PROCEEDINGS OF
INTERNATIONAL CONFERENCE ON ADVANCED TECHNOLOGIES
https://proceedings.icatsconf.org/
11th International Conference on Advanced Technologies (ICAT'23), Istanbul-Turkiye, August 17-19, 2023.

Defining The Decisive Factors on Purchase and Comparing Feature Importance Methods

Erman Demir
Graduate School of Science and Engineering,
Galatasaray University,
Çırağan Caddesi, Ortaköy , Türkiye
erman.demir@hepisburada.com

F. Serhan Daniş
Graduate School of Science and Engineering,
Galatasaray University,
Çırağan Caddesi, Ortaköy, Türkiye
sdanis@gsu.edu.tr

***Abstract*— Online retail companies focus on two activities for getting revenue in their businesses and to survive in the market. First activity is increasing traffic of the online shopping platform and second activity is converting this traffic to revenue for the company. Marketing facilities try to attract customers to the online shopping platforms at great costs. Because of the costs of getting traffic, it is crucial to make customers order. Online shopping platforms need to understand which factors are decisive on customer purchase decision. In this study, which factors are decisive on consumer purchase decisions will be studied on an e-commerce retail platform from Turkey, Hepsiburada. Which of these factors are most decisive try to be defined: traffic source type (google, campaign, direct etc.) of the customer, customer persona or segment, which types of page or page components has been seen, product position on the page, does the customer benefited from campaign or discount, product review scores and counts, has the product recommended or not. In this study, data will be gathered from Hepsiburada transactions stored in google's big query environment. Performance problems will be solved via SQL optimization and other methods. Data quality issues will be fixed to get consistent results. Then statistical methods, supervised machine learning and deep learning methods will be applied to data for getting feature importances. Importance value of the features will show which factor decisive on customer purchase decision. Feature importance values will be compared and evaluated according to method, model results. Hyperparameter tunings is applied to the methods. Also, the model performances will be compared and evaluated. This study uses and compares 7 methods and there is no comprehensive study in literature in terms of method variety.**

***Keywords*— feature importance, online retail, e-commerce, purchase, model performance**

I. Introduction

Pandemic hitting the world, physical shops closed in most of the countries. Even distant consumers to online shopping due to various reasons had to shop online. With the obligation to online shopping, profits have increased in the sector. In parallel to consumer interests increasing online, also the sector invested much more on the functionality and conveniences. They guaranteed to customers for quality products, totally free returns, right of choosing exact time of delivery and returns etc. Competition heated up with the new entrants to the sector.

In parallel to the increased competition in the sector, online retailers invested much more on their technology. Technology team head counts increased 3x-4x prior to the pandemic. Data has become very important in making decisions. Every movement of the customers started to be analyzed to get consumer insights, no design changes have applied without A/B tests.

Logging customer events simultaneously, store and being able to analyze them became very important. Company invested much more on traffic, funnel and flow analysis. They started measuring the performance of their platforms. Also, by measuring shopping behaviors they segmented their customers, presented special and personalized experiences.

Many customers visit online retailers using paid and unpaid channels. Using search engines or via advertisements they reach retailers' apps and web platforms. But it is not sufficient for a business that customers only land the platforms, they must make a purchase to get revenue for business survival and growth.

Marketing budgets are very high for the companies during the pandemic. Because, customer traffic was high and companies somehow get profit from customers. With the end of the pandemic customer traffic decreased and also many regulations came to life restricting marketing facilities. For this reason, getting revenue from the available customer traffic on the platform became much more important. Company had to present much more user friendly, effective platforms to the customer.

After customers reach the platform, they have to complete the conversion funnel for purchase. Conversion funnel steps for the purchase are landing, discovery, check out and buy.

- Landing: Customers land to online platforms from many points. These points can be search engines like google, bing etc, advertisements on many platforms, CRM mails and SMS, social media platforms etc, push notifications and friends' or influencers shared links. Apart from these, customers can reach the platform directly by opening the platform homepage on their devices.
- Discovery: In this phase, customers browse the products. After landing the platform customers browse the products on the homepage, navigate the category and brand pages or search the product names. Listing the products, they can add the products to the basket directly on the lists or open product detail pages and add the products to basket on product detail pages.
- Checkout: In this phase, customers view the product items on the basket and check the purchase details and confirm, then they choose payment and delivery address info.
- Payment: This is the last step of the conversion funnel. Customer confirms the payment and sees the purchase success page.

Customers can have any problem in any step of the conversion funnel and quit the journey towards the purchase. For these reasons, conversion funnel steps should be smooth, simple, viable and user friendly. So, customers can easily navigate the platform, browse products and complete the purchase processes. The success metric of this process is conversion rate. The conversion rate is the ratio of how much of the customers made a purchase. It is how many customers made purchases between how many customers landed on your platform. There is a high cost of attracting customers to the platform, for this reason it is crucial to convert customers. Conversion rate optimization is the area focusing on that area, it is interested in converting many more customers and generating much more revenue. Defining the factors of the purchase decision and getting actions on those factors will optimize the conversion rate.

What factors can be decisive in a customer purchase decision? These factors are generally effective even though their impact changes according to platform, customer type or the sector.

- Traffic source type: Is the customer coming to the platform from google search, price comparison website, push notifications campaigns, influencer links etc. The type of the interaction can be important for the purchase decision.
- Customer persona: Customer personas are defined according to shopping habits. Customers can be in "Moms - Women", "Digital Parents", "Gamers", "Fashonists", "Students".
- Customer segment: Customer segments are defined according to engagement of the customer with the platform. Customers can be in new, inactive, retained, reactivated, churn segments.
- Customer tag: Customer tags are defined according to the loyalty program that customer belongs to. Customer tags are new member, premium, efso, new member & premium, efso & new member & premium, efso & new member, efso & premium.
- Page Type: Type of the page that customers browse. Page types can be homepage, search results pages, product detail page, merchant and brand page.
- Page Component: Type of the component which customers interact with. Buybox, widgets and variants etc.
- Campaign Label and Label Type: There are labels on the product images that mention the offers. Offers can be about payment and delivery conveniences, benefits etc.
- Discount: Discount applied or not applied info can be decisive on purchase decisions.
- Product position on page: Whether the product is easily reachable on page or not.
- Notification permission status: Whether the customer permits the notifications or not. It means that the customer is open to campaign communications or not.
- Product review counts and scores: Customers trust the other customers' reviews about products.
- Recommendation: Is the product recommended to the customer or not according to personalized purchase habits.

In this paper, which factors are decisive on consumer purchase decisions will be studied on an e-commerce retail platform from Turkey, Hepsiburada. The data related to factors mentioned above will be gathered in a dataset. The factors which have impacts on purchases will be analysed.

The data gathering and pre-processing is the most challenging part of the study. Due to big data, there are performance issues regarding processing data. Optimizations need to be applied on SQL queries and pipelines should be applied for data feeding to the model.

Also, quality problems on the data is another big problem that needs to be solved.

In this paper, these steps will be applied.

- Data needed for the study will be gathered from Hepsiburada transactions stored in google's big query environment. Performance problems will be solved via SQL optimization and other methods.
- Data quality issues will be enhanced by fixing missing values and replacing unstandardized data with appropriate data for the model.
- Unbalanced data will be handled to get consistent results.
- Supervised ML models will be applied to data for feature selection and feature importance procedures.
- Quality and the significance of the results will be evaluated and results of different models will be compared.

II. LITERATURE REVIEW

Customer decisions on purchase can be related to many factors. Customers buying practices, how they interact with the platform and platform design features are very effective. There are many studies related to the topic in the literature.

Teye and Missah studied the usefulness relation with the purchase with other related factors.Their model consists of number of pages that a user visits in a session, duration of the page visits, bounce rate and exit rates. Apart from these usability factors, the model studies traffic type, visitor type, special day, weekend. They convert categorical data to the numerics with one hot coding. The target value in the model is purchase and non-purchase decision at end of the session. In their dataset non-purchase ratio is %85 in all transactions. The imbalance problem in data is solved with adaptive synthetic (ADASYN) sampling method. The oversampling is being applied to balance the data. Their dataset includes 18 features, the feature count increased to 57 by applying one hot coding method.

They use Sklearn's SelectKBest method to find the most impactful features on the dataset. Feature counts decreased to 15 at the end of the process. Also, outliers in the dataset are deleted from the dataset to increase accuracy of the model. The random forest classifier is used in the dataset. Teye and Missah have chosen a random forest classifier, because it minimizes risk of overfitting well, the performance is good in terms of training time on large datasets and handles with missing values. Also, hyper parameter tuning is studied to increase the accuracy.

Looking at the results, they achieve %95.12 accuracy, 0.83 True Positive Rate, 0.89 True negative rate, 0.79 F1 score. It is better than Sakar et al. [2] and Baati and Mohsil studies [3]. As the key result of the study, they observe that time spent on navigation and pages, special days are also important with the pricing and design components.

Sakar et al. [2] studied a real-time online shopping behavior analysis system. The system imitates the behavior of the real sales person in the physical store. The system understands customer intentions of purchasing The system includes two modules.The first module predicts the visitor's shopping intent, second module predicts Website abandonment likelihood. The first module calculates a score that measures purchase intention of the customer and offers the content to the customer if their buying intention exists. The second module is activated if there is greater value than the predetermined threshold. The second module decides whether the customer does abandon the website, then the first module offers the content to get the customer to continue browsing. They studied the first module using random forest, support vector machines and multilayer perceptron classifiers. They apply oversampling and feature selection methods to improve the processes. In their dataset there are categorical and numerical features. Numerical features include number of pages visited during the session, amount of time on visited pages, bounce rate and exit rates. Categorical features include operating system, region, traffic type, visitor type. They collect the data from 12,330 sessions using Google Analytics. %84.5 of the sessions end without purchase, the rest 1908 sessions end with the purchase behavior. The best performing classifier is multiple layer perceptron with the &87.24 accuracy. Decision tree method of random forest achieves %82.34 accuracy, support vector machine achieves %84.88. MLP classifiers have been chosen for the model in the first module. In the second model, the LSTM-RNN model was used. LSTM-RNN achieves %74.3 accuracy to understand when the customer leaves the website after a single action.

Similar study to Sakar et al. has been made by Baati and Mohsil in 2020. They use the same dataset. Their proposed system claims that they address the purchasing intention of the customer as soon as they interact with the website and offer the content only to those who intend to purchase. Their system consists of two parts. First parts offer marketing content if they detect there is purchase intention. After presenting the content, if the customer does not purchase, if the purchase intention continues, a more generous offer has been presented by the second system. They use Naive Bayes Classifier, C4.5 decision tree classifier and the random forest classifier. Naive Bayes Classifier handles input features independently with the target values, and is not interested in correlations between input features. The C4.5 decision tree classifier is based on information gain. The highest information gain is used to classify in the model. Random forest is constructed on different decision trees. Each decision tree bootstrap sample is set from the dataset. After solving imbalancing problem in the dataset with SMOTE method, the random forest method got the highest scores, accuracy is %86.78 and F1 score 0.60. Apart from Sakar et al. system, this system avoids the risk of customers leaving the website as soon as they interact with the website. Also, this study makes comparisons of different classifiers.

Topal [4] uses same dataset Teye, Missah [1] and Sakar [2] and apply different methods. For the feature selection, fisher score method has been applied. This method uses mean and standard deviation to calculate relationships for each class. The decision tree classifier for classification is used for the model. To handle imbalance data problems, the decision tree classifier is used with the K Fold Cross validation. At the end of the study, he finds that the most powerful feature in the data set is "Page Value". The study achieves %88.39 accuracy with 0.54 F score, %74,71 true positive rate, %89,55 true negative rate.

Conversion rate is the ratio of order count ration in the sessions. Fatta et al. [5] establish a hypothesis describing conversion rate effectors. First type of effectors are promotionals. Free shipping, free return service, discount policies and the sale season affects the conversion rate positively. The other effectors are related to quality. Page speed load is one of the quality factors. Other quality factors are luxury and mainstream products. Data is collected from six different SMEs ecommerce. These websites sell bags, shoes, accessories and apparels etc. Three of them sell luxury products, the other three sell mainstream products. Data includes daily visit counts, number of orders, average load speed of the pages, discounts, free shipping, free return and seasonality info. They made 1184 on a daily basis from the six websites for six months period. They use OLS (Ordinary Least Square) regression analysis to find the impact of independent variables. Also, they use qualitative comparative analysis (QCA) to find relationships within variables and the conversion rate. As a result of the study, they found that discounts and free shipping work together to boost conversion rate positively. Their interesting finding is that only free shipping has a negative impact on conversion rate. The customer feels that on the final prices they pay much. Their other finding is that discounts on sale season has a positive impact on conversion rate. The promotional factors have less effect on luxury products other than mainstream products. But the load speed all the time has a positive effect on conversion rate independent of any factor.

It is certain that there is a relationship between conversion rate and website design features. McDowel et al. made a study to test these hypotheses in 2016.

- Hypothesis 1: E-commerce website features that enhance purchase intention within the Visitor Greeting stage of the website are associated with conversion.
- Hypothesis 2: E-commerce website features that enhance purchase intention within the Catalog pages of the website are associated with conversion.
- Hypothesis 3: E-commerce website features that enhance purchase intention within the Shopping Cart page(s) of the website are associated with conversion.
- Hypothesis 4: E-commerce website features that enhance purchase intention within the Checkout page(s) of the website are associated with conversion.

Dependent variable on the study is conversion rate. Independent variables are recommended and featured products on greeting and catalog pages, shopping cart icon on navigation bar, shipping charge display in shopping cart, one-time credit card registration on checkout page, provision of links to site pages. They use Pearson correlation and p-values. Looking at the results of the study, early interaction with customers is important. Recommended and featured products, shopping cart icon in the greetings page have a positive impact on conversion rate. Provision links to other pages like investor relations and customer relations etc. have a negative impact on conversion rate. For the catalog pages, special offers and uses subject tabs have strong positive correlation with the conversion rate. Shopping cart and recommended products are not associated with the conversion rate. On shopping carts, instantaneous pricing, displaying shipping and other charges, offering related products have a positive impact on conversion rate. Return to main catalog and order tracking features have a negative impact on conversion rate. For the checkout page, the email address required impact negatively. Human contact information and instantaneous pricing has a positive impact. The study implies that there is a relationship between effective web design and the conversion rate. Flow-enhancing features have a commonly positive impact on conversion rate.

Cezar and Öğüt analyzed the role of customer reviews, recommendations and the rank order in search listings for conversion rate in hotel booking. They collected the data from Barcelona and Paris hotels via Booking.com. They use fractional regression models, quasi-maximum likelihood estimation and a regression model with beta distribution. As a result of the study, they found that high location ratings, high numbers of recommendations and high ranks in search listings have a positive impact on conversion rate. Also, low ranks in search listing affect conversion rate positively if there are high numbers of recommendations and high location ratings. Service and star rating do not have an impact on conversion rate. In addition to those, price has a negative impact on conversion rate.

The study focuses on the LSTM-RNN method to understand the purchase intention of the customer [8]. A web site server logs were used in the study. All actions of the customers categorized to types. These types are category, view product, home, ask question, order, contact, add cart, view cart, search, concerned, account, recommended. Concerned includes the logs of user check privacy policy, payment security, product shipping and return related pages. Ask questions includes the actions asking questions about the product or the service. Recommend includes the actions of recommending a product. %3.14 of the session ends with the purchase, %90.9 of the sessions are just for browsing so that users do not add any item to basket, the rest of the session's items added to cart but not purchased at the end of the session.

In the model, actions of the users in the session are handled in a sequence that represents user navigation. The LSTM method was used by forwarding time windows in the navigation for the session. With windows size of 20, they achieve %98.15 accuracy.

Another study using RNN method to predict purchase intent made by Sheil et al [9]. They use The RecSys 2015 Challenge (9.2 million user sessions) and Retail Rocket Kaggle (1.4 million user sessions) open datasets. Both datasets present the data sequentially with timestamps. A model is constituted from RNN layers after data embedding. Dataset is split into training and test in 90:10 ratio. The model is trained using Adam optimizer. They tested three types of recurrent cells: For the model comparison, they used AUC comparison. As a result of the study, the RNN model achieves %98.4 of current state of art performance.

Nazir et al [10] test whether artificial intelligence is influencing social media engagement and purchase intentions of the customers. They applied questionnaires to the customers who booked hotels or flights via booking.com in three different regions of Oman. The data collection period was 12 weeks. In the data collection process, they applied a convenience sampling method. 550 questionnaires were distributed to hotel customers with cover letters. 329 questionnaires were returned from customers, 21 were missing, so 308 questionnaires were in hand ready to use.

Five-point Likert scale was used in the questionnaire, the scale is ranging 1 (strongly disagree) to 5 (strongly agree). Questions measure the interest of the respondents about artificial intelligence technology, where engagement on social media teaches about brands, and whether digital promotions on social media are effective on purchase decisions. Also, there are questions to measure consumer habits and repurchase intentions. Besides, respondents' demographic information has been collected. These are age, income, gender, frequency of online purchases.

They used SmartPLS 3.2.9 (partial least square-based software) to analyze the results of the questionnaire. With the software, they tested the hypotheses they constructed for the study. According to results, AI technology is positively related with consumer engagement on social media. Also, there is a positive relationship between customer engagement on social media and customer satisfaction. Apart from that, customer satisfaction is positively related to re-purchase intention. As the result of the study, it is proven that usage of AI in social media campaigns is effective in converting users to customers.

Hu et al [11] applied feature extraction to understand shopping behavior. Features are divided into five categories.

- User characteristics: The count of various behaviors of the users before the prediction day.
- Product characteristics: the number of users visiting the product before the prediction day.
- Product category characteristics: Which category a product belongs to.
- User product characteristics: The count of various actions of users on a product before the prediction day.
- User - product category characteristics: The user preference on certain product categories

They use a deep forest model. This model is constructed from random forest structures in a manner of deep learning models. On each layer of the model, different types of random forests are used. This improves the fault tolerance of the model. At the end of the study, a 9.51 F1 score was achieved with 41 s. Training time. Also, support vector machine, random forest, Xgboost, deep neural network tried. Best F1 score achieved with deep forest compared to other methods. This study proves that the deep forest model is better in some cases from other models.

III. DATA PREPARATION

In the introduction section, we mentioned the factors that can affect customer purchase decisions in an online session. Customers add to cart action is used in data of this study. Add to cart means customers add products to the cart in the online platform. Add to cart is a certain interaction event of the customer with the platform. For this reason, add to cart action is used to form the data. In this dataset, if customers make an add to cart they are assumed as they are really interested with the session and the session features, then we evaluate the factors on the customer decision for purchase or not purchase.

To explain the rationale in the data, there are features

- related to the screen design characteristics: which pages and page components the customer has interacted with. In the study, add to cart action is assumed as interaction action. There are many components on the page, the customer scrolls the screens but we do not know which components the customer really interacted with. Add to cart action is a certain action that shows the customer action on the screen. The page type and page component that the product was added to the basket is included in the dataset. Product position on the page is another feature included in the dataset. Because the customer sees only two products on listing pages after opening the page directly, for others the customer has to scroll down to see other products.
- related to the customer characteristics: customer persona, tag and customer segment, communication preferences like notification permission status.
- related to how customers visit with the online platform: traffic source type.
- related to product characteristics: Is there a price discount on the product, product review counts and scores, is the product recommended by data science engine or not.

A. *Raw Data Collection*

The data has been gathered from the transactions from Hepsiburada, a leading e-commerce company in Turkey. First of all, add to cart data has been inquired with 1,348,826 the row count. IOS and android native app data is used in that study, web platform is not in the scope of the study. This data includes page type, location, campaing_label, product position on the page, if the price has discounted, if customer notification permission is on, review count and score on the product. Add to cart click data includes screen design product, device and platform characteristics.

Then all customers' persona, tag and segment info inquired from the system. Customer segment, tag and persona info has been merged to add to cart data. So, the add to cart click data with customer characteristics has been obtained. Customer segments are new, churn, inactive, retained, reactivated. Customer tags are new member, premium, efso, new member & premium, efso & new member & premium, efso & new member, efso & premium. Customer personas are gamer, student, no_persona, anne_kadın, fashionista, digital_ebeveyn.

Then all customers' traffic types info inquired from the system. Traffic type of the customer can be seo which means customer comes to online platform from searching in search engines like google, bing etc. Traffic type can be paid, which means customers can come to the online platform by clicking advertisements on various platforms. These platforms can be the search engines, social media platforms instagram, sms, e-mail or push notifications that the company sends to customers etc. For these customers, the company pays for the advertisement platforms. Another type of traffic is direct. This means that the customer directly opens the online platform, this is ios or android app in our case. Traffic types of sessions have been merged to add to cart data.

Finally, purchase info added to add to cart data using sessionId. If users make an add to cart and purchase in the same session, the is_order column in the dataset is set as true. If users make an add to cart and did not purchase in the same session, the is_order column in the dataset is false.

As a result, our raw dataset has been finalized. is_order column is the target column in the dataset and other columns include features that affect purchase action of customers in the session. In our raw dataset, we have 14 features and one target value (the customer made purchase or not). Our rows dataset includes 991,615 rows.

B. *Pre-processing the Data*

Checking the dataset, some feature values are null that they cannot be nulls. These features are page_type and location. 23 null values in these features in the dataset was dropped.

Then, data types of the features have been checked. Isnotificationon feature format was object type, then that is converted to boolean.

Some features in raw data are categorical. These values should be converted to numerical values to be able to establish a statistical model. These columns in our dataset are 'page_type', 'location', 'label_type', 'segment', 'tag', 'persona'. One hot coding method was used to convert these categorical values to numbers. One hot coding method is better than other methods in terms of performance. It encodes the values in the feature as a binary vector array in separate columns. After applying this method to our dataset, the column count in our dataset increased to 68.

Some features in the data are numerical. These features' data range is different. These values should be normalized for the model performance and the correct results. The data range should be in the same range. MinMaxScaler method has been used to normalize data. This method rescales variables into the range [0,1].

IV. Methodology, Experiments and Results Evaluation

A. *Methodology*

To define decisive factors on customer purchase decisions, feature importance methods will be analyzed, applied, compared in the scope of the study. Feature importance methods assign scores to the features according to how they are effective on defining target value. In our example, the target value is two class, the customer purchased or did not purchase.

We have defined probable features that have effect on purchase decisions with the domain knowledge. Feature importance methods will define the most important ones among these features with the power statistics and give important insights about our data.

There are 3 ways of defining feature importance generally.

- First way is calculating coefficients. Coefficients show the relationship between features and the target variable. They explain when a feature changes how much the target variable changes positively or negatively. Logistic regression and linear regression methods are using coefficients when predicting target values. While the linear regression method is predicting continuous target variables, logistic regression predicts categorical classes. In our study, our target variable purchase or not purchase, there are categorical binary classes, so we use logistic regression method in our study.
- Second is tree-based methods. Tree based methods are powerful on explainability. In this method, how much each feature reduces the impurity in a tree is defined. Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. In our study, we will use CART, Random Forest and XGBoost methods as tree-based methods. While CART uses one decision tree, Random Forest and XGBoost use an ensemble of many decision trees. Random forests combine many independently trained trees. XGBoost trains decision trees sequentially. Every tree is trained according to the learnings from previous trees, adjustments are made according to previous tree's errors.
- Third is the permutation-based methods. In permutation-based feature importance methods, some features values are shuffled in the model. If the model error increases much, the shuffled feature is assumed to be more important. This means, this feature impacts prediction much more. Permutation-based method cannot be used alone. Firstly, a model had to be trained with a classification method. In this study, we will try permutation method on Random Forest, XGBoost and Multilayer Perceptron classifiers. After a model has been trained with any of these classifiers, shuffling process is being applied to the model to find features the increases model errors much.

B. *Experiments*

After normalizing the data, data is split into train and test data. Logistic regression model was trained with the train data and then accuracy, f1, precision and recall score were found with the test data.

1) Logistic Regression: "Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables." [13]

In our dataset dependent variable is purchase decision of the customer in a session: purchase or no purchase, which is binary class. They are discrete value. They are not continuous value, for this reason we used logistic regression instead of linear regression. 68 features are independent variables that impacts customer purchase decision which is dependent variable. Logistic regression method is used when dependent variable is categorical. To define that the customer is likely to purchase or not, we need to put threshold between purchase and not purchase classes. As linear regression method is unbounded, it is not possible to define classes of customers. Logistic regression method predicts customers' purchase decision in a session using sigmoid function. This function uses all independent features and attain customer session to binary class: 1 (purchase) or 0 (no purchase).

Sigmoid function is a mathematical function maps linear regression output which is continuous value to between 0 and 1. If linear regression output tends to negative infinity maps the value to 0, which is no purchase in our example. If linear regression output tends to positive infinity maps the value to 1.

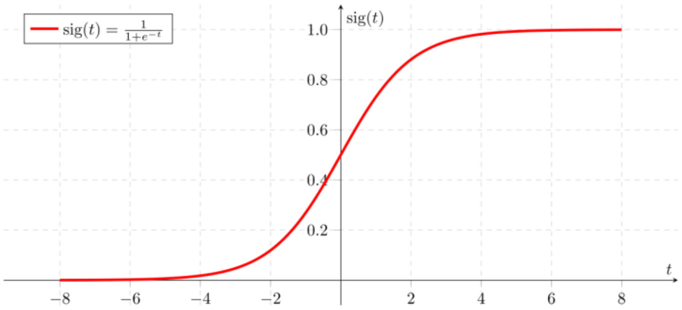

Fig. 1 Sigmoid Function

In the first try we used unbalanced data, accuracy of the model is 0.83. Since we used the unbalanced data, we checked the f1 score. In our case true positives and false negatives importance is the same, for this reason we did not control Precision and Recall, instead we checked f1 score. According to results, the purchase f1 score is low (0.34) , while no purchase f1 score is high (0.9). This means our results for the purchase class are not qualified. For this reason, we try the model with balanced data.

Undersampling method has been applied to data from imblearn library. After this procedure, accuracy of the model has dropped to 0.71. But, f1 scores of no purchase and purchase got close. No purchase f1 score ise 0.64, purchase f1 score is 0.76. Then oversample method has been applied but accuracy and f1 scores did not change.

MinMaxScaler has been applied to data. This method changed with StandardScaler method. But, accuracy and f1 score did not change.

Feature selection method has been applied to data. Count of features decreased to 10 with the PCA method. Then, logistic regression has been applied. But, the accuracy score did not improve. Accuracy score is 0.65. No purchase f1 score is 0.68, purchase f1 score is 0.62.

This method defined the most important 10 features and their importance values as in table below.

Feature_name	Importance
segment_Reactivated Customer	2.16
segment_Retained Customer	2.00
segment_New Customer	1.97
location_buy_ticket	1.69
page_type_add_basket	1.69
persona_Gamer	1.12
segment_Inactive Customer	1.09
persona_Student	0.93
segment_Churn Customer	0.86
tag_New Member & Premium	0.84

Table I. Logistic Regresssion Results

2) Decision Tree: "A decision tree is a hierarchical model used in decision support that depicts decisions and their potential outcomes, incorporating chance events, resource expenses, and utility. This algorithmic model utilizes conditional control statements and is non-parametric, supervised learning, useful for both classification and regression tasks. The tree structure is comprised of a root node, branches, internal nodes, and leaf nodes, forming a hierarchical, tree-like structure." [14]

For understanding decision tree method, some terminology must be introduced. Nodes are basically features in the dataset. Root nodes are starter node that splits the dataset. Every split creates sub-trees. Root nodes are split into several decision nodes. Then splitting goes on to find leaf nodes.

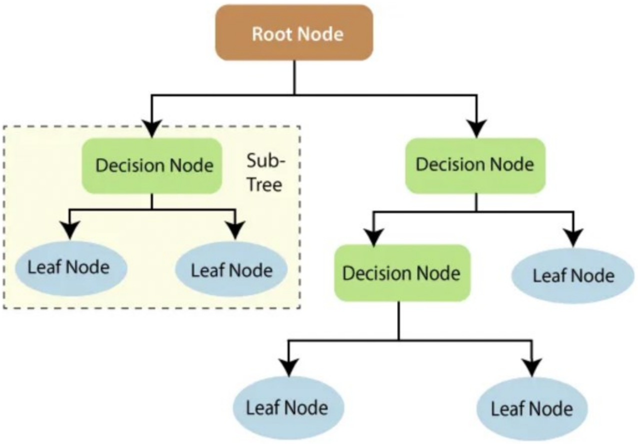

Fig. 2 Decision Tree

Firstly, algorithm defines the best feature in the dataset as root node using Attribute Selection Measure (ASM). Then split the root node into subsets. These subsets are decision nodes or leaf nodes. If the subset is decision node, split continues, if the subset is leaf node, split stops. By splitting, we create many decision trees.

Splitting is made according to Attribute Selection Measures (ASM). Attribute Selection Measures are Information Gain and Gini Indexes. Information gain measures the changes in entropy after splitting dataset. According to entropy change, find feature which gives much information about the class. Another Attribute Selection Measure is Gini Index. It looks to purity measure in dataset when splitting. Low Gini index is preferred in the process. When high information gain and purity has been gained in datasets, the splitting process is stopped. But, it should be noted that much higher information gain or impurity creates overfitting.

With the unbalanced data, accuracy of the model is 0.82. f1 score of no purchase class is very higher than purchase class. No purchase f1 score is 0.90, purchase f1 score is 0.37. This means our results for the purchase class are not qualified. For this reason, we try the model with balanced data.

Oversampling method has been applied to data. f1 scores got close to each other after balancing data. No purchase f1 score is 0.69, purchase f1 score is 0.78. But, accuracy of the model decreased to 0.74.

MinMaxScaler has been applied to data. But, accuracy and f1 score did not change.

To improve model accuracy, decision tree parameters have been tuned. "Gini" and "entropy" functions have been tried, but accuracy did not improve. Max_depth have been set to 10, max_features have been set to 0.8. Splitter parameter has been selected as "best" and "random". These tunings did not help to improve accuracy.

This method defined the most important 10 features and their importance values as in table below.

Feature_name	Importance
location_buy_ticket	0.27
segment_Inactive Customer	0.19
segment_Churn Customer	0.10
position	0.10
tag_No-Tag	0.06
segment_Reactivated Customer	0.05
tag_Premium	0.04
persona_No-Persona	0.03
segment_Retained Customer	0.03
customerreviewcount	0.02

Table II. Decision Tree Results

3) Random Forest: "Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result." [15]

Random Forest is ensemble of decision trees. It uses bagging method. Bagging gets the random subset from the dataset and train every subset independently and creates many models. From each training, separate outputs have been obtained. Every output selects a class after training. All outputs have been combined with majority voting method. The class selected by the majority of models defined as final output of the process.

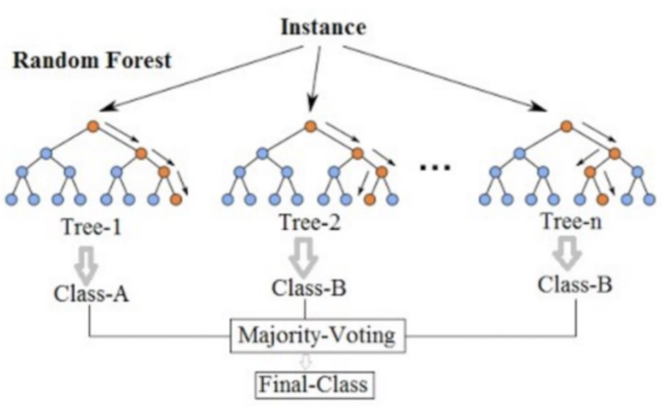

Fig. 3 Random Forest

- With the unbalanced data, accuracy of the model is 0.83. f1 score of no purchase class is very higher than purchase class. This means our results for the purchase class are not qualified. No purchase f1 score ise 0.9, purchase f1 score is 0.35. For this reason, we try the model with balanced data
- Oversampling method has been applied to data. f1 scores got close to each other after balancing data. No purchase f1 score is 0.67, purchase f1 score is 0.77. But, accuracy of the model decreased to 0.73.
- Random forest parameters have been tuned to optimize the learning process of the model. Tuning methods are trying hyperparameters in the model learning process and finding the ones which give best results. Mainly there are two methods for hyperparameter tuning: Grid Search and Randomized Search. Grid search tries every combination of hyperparameters. Randomized search method samples the hyperparameters and try in the model [12]. We used the randomized search method.

The best parameters have been find from the model: n_estimators = 115, min_samples_split = 2, min_samples_leaf = 1, max_features= 'log2', max_depth = 19, criterion= 'gini'

After applying best parameters model accuracy resulted with 0.73, did not change. No purchase f1 score is 0.67, purchase f1 score is 0.78.

This method defined the most important 10 features and their importance values as in table below.

Feature_name	Importance
segment_Inactive Customer	0.13
segment_Retained Customer	0.12
location_buy_ticket	0.11
segment_Reactivated Customer	0.10
page_type_add_basket	0.09
position	0.09
tag_Premium	0.07
segment_Churn Customer	0.06
tag_No-Tag	0.02
segment_New Customer	0.02

Table III. Logistic Regresssion Results

4) XGBoost: "Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. XGBoost is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems."

XGBoost is similar to Random Forest algorithm. While random forest trains each model at the same time, XGBoost trains sequentially. On each train, predecessors's error is being corrected.

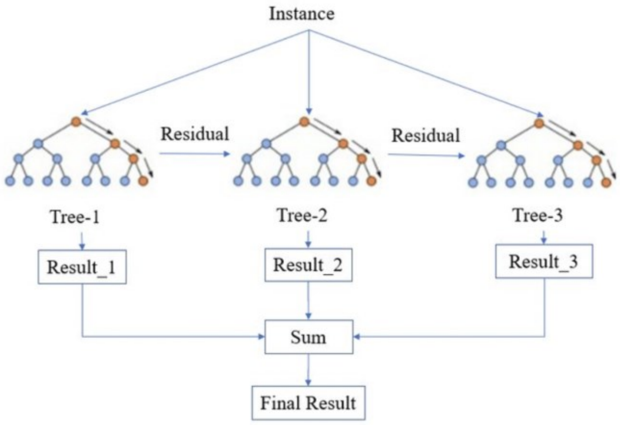

Fig. 4 XGBoost

With the unbalanced data, accuracy of the model is 0.83. f1 score of no purchase class is very higher than purchase class. This means our results for the purchase class are not qualified. No purchase f1 score is 0.9, purchase f1 score is 0.35. For this reason, we try the model with balanced data.

Oversampling method has been applied to data. f1 scores got close to each other after balancing data. No purchase f1 score is 0.66, purchase f1 score is 0.76. But, accuracy of the model decreased to 0.72.

To improve model accuracy, xgboost parameters have been tuned. Randomized search method has been used to find best parameters.

The best parameters have been find from the model: subsample= 0.7, n_estimators= 250, learning_rate= 0.1 ,max_depth=15, colsample_bytree= 0.8999999999999999, colsample_bylevel= 0.8999999999999999

After applying best parameters model accuracy has slightly increased to 0.74. Also, f1 scores have slightly improved. No purchase f1 score is 0.68, purchase f1 score is 0.77.

This method defined the most important 10 features and their importance values as in table below.

Feature_name	Importance
page_type_add_basket	0.531
location_buy_ticket	0.274
segment_Inactive Customer	0.066
segment_Churn Customer	0.036
segment_New Customer	0.015
segment_Reactivated Customer	0.009
segment_Retained Customer	0.007
persona_No-Persona	0.006
tag_No-Tag	0.003
tag_Premium	0.003

Table IV. XGBoost Results

5) Permutation with Random Forest: "The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature." [17]

We had trained a model using Random Forest, described in section 4.2.3. We will apply permutation method to his model. This method works based on predictions of the model. This method shuffles the values in selected features, then controls model error increase. If the model error increase is bigger on the selected feature, that feature is assumed as important.

This method defined the most important 10 features and their importance values as in table below.

Feature_name	Importance
segment_Retained Customer	0.0557 ± 0.0006
segment_Reactivated Customer	0.0514 ± 0.0005
tag_Premium	0.0321 ± 0.0009
position	0.0286 ± 0.0005
is_direct	0.0178 ± 0.0006
segment_Inactive Customer	0.0165 ± 0.0006
tag_No-Tag	0.0150 ± 0.0003
page_type_search	0.0103 ± 0.0003
segment_New Customer	0.0090 ± 0.0002
is_paid	0.0089 ± 0.0002

Table V. Permutation with Random Forest Results

6) Permutation with XGBoost: We had trained a model using Random Forest, described in section 4.2.4. Then, we applied permutation method.

This method defined the most important 10 features and their importance values as in table below.

Feature_name	Importance
segment_Retained Customer	0.0598 ± 0.0009
segment_Reactivated Customer	0.0390 ± 0.0004
position	0.0271 ± 0.0006
tag_Premium	0.0213 ± 0.0006
segment_Inactive Customer	0.0185 ± 0.0004
is_direct	0.0120 ± 0.0005
segment_New Customer	0.0104 ± 0.0001
page_type_add_basket	0.0095 ± 0.0004
segment_Churn Customer	0.0079 ± 0.0003
tag_No-Tag	0.0073 ± 0.0004

Table VI. Permutation with XGBoost Forest Results

7) Permutation with Multilayer Perceptron: We did not create multilayer perceptron model. For this reason, a multilayer perceptron model had to be performed before applying permutation method.

"The perceptron is very useful for classifying data sets that are linearly separable. The MultiLayer Perceptron (MLPs) breaks this restriction and classifies datasets which are not linearly separable. They do this by using a more robust and complex architecture to learn regression and classification models for difficult datasets." [18]

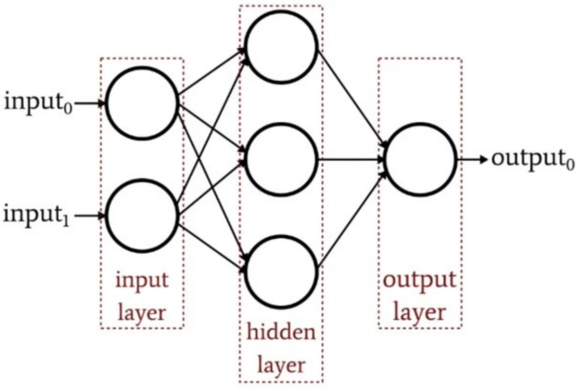

Fig. 5 Multilayer Perceptron

With the unbalanced data, accuracy of the model is 0.83. f1 score of no purchase class is very higher than purchase class. This means our results for the purchase class are not qualified. No purchase f1 score is 0.9, purchase f1 score is 0.35. For this reason, we try the model with balanced data.

Oversampling method has been applied to data. f1 scores got close to each other after balancing data. No purchase f1 score is 0.67, purchase f1 score is 0.75. But, accuracy of the model decreased to 0.72.

To improve model accuracy, multilayer perceptron parameters have been tuned. Grid Search method has been used to find best parameters.

The best parameters have been find from the model:

{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (50, 100, 50), 'learning_rate': 'adaptive', 'solver': 'adam'}

After applying best parameters model accuracy and f1 scores did not improve and change.

This method defined the most important 10 features and their importance values as in table below.

Feature Name	Importance
segment_Retained Customer	0.0415 ± 0.0006
segment_Reactivated Customer	0.0387 ± 0.0003
tag_Premium	0.0210 ± 0.0005
segment_Inactive Customer	0.0166 ± 0.0005
is_direct	0.0098 ± 0.0002
segment_Churn Customer	0.0081 ± 0.0004
persona_No-Persona	0.0079 ± 0.0002
position	0.0073 ± 0.0002
is_paid	0.0065 ± 0.0003
tag_No-Tag	0.0049 ± 0.0004

Table VII. Permutation with Multilayer Perceptron Results

C. *Results Evaluation*

7 experiments have been realized for defining most impactful features on purchase. In this section, impactful features according to results and model performances will be interpreted and compared.

1) Logistic Regression: Logistic regression method defined that all features have impact on purchase behaviour.

The most impactful 3 features are related to segment information of the customer. Reactivated, retained and new customers are most likely to purchase. Looking at the segment descriptions below, the customers in those segments have orders in last 30 days. We can say that the customers who made order recently are most likely to purchase.

The segment descriptions are these in the company:

- New Member: user that do not have order yet.
- New Customer: the user in a 30 days period of after first order.
- Retained: the user that have order in last 31-60 days, also have order in last 30 days.
- Reactivated: the user that have not order last 31-60 days, but have order in last 30 days.
- Churn: the user that have order in last 31-60 days, but do not have order in last 30 days.
- Inactive: the user that do not have order in last 31-60 days, also do not have order in last 30 days.

"location_buy_ticket", "page_type_add_basket" features belong "Scratch and Win" game of the Hepsiburada app. Customers buy tickets with small amounts for participating draws. According to logistic regression results, the customers who buy "Scratch and Win" tickets is much more likely to purchase.

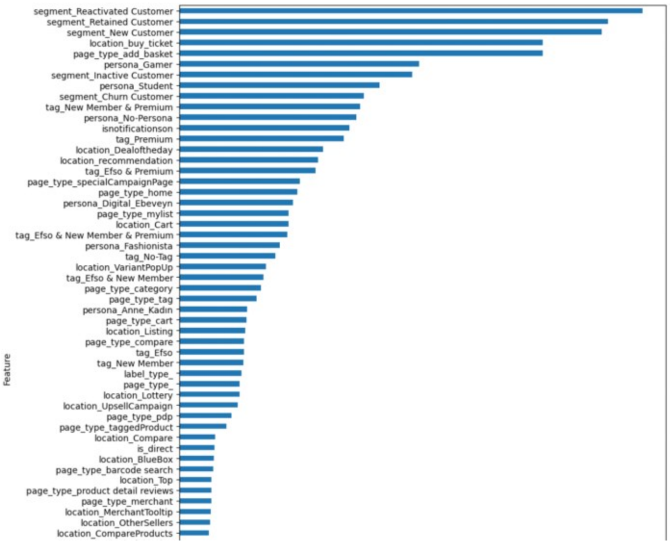

Fig. 6 Logistic Regression Results

2) Decision Tree: Unlike logistic regression, decision tree method does not find that all features have impact on purchase decision. Especially "location_buy_ticket",

"segment_inactive_customer", "segment_churn_customer" and "position" features have impact on purchase decision.

"location_buy_ticket" has been defined as impactful on purchase decision by logistic regression method too. This feature diverges from other features in decision tree method. This means, the customers who buy "scratch and win" tickets are much likely to purchase.

Unlike logistic regression, decision tree method found different segments related with purchase. Decision tree method found customers who did not purchase in last 30 days are more likely to purchase. Logistic regression method stated opposite of this finding.

Another important finding of this method is the position is important on purchase decision compared to other features. Logistic regression method did not define position related to purchase.

While Retained customer segment is the second important feature on logistic regression method, it is ninth important feature.

According to business domain knowledge, customer review counts and scores are important factors on purchase decision. Logistic regression method did not find these features as related with the purchase decision. Decision tree method defined as tenth important feature.

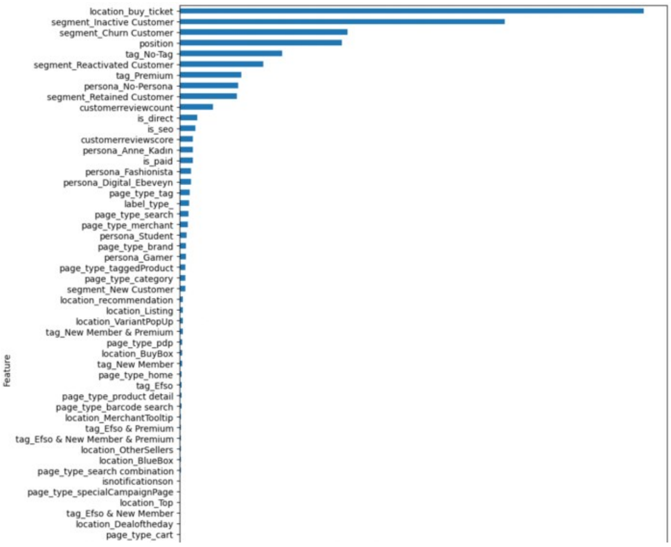

Fig. 7 Decision Tree Results

3) Random Forest: Random forest method found these feature is mostly related to purchase decision: "position", "is_price_discount","isnotificationson","customerreviewcount" , "customerreviewsscore", "is_direct", "is_paid", "is_seo".

Random Forest method findings are parallel to business domain knowledge. According to business domain knowledge, position of the product is very important, if the product position is small, it can be seen much more from customers and can be bought easily. As it is mentioned above, customer reviews are very important on purchase decision. When review counts and scores are higher, customers are most likely to purchase products. Another important finding that is parallel to business domain knowledge is price discount. In e-commerce sector, it is known that discounts are very effective on purchase decisions.

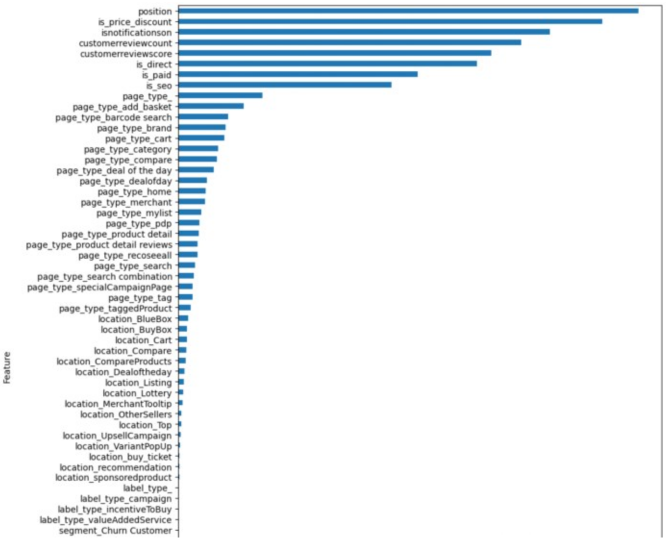

Fig. 8 Random Forest Results

4) XGBoost: XGBoost method finding are very parallel to Random Forest. They found the same features with same sequence as important. But the XGBoost method is much more clear on its outputs. Random forest found another features as related to order in small important ratios. XGBoost method attained most feature importances as zero.

Random forest method is cluster of decision trees. Every tree train sample of the dataset. Prediction of the majority of decision trees becomes final decision. Decision trees is trains at the same time. Unlike random forest method, XGBoost trains decision tree model sequential. Every model learns according to predecessor's mistakes and tries to correct. According to our findings, XGBoost presents much more clear results than random forests.

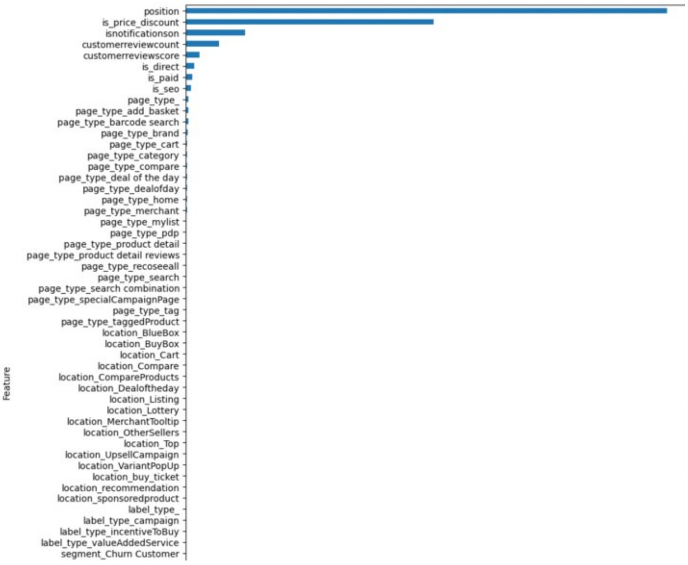

Fig. 9 XGBoost Results

5) Permutation: Permutation method has been tried with 3 different models: Random Forest, XGBoost and Multilayer Perceptrons. With all these models, permutation generated similar results.

Retained and reactivated customers have been found as the most important features parallel to logistic regression method. According to business knowledge, premium customers are much more likely to purchase, only permutaion method has defined this feature among important ones. Also, similar to random forest and XGBoost methods position feature has been found as important by permutation method.

6) Overall Evaluation: Random forest and XGBoost generated resulted similar results. But other methods generated different results.

Evaluating to results of the methods with business knowledge, random forest and XGBoost method generated results are most parallel to business knowledge.

Permutation method model has not accuracy and f1 scores, because it is based on random forest model. Tree based methods model performance scores are similar to each other and successful than logistic regression. Among tree-based methods XGBoost is the successful one in terms of accuracy, decision tree is successful one in terms of f1 scores.

Methods	Accuracy	No Purchase F1 score	Purchase F1 score
Logistic Regression	0.65	0.68	0.62
Decision Tree	0.74	0.69	0.78
Random Forest	0.73	0.67	0.78
XGBoost	0.77	0.68	0.77
Multilayer Perceptron	0.72	0.67	0.75

Table VIII. Model Performances Comparison

ACKNOWLEDGMENT

I would like to thank the Hepsiburada R&D Center for their support/contribution in the emergence of this study.

REFERENCES

[1] M. Teye, Y. M.Missah, Investigating Factors that Affect Purchase Intention of Visitors of E-commerce Websites Using a High Scoring Random Forest Algorithm, International Journal of Engineering and Technical Research (2020). https://www.researchgate.net/publication/351840990_Investigating_Factors_that_Affect_Purchase_Intention_of_Visitors_of_E-commerce_Websites_Using_a_High_Scoring_Random_Forest_Algorithm

[2] C.O. Sakar, S.O. Polat, M. Katircioglu, Y. Kastro, Real- time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks, Neural Comput. Appl. 31 (2019) 6893–6908. https://link.springer.com/article/10.1007/s00521-018-3523-0

[3] K. Baati, M. Mohsil, Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest, Artif. Intell. Appl. Innov. AIAI 2020. IFIP Adv. Inf. Commun. Technol. 583 (2020) 43–51. https://link.springer.com/book/10.1007/978-3-030-49161-1.

[4] I.Topal, Estimation of Online Purchasing Intention Using Decision Tree, journal of management and geonomics research (2019).https://dergipark.org.tr/en/pub/yead/issue/50655/542249.

[5] D. Fatta, D. Patton, G. Viglia, The determinants of conversion rates in SME e-commerce websites,Journal of Consumer and Retailing Services (2018). https://www.sciencedirect.com/science/article/pii/S0969698917306525?via%3Dihub.

[6] W. McDowell, R. Wilson, C. Jr, An examination of retail website design and conversion rate,Journal of Business Research (2016). Https://www.sciencedirect.com/science/article/pii/S014829631630203X

[7] A. Cezar, H. Öğüt,Analyzing conversion rates in online hotel booking, International Journal of Contemporary Hospitality Management (2016). https://www.emerald.com/insight/content/doi/10.1108/IJCHM-05-2014-0249/full/html.

[8] K. Diamantaras, M. Salampasis, A. Katsalis, K. Christantonis, Predicting Shopping Intent of e-Commerce Users using LSTM Recurrent Neural Networks, 10th International Conference on Data Science, Technology and Applications (2021). https://pdfs.semanticscholar.org/99ea/7a6a1c7aa63d752a8f33efa0128b7c36b330.pdf

[9] H. Sheil, O. Rana, R.Reilly, Predicting purchasing intent: Automatic Feature Learning using Recurrent Neural Networks, SIGIR eCom (2018).https://arxiv.org/abs/1807.08207.

[10] S. Nazir, S. Khadim, M. Asadullah, N. Syed, Exploring the influence of artificial intelligence technology on consumer repurchase intention: The mediation and moderation approach, Technology in Society (2023). https://www.sciencedirect.com/science/article/pii/S0160791X22003311

[11] X. Hu, Y. Yang, L. Chen, S. Zhu, Research on a Prediction Model of Online Shopping Behavior Based on Deep Forest Algorithm, International Conference on Artificial Intelligence and Big Data (2023). https://ieeexplore.ieee.org/abstract/document/9137436

[12] Arindam Banerjee (2022, November 2), Hyperparameter Tuning Using Randomized Search, https://www.analyticsvidhya.com/blog/2022/11/hyperparameter-tuning-using-randomized-search/

[13] Anshul Saini (2021, August 03), Conceptual Understanding of Logistic Regression for Data Science Beginners, https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/

[14] Anshul Saini (2021, August 29), Decision Tree Algorithm – A Complete Guide, https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

[15] Niklas Donges (2023, March 14), Random Forest: A Complete Guide for Machine Learning, https://builtin.com/data-science/random-forest-algorithm

[16] XGBoost, https://www.geeksforgeeks.org/xgboost/

[17] scikit-learn developers, Permutation feature importance, https://scikit-learn.org/stable/modules/permutation_importance.html#:~:text=The%20permutation%20feature%20importance%20is,model%20depends%20on%20the%20feature.

[18] DeepAI, Multilayer Perceptron, https://deepai.org/machine-learning-glossary-and-terms/multilayer-perceptro