in [5] , while CART technique revealed higher prediction success rate at 95.01% on training data and 91.22% on test data, C 5.0 showed low results. Therefore, CART is more useful to identify customer characteristics that significantly influences churn.

The article [7] reported that when compared in terms of improvement exactness, Artificial Neural Network model surpassed Decision Tree model by achieving 86.52% accuracy in anticipating clients who are high likely to terminate their loyalty to the bank. According to the authors in [8], Support Vector Machine is extremely effective for predicting VIP customer churn of a commercial bank. However, deciding the values for parameters of SVM regression and selecting the most appropriate kernel continues to be a challenging task.

The authors of conference paper [3] utilised Decision Tree, Random Forest, and Logistic Regression to conduct customer churn analysis. Random Forest classifier performed better in areas such as stratified sampling, without stratified sampling and 8-fold cross-validation. In [9] , it was explored by the researchers that customer segmentation lays no influence on prediction of customer attrition and Random Forest at 97.25% revealed the highest accuracy score. The researchers found the most appropriate model for churning in the banking sector based on ROC Curve. Random Forest Model with 86.312% accuracy acquired most of the area under the ROC curve than Logistic Regression and Support Vector Machine.

# Methodology

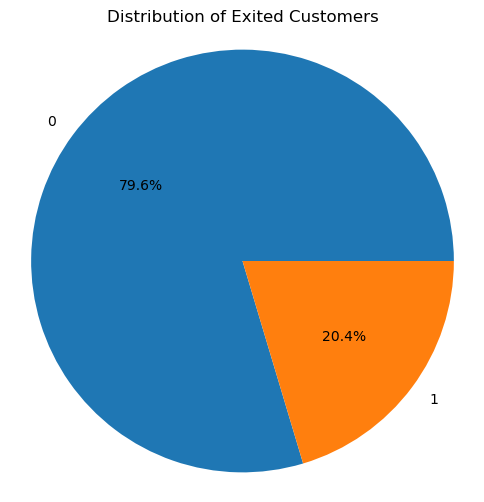
## Data : The data used in this project has been acquired from Kaggle. This dataset contains data of customers in the banking sector. It gives us insights about the patterns in banking churn and different features which may affect their decision. In this dataset we have data of 10,000 customers with the information about the features namely ‘RowNumber’, ‘CustomerId’, ‘Surname’, ‘CreditScore’, ‘Age’, ‘Geography’, ‘Gender’, ‘HasCrCard’ and ‘IsActiveMember’ , ‘Tenure’, ‘NumOfProducts’, ,‘Balance’,‘EstimatedSalary’ and ‘Exited’. Exited is our target variable which gives us the information whether the customers stayed or churned.

## Data Pre-processing: Prior to the data analysis and model building, the data needed to be pre-processed to ensure that it is clean, complete, and in a suitable format for analysis. We checked the dataset for themissing values and fortunately we had a clean dataset, and it didn’t have any missing values. Next for the feature selection part we identified few variables that didn’t have a significant influence on the dependent variable. As a result, we removed the redundant variables such as "RowNumber," "CustomerId," and "Surname".

Next in the preprocessing phase we encoded the remaining categorical variables into numeric to make them compatible with our models. We used Label encoding and One Hot Encoding fo the variables Gender and Geography respectively.

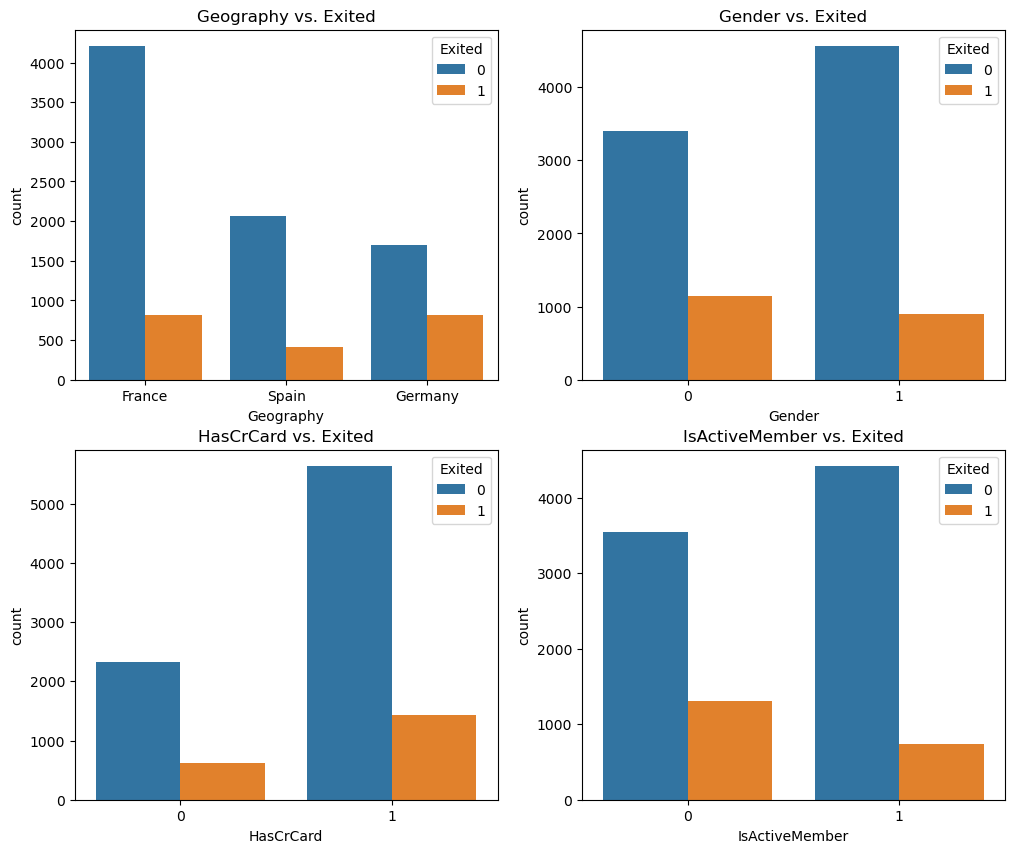
## Exploratory Data Analysis (EDA): To understand the distribution of data and relationship between its different variables, we performed EDA. First we checked the summary statistics of the data for the five point summary of the data.

## To get the clear distribution of our target variable and the representation of the proportion of customers who churned or stayed, we used a pie chart. The pie chart in Figure 1. Clearly displays the significant class imbalance we have in our dependent variable. The majority of the customers have not churned accounting for 80 % of the total data .



*Figure1*: Distribution of the target variable ‘Exited’

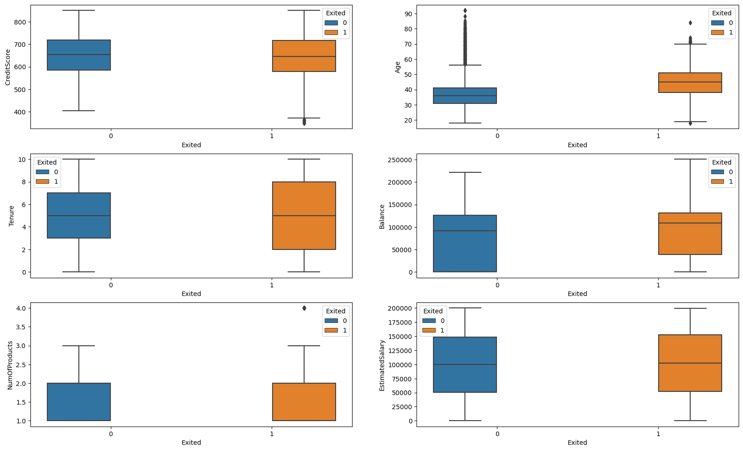
Next, we plotted count plots to visualize the distribution of some variables with respect to the target variable. Figure 2 shows the distribution Geography, Gender, Credit card owners, and Member activeness against the Exited variable to get insights about these variables.



*Figure2:* Distribution of Categorical Variables vs Exited.

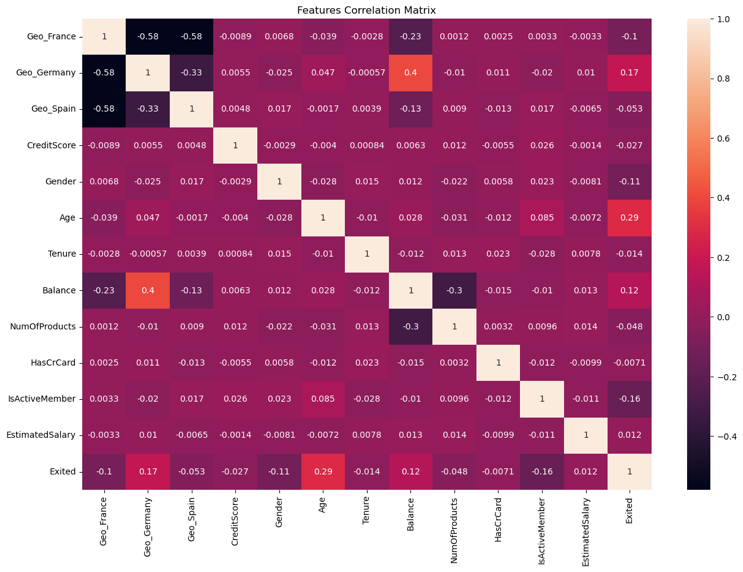
For our numerical variables, we plotted them in boxplots to visualize their distribution with respect to the dependent variable. This helps us to visualise the difference between the distribution of a particular feature between customers who churned and those who stayed and look for outliers in the data. Figure 3 clearly shows that credit score, Number of products and salary variables don’t have any significant

effect on dependent variable whereas Age and Tenure do affect the churn rate in a bank.



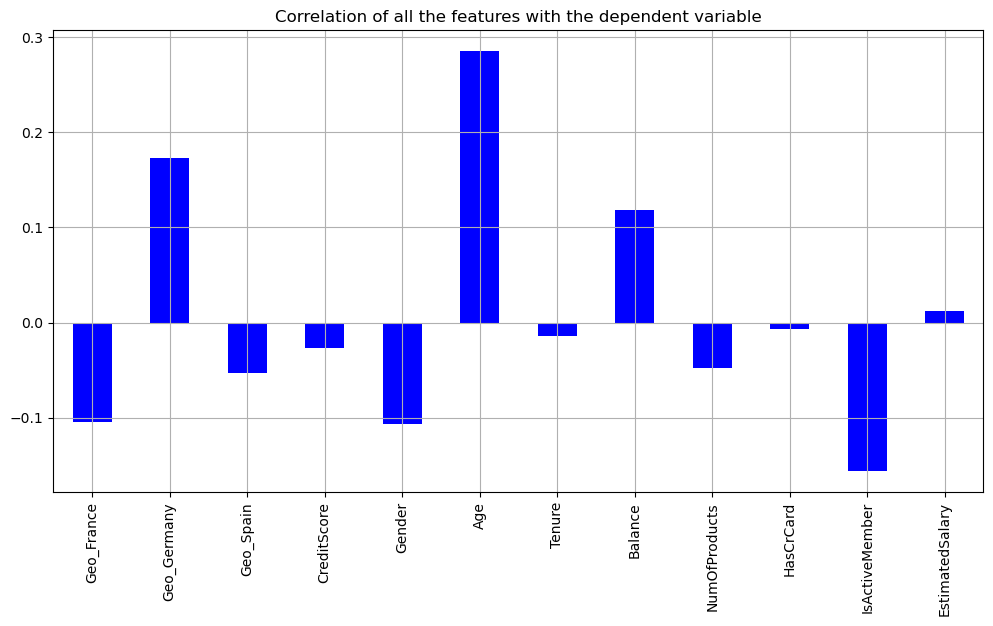
*Figure3:*  Distribution of Categorical Variables.

To visualize the distribution among different variables we made a correlation heatmap. The heatmap confirms that Age and IsActiveMember have the highest correlation with the dependent variable. Figure4 below gives us the precise numeric values representing the relationship between different features.



*Figure 4:* The correlation matrix heatmap

Next, we find the correlation between every feature and the target variable i.e Exited. This will help us in understanding the features which have stronger relationship with the target variable both positive and negative. Figure 5 confirm again that the age has the highest positive correlation with the Exited variable. IsActiveMember has the second highest correlation with the target variable but in negative direction.



*Figure 5:* Correlation of all features with Exited

*D. Model Building* : We start this step by splitiing the data. First we separate the the dependent and independent variables and then then split them into training and testing sets. We split our dataset in the ratio of 80-20. Our models train on the 80% training sets and then predict the values for the 20% testing data so that we can evaluate their performance based on different metrics.  
  
Our next step is to train the models by fitting them over the training set with the default parameters. We use Logistic regression, Random Forest, Decision Tree and support Vector Machine for this project. We choose these algorithms keeping in mind the class imbalance, binary classification and overfitting chances in our data. This phase is known as model training. During this phase, the models examine the training data to find patterns and relationships within the dataset. These patterns enable the models to make predictions on new, unseen data.

*E. Model Evaluation*

Once our models are trained, we move on to the prediction phase where the values for the class are predicted for the testing data. The relationships that the model has learnt in the trining phase are used to find the similar connections in the testing data variables. The models output the predicted values with respect to the patterns it learnt from the data during the training phase.

Then we compared the the predicted values of the model with the actual values for the classes and evaluate the model performance and effectiveness of the trained models. Due to the imbalance in our target variable, we couldn’t rely solely on accuracy . In this project we used Accuracy and F1 score to evaluate our models performance. Table 1 gives the classification report for all the models.

Table 1: Classification Report

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Precision** | | **Recall** | | **F1 -score** | | **Accuracy** |
|  | Class0 | Class1 | Class0 | Class1 | Class0 | Class1 |  |
| LR | 0.83 | 0.58 | 0.96 | 0.24 | 0.89 | 0.34 | 81.1% |
| DT | 0.89 | 0.51 | 0.86 | 0.58 | 0.87 | 0.54 | 80.2% |
| RF | 089 | 0.75 | 0.95 | 0.52 | 0.92 | 0.62 | 86.7% |
| SVM | 0.87 | 0.8 | 0.97 | 0.44 | 0.92 | 0.57 | 86.5% |

Based on the Table 1 values, we identify Random forests and SVM are the best performing models, but we select Random Forests as the best model for this stage of the project as it has highest accuracy of 86.7%, highest F1 score and a good precision and recall balance. Support Vector

Machine stood close to RF with almost similar accuracy, but Random Forest has a higher F1 score (0.62) than SVM (0.57), which suggests that it is a better model for this task.  
  
Next we perform Hyper parameter tuning on all our moels to improve their performance. This will help us to build a robust model which will work efficiently on new unseen data.

*F. Hyper parameter Tuning:* Hyper parameter tuning is the process of getting the best set of Hyperparameters for any model. We selected a range of values for our hyperparameters and optimize the models with the best combination of all these hyper parameters that results in the best model performance.

We used both GridsearchCV and RandomizedsearchCV to optimize our hyperparameters. We created a grid for different combinations of hyperparameters and then used cross validation method to evaluate the model with each of those combinations.

The optimal parameters we recieved due to hyperparameter Tuning are as follows:

* Logistic Regression : C= 1, penalty = l1, solver = liblinear
* RandomForestClassifier: max\_depth=100, max\_features=3, min\_samples\_leaf = 3, min\_samples\_split = 8
* DecisionTreeClassifier : criterion = entropy, max\_features = 6, min\_samples\_leaf = 6
* SVC : C=1, gamma = 0.5, kernel = 'poly'

Table 2: Results of hyperparameter tuning

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Grid Search Cv | | Randomized Search Cv | |
|  | Accuracy | F1 score | Accuracy | F1 score |
| LR | 81.1% | 0.336 | 81.1% | 0.336 |
| RF | 86.90% | 0.610 | 86.95% | 0.612 |
| DT | 82.95% | 0.568 | 83.90% | 0.566 |
| SVM | 86.5% | 0.597 | 86.45% | 0.59 |

*G. Best Model Selection :*  Based on the results of Hyper parameter Tuning Among the models considered, the Random Forest Classifier (RF) stands out as the best standalone model in terms of model performance, with an accuracy of 87% and an F1 score of 0.62. RFis a combination of several decision trees is a powerful emsenble technique and a very intutive way to deal with class imbalance. It is excellent at managing overfitting, handling intricate relationships in the data, and producing reliable findings. RF with the optimized hyperparameters, offers a robust solution for the predicting the banking churn. The highest accuracy and F1 score achieved among the models tested, making it the best standalone model for this specific problem.

# Results

In this section, we present the results of our churn prediction model. We fit our models with the optimal hyper parameters that we got from the hyperparameter tuning. Table 3 gives us the accuracy and the F1 score obtained from best parameters for all the four models :

|  |  |  |
| --- | --- | --- |
| Models | Accuracy | F1 score |
| LR | 81.1% | 0.34 |
| DT | 83.4% | 0.56 |
| RF | 87% | 0.61 |
| SVM | 86.5% | 0.60 |

Table 3: Best Accuracies of different models after hyperparameter tuning

While analyzing the report, we can observe that Random Forest outperformed other models with accuracy rate and F1-score documented at 86.7% and 0.62 respectively.

To understand the impact of different train-test splits on model performance, we evaluated the Random Forest model with various test split sizes. The results are as follows:

|  |  |
| --- | --- |
| Split | Accuracy |
| 90:10 | 86.4% |
| 80:20 | 86.8% |
| 75:25 | 86.5% |
| 70:30 | 86.8% |
| 65:35 | 87% |

*V. CONCLUSION AND FUTURE WORK*

*Finally, this* research report has dived into the critical area

of bank churn prediction, which is becoming increasingly

relevant in the banking industry. Customer churn is a big

concern due to its negative influence on the financial

capacity of the bank. Churn prediction models were created

in our project to solve this problem.

Several visualization techniques and popular machine

learning algorithms such as Logistic Regression (LR),

Support Vector Machine (SVM), Decision Trees (DT), and

Random Forest were incorporated in our project for the

purpose of anticipating customer churn in the banking

industry. Each algorithm was carefully evaluated and

compared in terms of predictive accuracy and overall

performance. We also observed that cross-validation did

not change the accuracy and F1-scores of three of the

models.

Our analysis revealed that the Random Forest algorithm was the best-performing model, with an impressive accuracy rate of 86.7%. This finding exhibits Random Forest&#39;s ability t0 efficiently analyze a diverse set of features while adapting to

complex, non-linear relationships in the data. The

hyperparameter tuning for each model revealed the best

combinations of hyperparameters.

The findings of the project have significant implications for

the banking industry, as they provide an excellent

opportunity for reducing customer churn and its financial

consequences. In conclusion, this study accentuates the

advantages of machine learning algorithms, particularly

Random Forest, in the prediction of bank churn. These

findings [15] are useful for banks looking to improve

customer retention and reduce the negative effects of churn.

Since the fields of data science and machine learning are

dynamic, we can expect even more sophisticated models

and techniques that further enhance our ability to address

the financial sector&#39;s complex challenges.

References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**