Artificial Intelligence Ethics Manifesto

TÜRKİYE İŞ BANKASI

Introduction

As İşbank, artificial intelligence(AI) technologies are the centerpiece of many of our operations. We are aware that these technologies should build trust as well as material advantage towards us and our customers. Building trust requires good intentions and consistency. Therefore, we are proud to share our principles that guide us in this path together with examples of concrete actions.

With what we share, our goal is to raise awareness on AI ethics, specifically within our industry and country. We hope that this transparent demonstration of our examples and standards will resonate with all developers who work on AI, creating a positive impact. Wishing a future that is safe, fair and livable for the society altogether...

Our Ethical Principles

To properly manage our AI technologies and to be able to achieve the desired impact, our ethical principles that we have been committed to follow are listed below.

- Sustainability and Beneficence
- Transparency and Explainability
- Accountability
- Fairness
- Robustness
- Privacy

Sustainability and Beneficence

In our AI applications, we hold environmental respect as a priority as it provides beneficence in economy, education and culture in addition to sustainability.

We organize endeavors that have a positive impact socially, beyond us and our customers. Our sustainability initiative aims to protect our planet by minimizing the harm done to the environment.

With our collaborations, we build bridges between the industry and academia. To develop the AI culture in Türkiye, we have established Koç University & İşbank Artificial Intelligence Center (KUIS AI). With KUIS and the rest of our collaborations, we provide a place for the real life applications of our professional experience and academic knowledge to enrich each other. By meeting with students in seminars, we support the proliferation of these skills. During the seminars, we pass on our knowledge by answering questions on technical expertise and career guidance.

With our AI competitions, we create the opportunity for young people who want to advance their skills on the field. In our prized Machine Learning Challenge competition, we provided a chance for the attendees to experience data science applications with real data while keeping to our privacy measures. In the end of the competition, we shared the most successful applications with all attendees and gave our feedback. Shared solutions and the data we have provided paved the way for the enhancement of skills on the AI field among the next generation.

We manage our high computation load in our own data centers which we also use for our AI applications. We use dynamic resource management to maintain high efficiency. To this end, our cloud servers operate on our Tier 4 certificated data center which allows us to use processor allocation limits and resource virtualization to achieve efficient scaling. With this approach, we save on energy and minimize our carbon footprint.

Transparency and Explainability

We fully understand the requirements and explain the results to related parties on time and in an understandable, complete, and reproducible manner.

When black-box models are used, the algorithm generating the results is not fully explainable, which leads to less controllable, explainable, and auditable models. Figure 1 shows the impact of the complexity of black-box and transparent models on explainability [1]. Explainable AI applies special techniques and methodologies to monitor and explain every decision taken throughout the machine learning process.

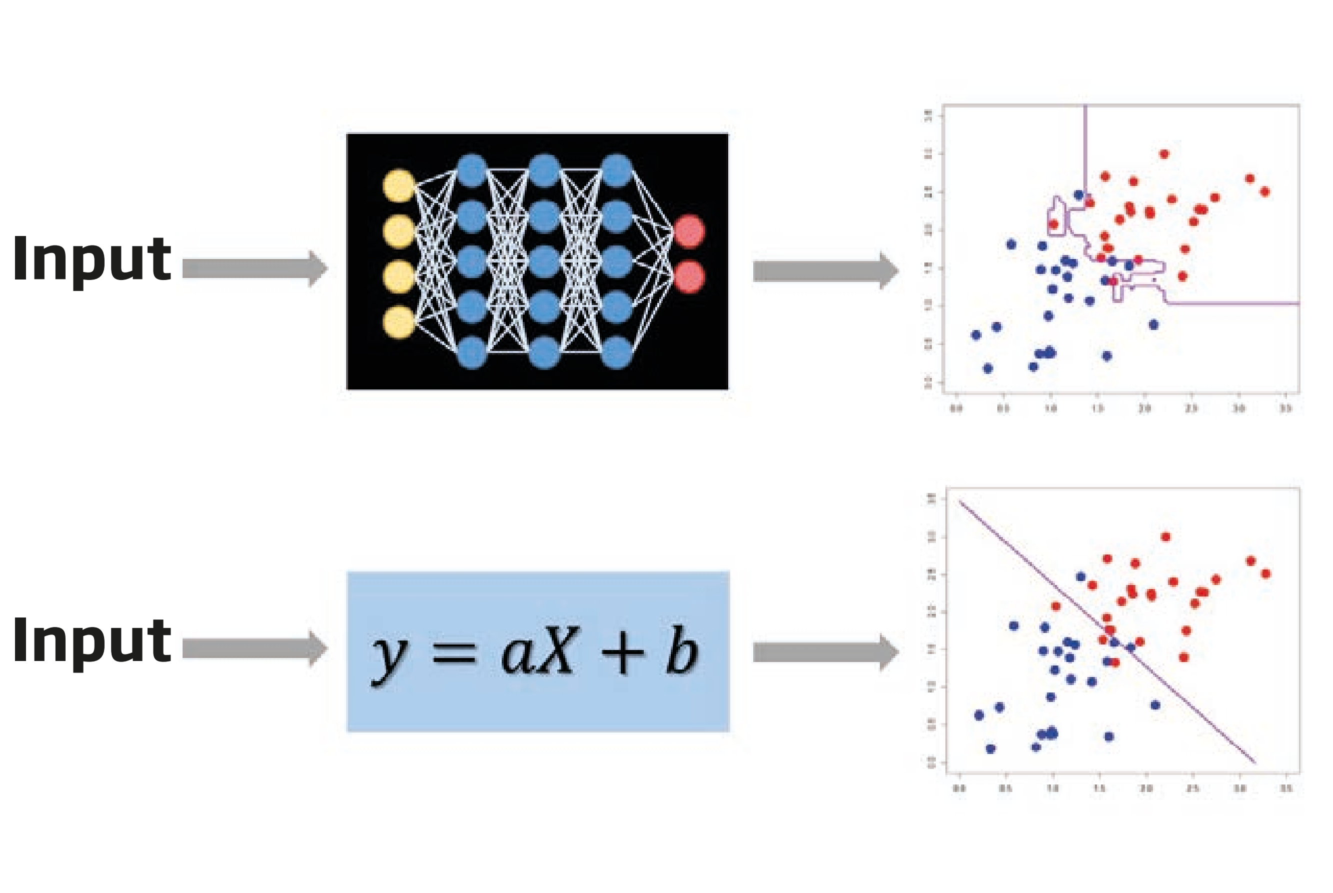


Figure 1: Black-box Model and Transparent Model [1]

Explainable AI models enable users to understand the output and trust the process of generating this result. In addition, the ability to produce explainable results enables the detection of potential bias.

Transparent models enable deterministic and interpretable results. However, black-box models such as artificial neural networks may be necessary depending on the problem. To clarify these black-box models, game-theoretic methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) calculate the individual and collective contribution of each variable.

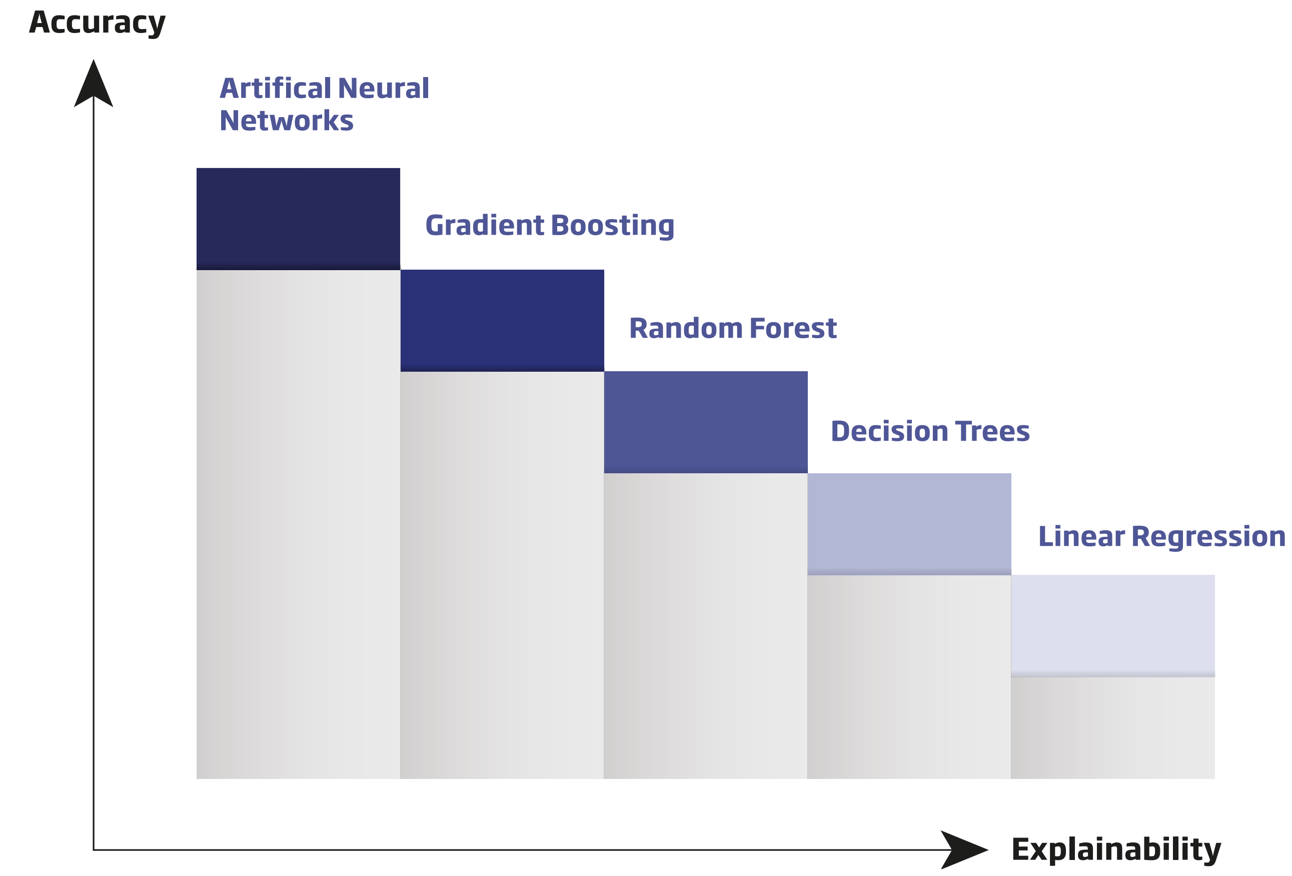


Figure 2: Accuracy – Explainability [3]

Explainability allows developers to be sure that the system functions as designed. Also, it facilitates questioning and changing of outcomes that affect users who are utilizing the models as a decision support system [2].

AI models being explainable is also important in accounting for and monitoring the changes that might occur in time. Continuous review and explainability of models in the production environment allow detection and correction of any issue in the model performance.

As Figure 2 shows, if our models achieve the same level of performance, we prefer simplicity to increase explainability [3]. We compare different models, and if there is no significant performance improvement, we do not switch to more complex models. Furthermore, we care not to use irrelevant or correlating features as input to lower complexity. We follow the recursive feature elimination approach while removing these features.

We disclose "Feature Importance Ranking" and share it with all stakeholders to observe the importance of features on the model output. Also, we adopt the SHAP approach in global (Figure 3) and local (Figure 4) explainability.

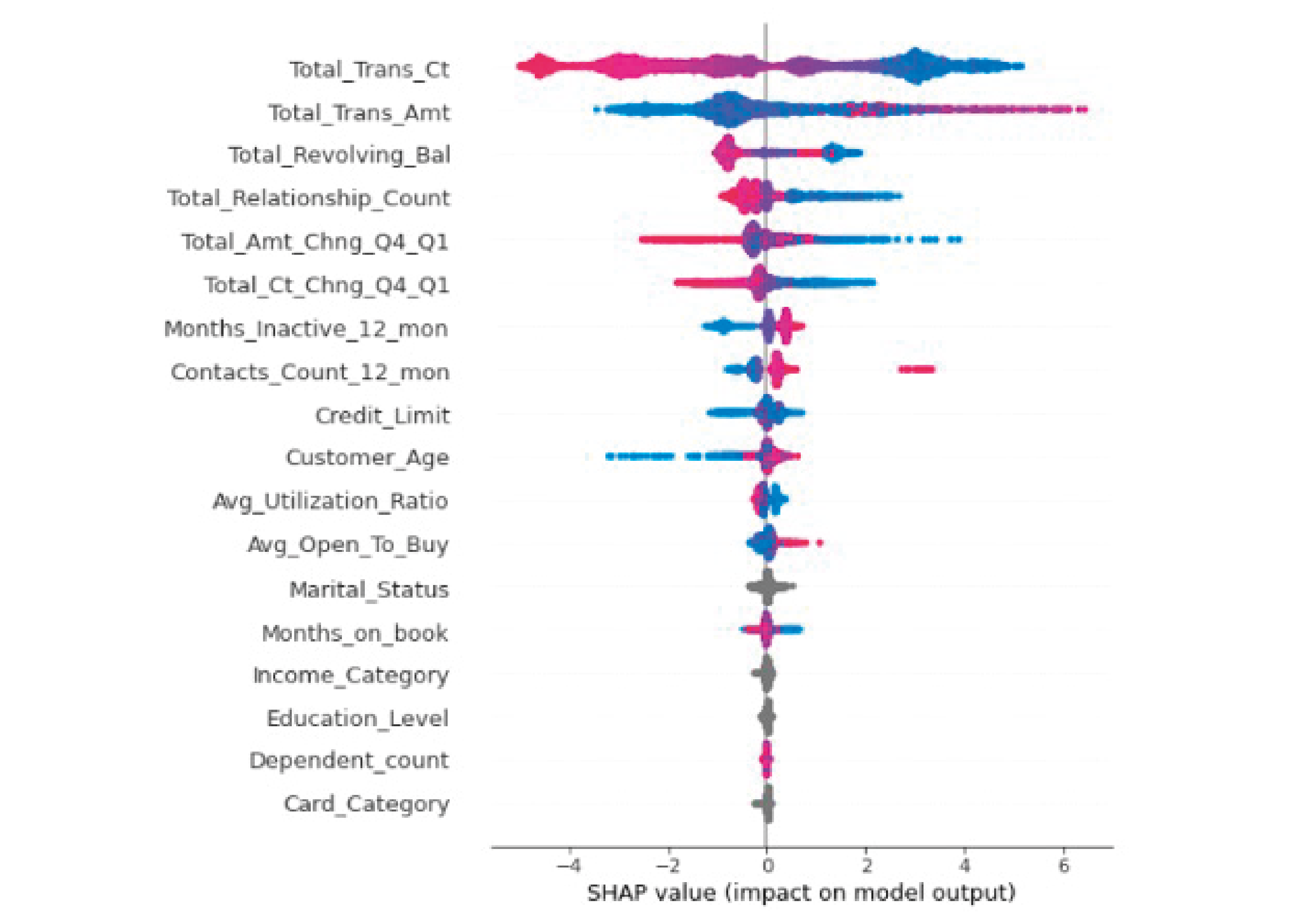


Figure 3: SHAP Global Explainability

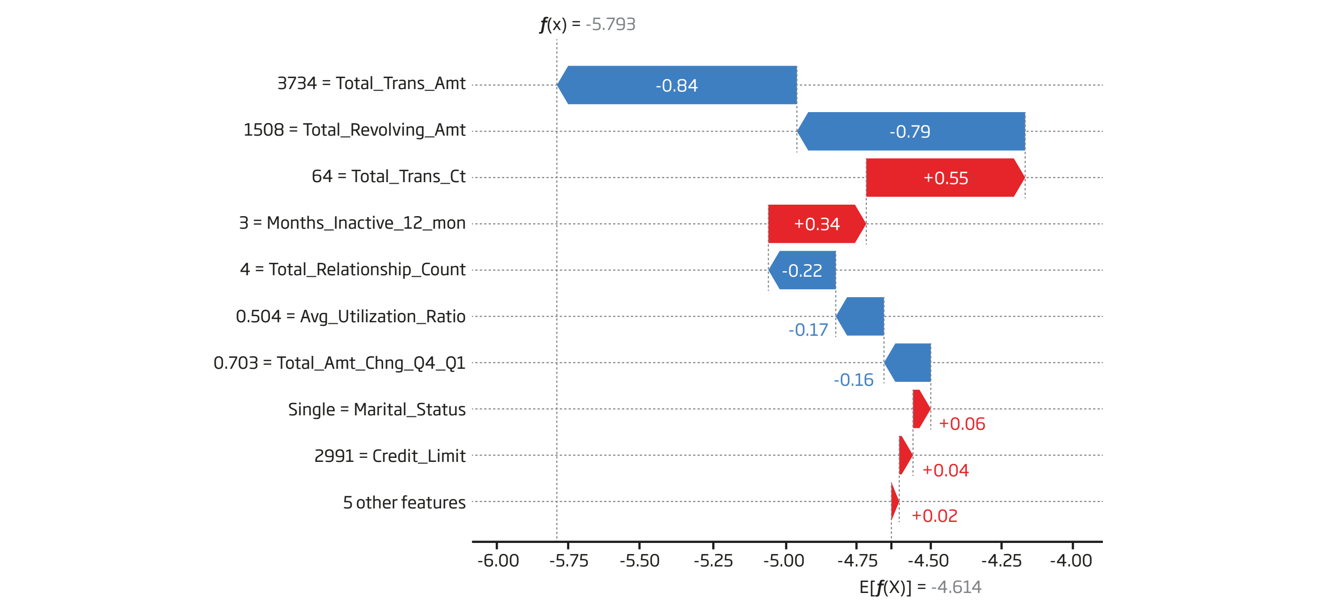
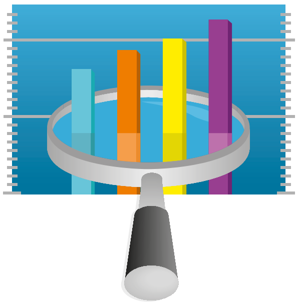


Figure 4: SHAP Local Explainability

Before deploying our models in the production environment, we compare model outputs by performing impact analysis.

Accountability

We execute and archive our AI applications in accordance with legislation and in line with their purpose. We are accountable for implementing AI applications based on our duty and role within the defined legislation and banking principles. We administer all of our applications through the İşbank Mozaik platform* we developed.



In general, accountability means that the individuals who develop and design the AI systems are responsible for the design, decisions, and outputs of those systems [4]. It is also the responsibility of these individuals to establish the necessary systems, standards and practices for the proper tracking and management of AI outputs. Clear identification of technical and business ownership for each AI application is critical to address the potential risks and issues which might occur in monitoring, maintenance and governance. To design accountable AI systems, the following criteria should be considered when evaluating AI applications [5]:

Guidance and Awareness: Proper principles, standards, and policies for AI applications should be established.

Evaluation and Monitoring: AI applications should be evaluated based on the company's predefined risks and values before being deployed into the production environment and further should be continuously monitored by responsible parties.

Auditing: Monitoring and alarm mechanisms should be put in place to ensure that AI applications continue to produce accurate, valuable, and sustainable outputs.

Standards mentioned above have been established for the accountability pillar in our banking AI ecosystem, and models developed and put into production are produced in compliance with these standards and managed from our AI platform called Mozaik. Mozaik enables the tracking of models from the early stages of experimentation to deployment in production environments. Authorization mechanisms have also been established on the platform to ensure that only authorized personnel can intervene on a model.

The technical and the business owners of each AI model together with its application have been assigned. Thus only the model owners and the authorized individuals can perform any changes to the model. The access to data and the model deployment can only be done by authorized individuals. Further, all actions taken by these individuals are recorded and can be audited.

All the details regarding model experimentation are stored historically on the AI platform. Similarly, all models trained in the production environment are recorded on the AI platform. Further, the deployment of these models is subject to the approval process conducted by the model owner.

All AI applications are reviewed by the Artificial Intelligence Committee at various stages of the AI lifecycle including early development to deployment into the production environment. This review is done using a comprehensive checklist which includes standards regarding AI architecture, data engineering and data science. Applications that have critical findings during the review are not put into production until those findings are being addressed.

Agile methods are used in the development process of AI applications, and regular reviews are being conducted with the participation of the corresponding business unit. This way, the process is carried out with the knowledge of the business unit and their feedback is received.

AI models are subject to verification and audit according to the bank's risk management standards. Findings are reported, audited and addressed.

*İşBank Mozaik platform is developed by our Bank to manage and standardize our AI-based systems.

Fairness

Our AI-based decisions are independent from attributions such as race, nationality, belief, social status and gender as we reject any type of biased approach.

Fairness ensures the protection of individuals and groups from discrimination or unequal treatment that may arise from biased decisions and behaviors based on their personal characteristics and the social classes they are in. In technical terms, fairness prevents bias by determining the quantities such as allocation, representation and error rates for individuals and groups in a fair or equal way [6].

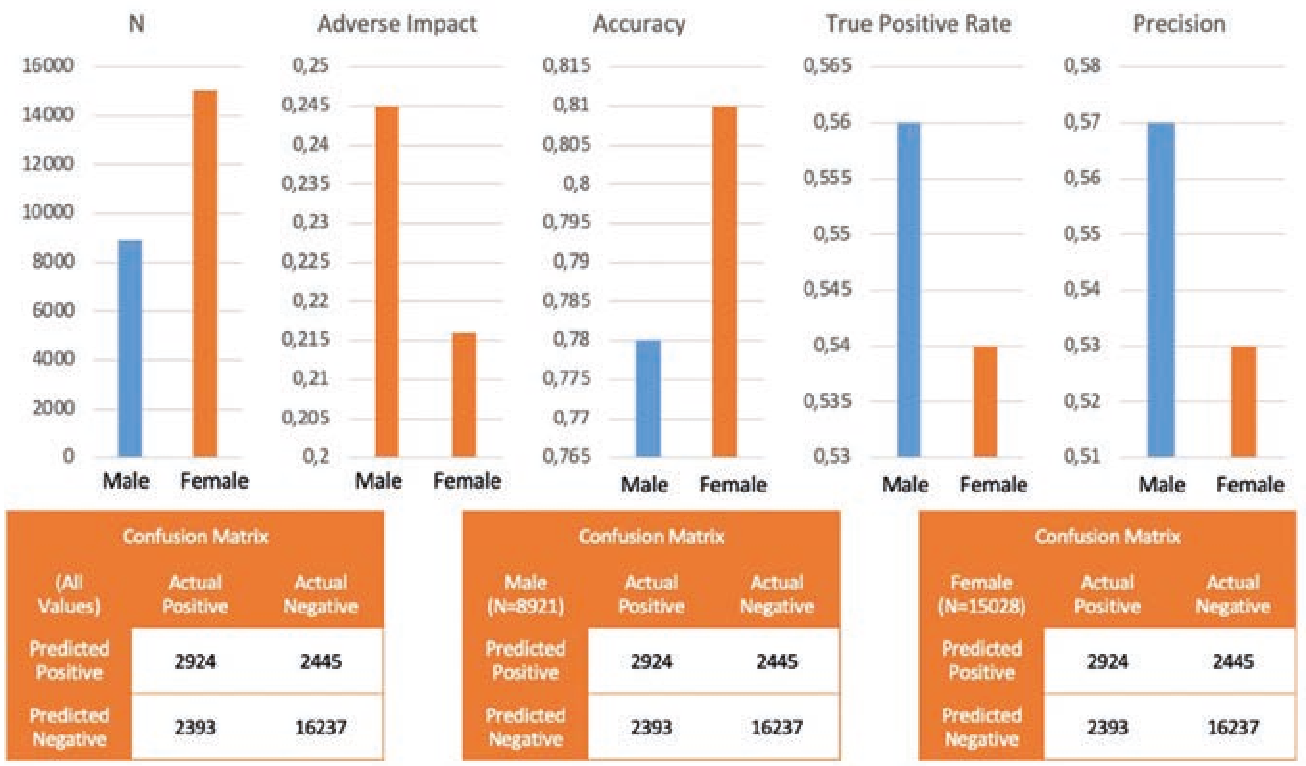
In our artificial intelligence applications, only personal data that people give explicit consent to and also approved by Personal Data Protection Law (KVKK) is used. Additionally, our bank has established a specific policy for the protection of sensitive personal data with special characteristics defined as "data that may result in discrimination or harm if learned" [7] under KVKK. In AI models; personal data relating to the race, ethnic origin, political opinion, philosophical belief, religion, religious sect or other belief, appearance, membership to associations, foundations or trade-unions, data concerning health, sexual life, criminal convictions and security measures, and the biometric and genetic data are listed as a special category in KVKK [7].

In this context, AI-based decisions are given independently of the listed attributes and any bias is rejected.

For example; in the applications that we have developed, gender information is not used in order to mitigate gender bias. Besides that, bias based on the region or location where the person was born, residind and working is prevented in the design of AI models.

In addition to avoiding the usage of data that may cause bias; correlation and effect analyses are performed. Using risky data with these analyzes, it is examined whether it causes bias in the model over other data.If, as a result of this study, a bias in risky areas emerges in the model, studies are carried out to eliminate this bias in the model. For example, if the model makes a biased decision by making gender discrimination by using other attributes without knowledge of gender, the attributes that cause this are removed from the application.

Disparate Impact Analysis is the study of evaluation and quantitative measurement of fairness in processes, applications and models that seem fair on impacted groups or individuals but create unintentionally biased or negative results. In Figure 5, an example study regarding disparate impact analysis is presented.



Confusion Matrix

(All Values)	Actual Positive	Actual Negative
Predicted Positive	2924	2445
Predicted Negative	2393	16237

Confusion Matrix

Male (N=8921)	Actual Positive	Actual Negative
Predicted Positive	2924	2445
Predicted Negative	2393	16237

Confusion Matrix

Female (N=15028)	Actual Positive	Actual Negative
Predicted Positive	2924	2445
Predicted Negative	2393	16237

Figure 5: Disparate Impact Analysis

Correlation Analysis is the study of quantitative measurement of relation in any feature and prediction tuple in processes, applications and models in question. In Figure 6, an example study regarding correlation analysis is presented.

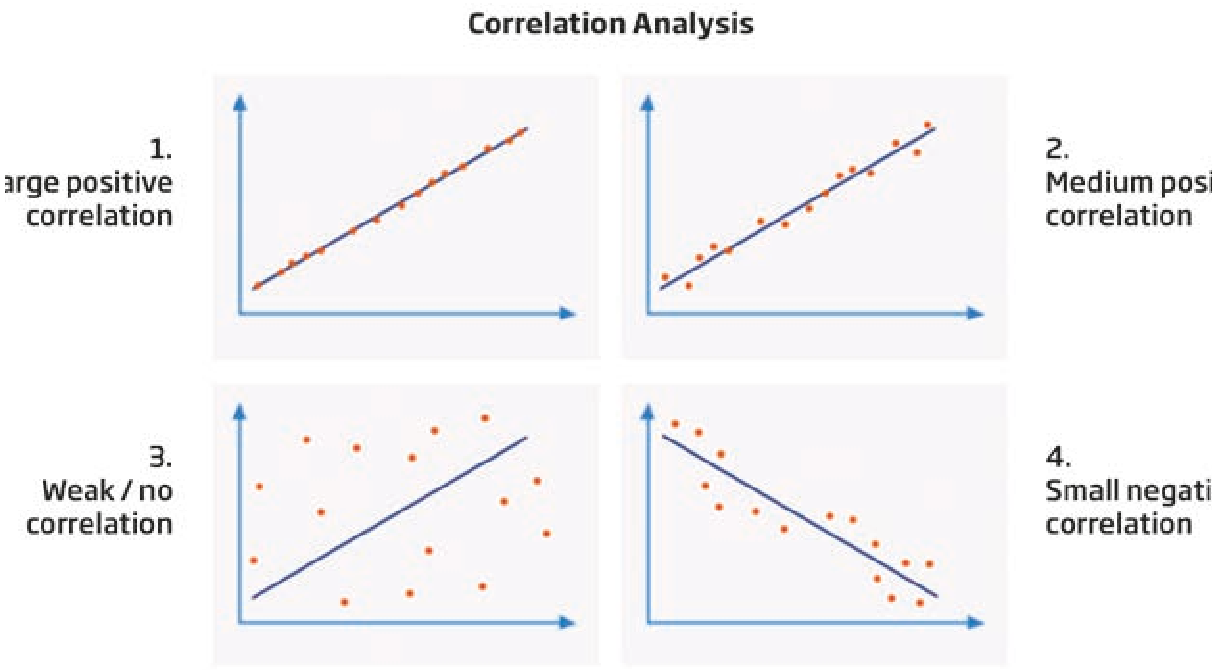


Figure 6: Correlation Analysis

Throughout the deployment of artificial intelligence applications, compliance of inputs and outputs with our bank's ethical principles and KVKK are checked by parties isolated from development teams. These controls, carried out by the Artificial Intelligence Committee, Risk Management, Internal Control and Inspection Board teams, respectively, can be exemplified as follows:

- Input data categories of the model,
- Compliance of the input fields of the model with KVKK and our bank's ethical principles,
- Inputs inducing the highest contribution to explain the target variable,
- Whether gender information is used as an input in the model,
- Underestimate/overestimate controls by gender in model estimations,
- Correlation / discrete impact analysis between target and input variables in business perspective,
- Compliance of model results with our bank's ethical principles.

Robustness

Our AI models running on our Bank's cloud infrastructure are equipped with the most prominent technology. These models are functioning in a fully automated manner open to human auditing and closed to human intervention. We provide all the possible security precautions and necessary controls for our models.

Robustness refers to the capacity of AI systems to successfully tolerate potential functionality damaging disruptions [4].

Our AI applications run on our on-prem, redundant, cloud-compatible infrastructure. The redundant structure ensures high availability by minimizing any service interruptions. Whereas, the cloud-compatible infrastructure allows our applications to scale dynamically according to the increasing workloads.

All of our input data goes through a quality control process. This guarantees a healthy and qualified data flow which is further served to the models to ensure their trouble-free employment.

Our applications are put into the production environment after passing universally accepted code and security analyses. Further, authentication and authorization of our applications are governed where unauthorized access is prohibited both at the infrastructure and application level.

AI applications are monitored in real-time which enables potential service interruptions to be quickly detected and intervened. In addition, health and the performance of AI models are further monitored. Also, interventions and revisions are made when necessary.

Our applications go through necessary testing steps before deployment and applications that do not pass these steps are not deployed into the production.

The entire model life cycle is governed within our bank using our custom Artificial Intelligence Platform Mozaik which guarantees standardization.

Privacy

We regulate our AI applications in accordance with the national, international and institutional policies. We consider and apply all the necessary measures to prevent realization of illegal actions. Regarding this issue, we are establishing cooperation with other banks and institutions; we protect the confidential customer and bank data in addition to personal data within the framework of our privacy policies.

İşbank shows special consideration in ensuring the safety and privacy of the information belonging to our customers [8]. This matter is a legal obligation of our Bank in accordance with the relevant legislation as well as it is a priority due to our Bank's sensitivity on the issue. In every application and process we pay due attention to this responsibility and we raise the awareness of our employees.

Personal data of our customers is not shared with any third parties except for official institutions and authorities with legal access to such data.

Accordingly, it is our basic approach to protect all information belonging to the Bank and to our customers from unauthorized access, false use and alteration, corrosion and destruction; ensure the privacy, integrity and availability of the information.

Confidentiality of our customers' data while collecting, processing, transferring and retaining is protected by KVKK.

Our bank ensures that organizations providing support services comply with our privacy standards and requirements.

Our bank ensures that only authorized users can access the data through data authorization.

Our teams responsible for data authorization continuously follow the rules of privacy and security.

References

[1] Takashi J. OZAKI, “Comparing machine learning classifiers based on their hyperplanes or decision boundaries,” Data Scientist TJO in Tokyo, February 6, 2014.
https://tjo-en.hatenablog.com/entry/2014/01/06/234155

[2] The Royal Society, “Explainable AI,” Royalsociety.org, November 28, 2019.
https://royalsociety.org/topics-policy/projects/explainable-ai/

[3] A. Duval, “Explainable Artificial Intelligence (XAI),” Scholarly Report, Mathematics Inst., The Univ. of Warwick, Coventry, UK, April 2019.
https://www.researchgate.net/publication/332209054_Explainable_Artificial_Intelligence_XAI

[4] “The OECD Artificial Intelligence (AI) Principles,” Oecd.ai, May 2019.
https://oecd.ai/en/ai-principles

[5] “Ethics guidelines for trustworthy AI,” European Commission, April 8, 2022.
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[6] G. Smith, N. Kohli and I. Rustagi, “What does ‘fairness’ mean for machine learning systems?,” Center for Equity, Gender & Leadership (EGAL), the Haas School of Business at the Univ. of California, Berkeley, 2020.
https://haas.berkeley.edu/wp-content/uploads/What-is-fairness_-EGAL2.pdf

[7] “KVKK, Personal Data Protection Law (Law Number 6698),” Turkish Offical Gazette, no. 29677, 2016.
https://www.kvkk.gov.tr/Icerik/6649/Personal-Data-Protection-Law

[8] “Privacy Policy” Türkiye İş Bankası, February 14, 2017.
https://www.isbank.com.tr/en/privacy-policy

TÜRKİYE İŞ BANKASI