Cemil Çelik

(+90) 541 918 9346 | <u>cemilcelik1199@gmail.com</u> | LinkedIn | Project Portfolio |

MOTIVATION

I am passionate about **solving business problems using Data Science & Machine Learning.** I **systematically** use my skillset to **add tangible value** to the team, the business, and the end-user. I am constantly learning, and always looking to improve.

SKILLS & TOOLS

Programming: Python (Base,Pandas,Numpy,Matplotlib,Scikit-Learn,Keras),SQL

Machine Learning: Linear Regression, Logistic Regression, Decision Trees,Random Forest,KNN,K-Means,PCA,Association Rule Learning,Causal Impact Analysis, Deep Learning

Other: Github,Statistics,Data Visualization,MS Office,Tableau,Jupyter Notebook, AWS

Languages: Turkish: C2(Native Language), English: C1(Full Professional Proficiency), German : B2(Professional working proficiency), French :B2(Professional working proficiency)

EXPERIENCE

Part-Time Data Scientist - Turkish Technology

July 2023 – Present

- Solving technical maintenance problems of the Turkish Airlines using Data Science & Machine Learning.

PROJECTS

Predicting Customer Loyalty Using ML | *Customer Loyalty, Machine Learning, Regression, Python*

- Used respectively Linear Regression, Decision Tree and Random Forest in Python to predict the loyalty score for the customers of a grocery that could only tag around half of its customers.
- This project helps the grocery owners understand true customer loyalty, leading to a more accurate and relevant customer tracking, targeting, and comms.
- **Metric 1: Adjusted R-Squared (Test Set)**
  Random Forest = 0.955, Decision Tree = 0.886, Linear Regression = 0.754

  Metric 2: R-Squared (K-Fold Cross Validation, k=4)
  Random Forest = 0.925, Decision Tree = 0.871, Linear Regression = 0.853

- As the most important outcome for this project was predictive accuracy, rather than explicitly understanding weighted drivers of prediction, we chose the Random Forest as the model to use for making predictions on the customers who were missing the loyalty score metric.

Assessing Campaign Performance Using Chi-Square Test For Independence | *AB Testing,Hypothesis Testing,Python*

- Applied Chi-Square Test for Independence for a grocery retailer who wants to know if there is a significant difference in signup rate for their campaign between groups of customers receiving low quality and high quality mailers.
- Low quality mailers are cheaper than high quality ones. With this project the retailer is going to make more informed decisions in the future, with the overall aim of optimising campaign ROI.
- **Chi-Square Statistic = 1.94**
  p-value = 0.16

- The resulting p-value is greater than acceptance criteria of 0.05. We retain the null hypothesis and conclude that there is no significant difference between the signup rates of Mailer 1 and Mailer 2 even though their signup rates are different.

Developing An Image Search Engine Using Deep Learning | *CNN,Data Science,Computer Vision,Python*

- Implemented the pre-trained VGG16 convolutional neural network to find similar products to the one at hand.
- After this project, customers are able to find cheaper products similar to the ones that they like.
- The model returns the products that are really similar to the one of interest with probabilites close one to another.

Fruit Classification Using A Convolutional Neural Network | *CNN,Data Science,Computer Vision,Python*

- Used a Convolutional Neural Network from Keras Deep Learning library to classify fruits for a client who wants to use a robotic sorting arm that will pick up and move products off a platform and it would use a a camera to "see" the product.
- Made use of Dropout to tackle the issue of overfitting. Added Image Augmentation to the data pipeline to increase the variation of images for the network to learn from which also helped with the overfitting.
- **Classification Accuracy On The Test Set:**
  Baseline Network = 75%, Baseline+Dropout = 85%, Baseline+Image Augmentation = 93%, Optimised Architecture+Dropout+Image Augmentation = 95%, Transfer Learning Using VGG16 = 98%

Customers' Nutrition Segmentation | *Customer Segmentation,Machine Learning, Clustering, Python*
- Used K-Means Clustering to segment up the customers based upon their their engagement with each of the major food categories, aiding business understanding of the customer base, and to enhance the relevancy of targeted messaging & customer communications.
- Made use of feature scaling which is important for distance based algorithms such as k-means. We use a process known as Within Cluster Sum of Squares (WCSS) to explore a suitable number of clusters.
- One of the clusters allocated its spendings to Fruit&Vegetables but very little to Dairy&Meat. We can make an early assumption these customers are vegan. One of the other clusters consume a lot of Dairy,Fruit&Vegetables but very little of Meat. We can assume these customers are vegetarian.

Enhancing Targeting Accuracy Using ML | *Customer Targeting,Machine Learning, Classification, Python*
- Used respectively Logistic Regression, Decision Tree ,Random Forest and KNN to classify the customers of a grocery retailer based upon their likelihood to sign up for the campaign.
- In addition to several data preprocessing tehcniques, feature & permutation importances are also used to understand the key parameters behind the signups.
- **Metric 1: Classification Accuracy**

  KNN = 0.936, Random Forest = 0.935, Decision Tree = 0.929, Logistic Regression = 0.866

  Metric 2: Precision

  KNN = 1.00, Random Forest = 0.887, Decision Tree = 0.885, Logistic Regression = 0.784

  Metric 3: Recall

  KNN = 0.762, Random Forest = 0.904, Decision Tree = 0.885, Logistic Regression = 0.690

  Metric 4: F1 Score

  KNN = 0.865, Random Forest = 0.895, Decision Tree = 0.885, Logistic Regression = 0.734
- Random Forest is the chosen model because it is the most consistently performant on the test set across the metrics and the feature & permutation importance helps the client understand the key drivers behind the signups.

Compressing Feature Space for Classification Using PCA | *PCA, Machine Learning, Python*
- Used Principal Component Analysis (PCA) to compress a big data of customers while classifying them as to whether they are likely to purchase a new album which is going to be promoted by our client.
- Plotted the percentage of explained variance for every number of components. Called a limit tou our dataset to the number of components that make up 75& of the variance of the initial feature set which is 24. Applied Random Forest Classifier
- **Metric: Classification Score** = 0.93

EDUCATION

Galatasaray University   Istanbul, Turkey
Bachelor of Computer Engineering   *Sep. 2018 – June 2024*

COURSES & CERTIFICATIONS

DSI Data Science Professional Certification
- **Actionable Learnings:** Extracting&Manipulating data using SQL.Applications of statistical concepts such as hypothesis tests.Utilising Github for version control and collaboration.Using Pyhon for data analysis,manipulation&visualization.Applying data preparation steps for ML including missing values,categorical variable encoding,outliers,feature scaling,feature selection,model validation.Applying Machine Learning for regression,classification,clustering,association rule learning and causal impact analysis for measuring the impact of an event over time.Machine Learning pipelines to streamline the ML pre-processing&modelling phase.Deployment of a ML pipeline onto a live website using Streamlit. Using Tableau to generate powerful Data Visualizations. Turning business problems into Data Science solutions.