**“Pyton Script”in izahı**

İlk olaraq, count, info, shape, column və s. kodlar vasitəsilə dataya ilkin baxış keçirdim, data ilə tanış oldum. Data cleaning mərhələsində variable-ların dəyərlərində bəzi səhvliklərin olduğunu gördüm. Məsəl üçün, daşınma rejimi variable-ında idxal dəyəri iki dəfə fərqli cür yazılmışdı, düzəlişlər etdim. Bir neçə variable-da eyni səhvlər var idi, düzəliş etdim.

Daha sonra missingno library-nin köməyi ilə missing dəyərləri gözdən keçirdim. Dataya da baxış keçirdikdən sonra gördüm ki, göndərən stansiya və göndərən ölkələr variable-ları demək olar ki, bir-biri ilə eynidir. Eyni ölkələr üçün eyni stansiya adları qeyd olunub. Bu səbəbdən göndərən stansiya variable-ındakı missing deyərləri göndərən ölkələrə uyğun şəkildə doldurdum. Təyinat ölkəsi və təyinat stansiyası arasında da eyni əlaqə olduğundan, təyinat stansiyası variable-ındakı missing dəyərləri təyin ölkəsi əsasında doldurdum. Vaqon növünü sadə mode ilə doldurmağın məntiqli olacağını düşündüm, çünki heç bir variable ilə direkt əlaqəsi yox idi. Daşınma rejimi birbaşa hansı ölkədən daşındığı ilə bağlı variable olduğu üçün, daşıma rejimi variable-ını gönderen və təyinat ölkələrin əsasında doldurdum. Yük qrupu variable-ı isə yükün adı ilə demək olar çox oxşarlıq təşkil edirdi. Eyni yük adları bir qrupa aid idi, amma bəzilərində yük qrupu missing idi. Ona görə də eyni yüklər hansı qrupa aiddirsə, missing value-larda da eyni şəkildə doldurdum. Məsələn, qorox bir yerdə taxıl bitkiləri yük qrupuna aid edilmişdi deyə missing value-da da eyni yerə taxıl bitkilərini avtomatik kod vasitəsilə əlavə etdik. Bu şəkildə bütün missing dəyərləri doldurduq.

Daha sonra data cleaning mərhələsində duplicated dəyərləri yoxladım. 75 duplicated dəyər çıxdı. Lakin həm data az idi deyə duplicated dəyərləri silmədim. Həm də eyni gün, eyni tarixdə, eyni yük ağırlığında iki qatarın yola çıxması normal hal kimi gördüyüm üçün duplicated dəyərləri silmədim.

Categorical variable-ları 1 və 0 olaraq kodlaşdırdım. Bunu edəndə isə, məsələn, yük qrupu variable-ı 4 dəyər alırdısa, meyvə, qıda, taxıl variable-ları 1 və 0 formasında ayrı variable-lar kimi kodlanır, kimya məhsulları dəyəri isə multicollinearity problemi yaratmasın deyə silinir.

Bu şəkildə kodlama apardıqdan sonra MinMaxScaler ilə bütün variable-lar üçün deyərlər 1 və 0 arasına gətirilir. Əgər bu edilməsə, onda yüksək ədədləri olan variable-lar target-ə daha çox, kiçik ədədləri olan variable-lar isə target-ə daha az təsir edir. Belə olan halda model o qədər də yaxşı öyrənə bilmir. Bütün variable-lardakı ədədləri eyni rəqəm aralığına gətirmək modelin nəticəsini yaxşılaşdırır. Daha sonra data train və testə bölünür ki, modeli train edəndən sonra onun keyfiyyətini ölçək.

Qərara gəldim ki, GradientBoostingRegressor ensemble learning modelini seçim hansı ki, bir neçə modeli (əsasən decision tree) ard-arda quraraq bir əvvəlki modeldəki səhvin üstündə işləyir, onu daha da təkmilləşdirir. Sadə bir modelə görə daha yaxşı bir işləmə sistemi olduğu və etdiyi səhvlərin üzərində işləyib daha daha onu təkmilləşdirdiyi üçün dəqiq accuracy əldə etmək olur. Buna görə bu modeli seçdim.

Ən uyğun parametrləri tapmaq üçün BayesianSearch-dən istifadə etdim. Bayes teoreminə əsaslanan bu hyperparameter tuning methodu təxmini bir dəyər (məsələn max\_depth parametri üçün) seçir və ona görə nəticəyə baxır, nəticə yaxşılaşırsa həmin istiqamətdə davam edir, nəticə pisləşsə əgər, əksinə digər etimallara baxır (səsli izah etmək daha rahat olardı). Bu şəkildə parametrlərin ən optimal halını seçir. Bu dəfəki case-də mən number of estimators (yeni boosting model nə qədər model quracaq səhvləri düzəltmək üçün), learning rate (step sizə hansı ki model datanı öyrənir, çox olsa bu qədər underfitting, az olsa overfitting olacaq), max\_depth (hər qurulan tree-nin maximum dərinliyi, sualların sayı), min\_samples\_split (yeni ki ən az neçə sample olmaldır ki decision tree bölünmə etsin, yeni sual soruşsun), minimum samples leaf (yəni decision tree qurulanda hər qolunda ən az nə qədər sample olmaldır ki decion tree qurulsun), subsample (yeni datadan nə qədərlik sample götürüb decision tree quranda model daha yaxşı nəticə göstərir) parametrlerini seçdim. Qurulan modelleri r2 score-a görə müqayisə etməsini istedim, 10 iterations (neçə müxtəlif hyperparameter kombinasiyası qurulmalıdır). Modeli data üzərində train etdikdən sonra yenidən ən yaxşı data ilə modeli train etdim.

Modeli dəyərləndirmə hissəsində isə train MAE (36.8 - actual və predicted value-lar arasında olan average absolute difference 36.8-dir), RMSE (63.43 - RMSE MAE-ə nisbətdə large errorlara daha çox sensitive-dir), R2 (0.87 - target variable-da olan variance-ların 87 faizi model tərəfindən interpret edilə bilir) modelin yaxşı nəticə verdiyini göstərir. Test datadakı 43-lük average absolute error, 92-lik RMSE və 0.63-luk R2 göstərir ki, modeldə azca overfitting göstəriciləri olsa da datanın keyfiyyətini və kəmiyyətini (sadəcə 500 sətir olması) nəzərə alsaq yaxşı nəticə kimi görmək olar.

Sonda isə true və predicted dəyərləri vizual şəkildə scatter plot və line chart ilə vizuallaşdırdım.

Nəticə olaraq, modeli deploy etdikdən sonra yüklərin həcmi barədə real rəqəmlərə yaxın predictionlar əldə etmək olar.

**Vizual hesabatın yazılı izahı**

**Qrafik 1 - "Bump Chart"**

"Bump Chart"da 2014-cü ilin hər ayı üzrə ən çox yük göndərən 10 ölkənin sıralaması göstərilmişdir. Bu qrafikə parametr də əlavə etdildiyi üçün ən çox yük göndərən neçə ölkəni görmək istəyiriksə parametrlərdən göstəricini sürüşdürməklə həmin rəqəmi seçə bilərik. "Default" olaraq həmin rəqəm 10 olduğundan biz qrafikdə "top" 10 ölkəni götürürük. Rusiya və ABŞ ən çox yük göndərən ölkələrin içində il boyu birinci və ikinci yeri bölüşür. 3-cü ölkələr mart, aprel, may və iyun aylarında "ranking"də ilk üçdə olmaqlarına baxmayaraq, digər aylarda daha az məhsul göndərirlər. Litva ilin əvvəlində az məhsul göndərsə də, payız aylarına doğru sıralamada üst pillələrdədir. Kanada da həmçinin qış aylarında çoxlu yük göndərməsinə baxmayaraq, digər aylarda sıralamada aşağıdadır. Ümumilikdə qrafik bizə ən çox yük göndərən ölkələrin aylar üzrə sıralamasını görmək üçün olduqca yaxşı imkan yaradır. Bu qrafikə baxaraq, hansı aylarda hansı ölkələr üzrə daha çox (yaxud az) qatar "allocate" etmək olar, onu müəyyənləşdirmək olar.

**Qrafik 2 - "Donut Chart"**

"Donut Chart"da 2014-cü il üzrə daşınan məhsulların növlərini müqayisəli şəkildə görə bilərik. Belə ki, daşınan yüklərin 45.02%-ini Təzə meyvə, giləmeyvə, tərəvəz və fındıq təşkil edib. Qida sənayesi məhsulları isə ümumi daşınan 40.53%-i qədərdir. Taxıl məhsulları və dənli bitkilər 13.36%, kimya sənayesi məhsulları isə 1.09% təşkil edir.

**Qrafik 3 - "Reversed Lolipop Chart"**

"Lolipop Chart"da 2014-cü il üzrə ən çox yük qəbul edən ölkələrin siyahısı sıralama ilə (çoxdan aza) verilib. İlk üçlükdə ~110 min ton yük qəbul edən Azərbaycan, ~40 min ton yük qəbul edən Türkmənistan və ~29 min ton yük qəbul edən Tacikistan yer alır. Bu ölkələri Qazaxıstan (~17 min ton yük), Gürcüstan (~8 min ton yük), Özbəkistan (~3 min ton yük), Rusiya (~2 min ton yük), 3-cü ölkələr (440 ton yük) və Əfqanıstan (280 ton yük) izləyir. Ən çox yük qəbul edən ölkələri görmək və müqayisə etmək üçün qrafik olduqca yaralıdır.

**Qrafik 4 - "Tree Map"**

"Tree Map"da ən çox yük daşıyan ekspeditorların siyahısı, eyni zamanda ən çox yük sifariş verən şirkətlərin siyahısı göstərilmişdir. Parametrlər bölməsindən ən çox yük daşıyan ekspeditorların siyahısına, yaxud ən çox yük sifariş verən şirkətlərin siyahısına baxacağımızı seçə bilərik. Daha çox yük sifariş verən şirkətlərə, yaxud digər siyahıda daha çox yük daşıyan şirkətlərə daha böyük sahə ayrılıb, üzərlərində yükün miqdarı və şirkətin adı yazılıb. Ən çox yük daşıyan şirkətlərin siyahısına "Aztransrail" MMC (95 620 ton), "Azərbaycan Şəkər İstehsalat Birliyi" MMC (66 620 ton) və Marketinq və İqtisadi Əməliyyatlar İdarəsi (25 150 ton), ən çox yük sifariş verən şirkətlərin siyahısına isə "Aztransoil" MMC (130 510 ton), "Trans Kafkaz Transport" MMC (53 820), "Aztransoil" MMC (10 940 ton) başçılıq edir.

Qeyd edim ki, kursoru şirkətlərin adlarının üzərinə gətirdikdə onların yüklərinin daşınma rejimi "Pie Chart" kimi "tooltip" formasında ekranda görünür. Buradan biz hər şirkət üzrə yüklərin daxili, ixracat, idxalat, yaxud tranzit yükdaşıma olduğunu faizlə göstərir. Məsəl üçün, ən çox yük daşıyan şirkət - Aztransrail" MMC ən çox idxal edilən yükləri (80.01%), daha sonra isə tranzit yükləri (19.01%), daxili (0.91%) və ixracat yüklərini (0.46%) daşıyıb. "Tooltip"dəki "Pie Chart"a ayrılıqda baxanda isə ümumi yüklərin 52%-nin idxalat, 46.3%-nin tranzit, 1.22%-nin ixracat və 0.48%-nin daxili yüklər olduğunu görürük.

**Qrafik 5 - "Box Plot"**

"Box Plot" hər vaqon növü üzrə daşınan məhsulların çəkilərinin paylanmasını müqayisəli formada göstərir. Qrafikdən məlum olur ki, çən növlü vaqonlarda daşınan yüklər daha ağır çəkiyə malik olur, əksəriyyəti isə 680 ton civarında olur. Qrafikə əsasən deyə bilərəm ki, çən vaqonda daşınan ən yüngül yük konteyner, platforma, yarımvaqon tipli vaqonlarda daşınan ən ağır yükdən daha çox çəkiyə malikdir. Qrafikə baxdıqda həmçinin o nəzərə çarpır ki, konteyner vaqonlarda daşınan yüklərin çəkilərinin aralığı olduqca aşağıdır (3700-4200 ton). Lakin qapalı tipli vaqonlarda bu rəqəm 50-900 ton arasında dəyişir. Yeni bu tipli vaqonda daşınan ən yüngül yük 50 ton, ən ağır yük 900 tondur.

**"Dashboard" üçün qeyd**

"Dashboard"da "action"dan istifadə etmişəm. "Lolipop Chart"da hansı ölkənin adını seçsək, digər qrafiklərdəki məlumatlar o ölkənin datasına görə yenidən sıralanır. Məsəl üçün, "Lolipop Chart"da Azərbaycanı seçdikdə, "Bump Chart"da Azərbaycana göndərilən məhsulların ən çox hansı ölkədən gəldiyini, "Donut Chart"da Azərbaycana daşınan məhsulların növlərinin faiz göstəricilərinu, "Tree Map"da Azərbaycandan ən çox yük daşıyan və ya ən çox yük sifariş verən şirkətləri, Azərbaycana yük daşıyan vaqonların növlərini görə bilərik.